INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

TRANSFER LEARNING AND LORA ADAPTATION OF MBART-50 FOR LOW-RESOURCE BRAHUI RESPONSE GENERATION: A PERFORMANCE AND RANK ABLATION STUDY

Naseer Ahmed[*1], Zeeshan Ali[2], Muhammad Geand[3], Samreen[4], Zahid Latif[5], Saba Gull[6]

[*1,2,3,4,5,6]*Computer Science Department, Balochistan University of Engineering and Technology, Khuzdar, Balochistan, Pakistan.*

DOI: https://doi.org/10.5281/zenodo.24218118

Keywords
Brahui language; low-resource NLP; chatbot; conversational AI; mBART-50; transfer learning; LoRA, multilingual transformer PEFT.

Article History
Received: 05 August 2026
Accepted: 19 October 2026
Published: 10 November 2026

Copyright @Author
Corresponding Author: *
Naseer Ahmed

Abstract
Languages with limited computational resources have been poorly represented in conversational AI due to the need for extensive training data and the resource requirements of neural dialog systems. Brahui, a Dravidian language spoken mainly in Balochistan, Pakistan, stands out as an especially difficult challenge because it has limited computational resources and conversational datasets. This paper examines the ability to adapt the multilingual pretrained model mBART-50 to generate Brahui conversational responses using transfer learning and parameter-efficient fine-tuning. 1,113 pairs of questions and answers in Brahui were split into 890 training samples, 111 validation samples, and 112 test samples. Full fine-tuning was contrasted with Low-Rank Adaptation (LoRA), with an analysis of ranks 4, 8, and 16. The evaluation was performed using METEOR, Translation Edit Rate (TER), and semantic similarity. Full fine-tuning scored 78% METEOR, 23% TER, and 89% semantic similarity. The LoRA with rank 8 scored the highest of all configurations tested with 80% METEOR, 20% TER, and 89% semantic similarity, whereas ranks 4 and 16 did not perform as well. It has been concluded that multilingual transfer learning may be a practical basis for developing conversational AI in Brahui, and the choice of LoRA rank apparently affects the performance of the model. The results should be considered preliminary because each configuration was evaluated in one run on a relatively small test set. The research provides an empirical basis for future work in dialogue generation in Brahui.

INTRODUCTION

Improvements have been made in natural language processing (NLP) from rule-based methods and statistical models to transformer-based architectures that are capable of self-attention for efficient language modeling and generation [1]. Nonetheless, the development in this area has been uneven across languages. Efficient models are trained using large quantities of data that are only available for rich resource languages, while most of the other languages do not even have texts or computing facilities [2]. Lack of data is considered the major limitation in developing efficient models in low-resource machine translation [3],[4],[5], and the same applies to conversational and question-answering systems for low-resource languages. This study develops an efficient Transformer for Brahui, a

language of the Dravidian family spoken mostly in Balochistan, Pakistan. Brahui has an extensive history of literary works and translation but has not seen much technological advancement and linguistic documentation [6], [7], [8]and lacks the computational resources to develop modern NLP models. As such, training the generative Transformer from scratch is not a practical solution in our case. Research has shown that utilizing information from multilingual or similar languages is a useful way in the case of lacking data [9], [10], [11], [12].

A complete fine-tuning method revises all model parameters, which incurs a high computational cost. A parameter-efficient fine-tuning (PEFT) approach trains only a part of the constraints [13], [14],[15]. One of the most used PEFT approaches is Low-Rank Adaptation (LoRA), which models weight updates via a trainable low-rank matrix [16]. Experiments in low-resource translation suggest that PEFT can achieve the same level of performance as full finetuning [17], [18],[19]. Nevertheless, the effect of LoRA depends on the specific configuration, on the rank of the update, for which adaptive selection was recently proposed because not all tasks may require the same fixed rank [20]. Thus, there is a need to investigate various ranks rather than one arbitrarily chosen rank.

Despite the advances made, two problems still exist. First, there is a lack of experimental studies about the adaptation of multilingual pretrained sequence-to-sequence models to Brahui conversation response generation, and second, there are no experiments regarding the influence of LoRA rank in this context. In this paper, the problems are addressed by conducting experiments on Brahui conversation response generation using a Brahui corpus with 1,113 question-answer pairs and mBART-50. The adaptation of the model via full fine-tuning and LoRA is studied; an ablation study on LoRA is conducted for ranks 4, 8, and 16, and the first experimental baseline with some limitations is established to be used by future researchers working with Brahui dialogue. The structure of the paper is as follows. Literature Review is given in Section 2, Methodology in Section 3, Experimental Results & Limitations in Section 4, while Conclusion in Section

1.1 Contributions of this study

The proposed study uses a Brahui conversational dataset of 1,113 question-answer pair samples that may be used for conversation response generation. Instead of training the model from scratch, a transfer learning approach is employed, in which the pre-trained mBART-50 multilingual sequence-to-sequence Transformer is adapted for Brahui conversational generation. The full fine-tuning of mBART-50 is regarded as an important benchmark for this task, while LoRA is explored as a parameter-efficient adaptation technique for the very same multilingual model. Three ranks for LoRA (4, 8, and 16) are tested, and the best results are shown by rank 8. In terms of comparison of the adaptation techniques, full fine-tuning demonstrates better results compared to all LoRA ranks regarding the quality of generated samples.

2. Related Work and Research Background

Creation of a conversational model in an under-resourced language like Brahui will need to consider the following three interrelated areas of research: neural language processing in low-resource conditions, multilingual transfer learning, and efficient fine-tuning of pre-trained transformers. This part provides the necessary background on these topics with a focus on the papers dealing with low-resource languages, multilingual seq2seq models, dialogues, and parameter-efficient fine-tuning.

2.1 Neural Language Processing in Low-Resource Contexts

The key issue in NLP within the low-resource context is the scarcity of training materials for particular languages. Unlike the case with resource-rich languages, there might not be enough parallel corpora, annotated datasets, standard computing tools, and performance benchmarks for low-resource languages. As a result, neural models trained to work well with resource-rich languages can suffer a significant

drop in performance if they are applied to low-resource languages. This issue is especially apparent in the research related to Manipuri. Singh and Singh explored the task of translating from English into Manipuri in a highly resource-scarce environment and found that semi-supervised learning can help improve translation performance by leveraging extra monolingual data. They demonstrated experimentally that self-training and back-translation techniques can outperform both supervised learning methods and multilingual models [21],[22].

Another relevant study considered English-Khasi translation, which is another relatively under-resourced language. The authors have created parallel corpora and compared systems based on LSTM, GRU, and Transformer, and researched the possibilities of transfer learning. The results obtained highlight the necessity of resource development together with pre-trained or transferred models in the creation of NLP systems for under-resourced languages [23].

In addition to the exploitation of text resources, the research on low-resource languages also paid attention to the possibility of using additional information. Created multimodal resources for Manipuri and studied whether correlated modalities can improve translation in the low-resource environment. The results obtained showed that extra information resources can compensate for the lack of textual data to some extent [24]. Further research on multimodal translation shows that the exploitation of additional information resources can help in developing low-resource translation [25]. One more area of investigation was the choice of the most suitable languages and their resources for the multilingual transfer. Mi and Xie proposed a framework for the evaluation of language relatedness considering linguistic, semantic, and translation-related characteristics. The results obtained show that the composition of multilingual training data affects the effectiveness of cross-lingual transfer, especially in the case where the target language is under-resourced [26]. This problem is relevant for Brahui since the language cannot rely on the presence of a big language-specific corpus. In this way, the creation of a Brahui conversational system should involve the exploitation of information obtained through multilingual pretrained models while adapting it to Brahui conversational data.

2.2 Multilingual Pretrained Models for Low-Resource Languages

The rise of multilingual pretrained sequence-to-sequence models transformed the strategy of developing low-resource NLP solutions. Instead of training an encoder-decoder model completely from scratch using randomly initialized parameters, one can start with a pre-trained multilingual model and adapt it using a limited dataset of the target language. Such an approach was studied in several South Asian languages. Munaf et al. introduced a framework for low-resource summarization that uses pretrained multilingual models for the Urdu language and tested mBERT and mT5 models in resource-constrained settings. Their results show that it is possible to adapt a pretrained multilingual architecture to a language with considerably less data than English while keeping its useful contextual representations [27]. The benefits of multilingual transfer were also confirmed in the low-resource translation research. The English-Khasi language pair demonstrated the possibility of implementing transfer learning in neural translation of a language with limited resources, and analysis of different architectures revealed that the Transformer-based models provide a good starting point for further fine-tuning [22]. At the same time, multilingual transfer cannot be considered a recipe for success for every low-resource language. The applicability of the transfer method depends on the linguistic proximity of languages, writing system, domain, the amount of data used for training, and target language representation in the pretrained model. This idea is highlighted by language relatedness analysis by Mi and Xie, which shows that choosing relevant languages for multilingual training involves more than just the number of languages [25].

The problem at hand is particularly pertinent to Brahui. Even though mBART-50 provides

multilingual sequence-to-sequence learning, Brahui is not among the languages that the model explicitly supports. Therefore, the current work does not regard mBART-50 as the pretrained Brahui model. It rather examines whether the multilingual knowledge embedded in the pretrained model can be transferred to Brahui via fine-tuning in supervised conversations.

2.3 Conversational and Dialogue Systems for Low-resource Languages

In contrast to classical NLP tasks, conversational AI requires a response to be relevant to the user's input, semantically correct, and grammatically well-formed. In low-resource languages, these criteria pose even more difficulties due to the lack of data and evaluation sources. Studied task-oriented dialogue between humans and computers under the conditions of low-resource languages and stressed the significance of correct language understanding for intention recognition and slot filling. Their findings reveal that dialogue systems working in low-resource languages experience difficulties in semantic interpretation that cannot be resolved with only more complex models [28]. Efficient parameter adaptation was also studied for task-oriented dialogue specifically. Mo et al. studied PEFT in Transformer-based dialogue systems and pointed out the high computational cost connected to tuning the whole parameter set of pretrained models. It was shown in their study that adaptation techniques can be helpful when one needs to specialize a dialogue model for a downstream task without fine-tuning all parameters repeatedly [29] [30]. In the area of dialogue generation, it is also shown that automated evaluation should pay attention to deeper properties of a response rather than its surface structure. Li et al. created a response generation approach based on discourse features and response assessment to ensure semantic and affective consistency. It was shown in their research that evaluating the quality of generated responses in terms of semantic and contextual properties is necessary, since a response may have adequate word overlap with no proper semantics and context at all [31].

However, more recently there have been studies that focused on the semantic variation among languages during multilingual dialogue. Wu et al. looked at multilingual open-domain dialogue generation and found the existence of semantic variance among languages to be a problem. This shows that there can be language-specific issues in terms of dialogue response quality and semantics for multilingual dialogue systems, which means that there cannot be a uniform approach to multilingual dialogue generation systems [32]. Such problems justify the evaluation method used in the current Brahui dialogue system. Given the fact that the aim is dialogue response generation, rather than classification, METEOR, TER, semantic similarity, and perplexity, where appropriate, are evaluated.

2.4 Parameter Efficient Fine Tuning

While full fine-tuning trains all parameters of a pretrained model on the task, this approach becomes increasingly computationally demanding as the size of the model increases. This issue is addressed by parameter-efficient fine tuning, which requires optimizing a fraction of the parameters in a pretrained model. Ding et al. performed an extensive empirical evaluation of parameter-efficient adaptation methods and concluded that only a fraction of the parameters of a pretrained model needs to be adapted for efficient downstream adaptation. They conducted a comparison of various adaptation methods on over 100 NLP tasks and concluded that parameter-efficient adaptation can achieve comparable results while greatly decreasing memory and computational demands. It is important to note that their results are especially applicable to the field of low-resource NLP since the computational demands of large pretrained models can become a more restrictive factor when one works with limited computing resources and a small amount of data. What is also notable is that their experiments show that the success of parameter-efficient adaptation does not solely depend on the number of trainable parameters but rather on the adaptation mechanism itself [33] [34].

LoRA forms a very significant approach within this family. Instead of performing updates on the original weight matrix, the method uses low-rank trainable matrices for parameter updates. As a result, most parts of the pre-trained model can remain untouched while the model gains task-specific knowledge through low-rank parameter updates. One justification for the importance of the PEFT approach for dialogue agents comes from Mo et al. The authors have shown how parameter-efficient adaptation can be applied for task-oriented dialogue modeling and revealed the benefits of computational efficiency and memory requirements [29],[35]. Later research in this area has also focused on efficient adaptation techniques that apply to ever-growing large language models and even allow selecting specific parts of the model for adaptation Biswas et al., for instance, have found out that partial fine-tuning through layer selection technique reduces the number of trainable parameters significantly without sacrificing task performance across multiple NLP tasks [36].

2.5 LoRA and the Significance of Adaptation Rank

The rank of trainable low-rank matrices of LoRA is an important property that affects the dimensionality of the adaptation subspace and, thus, helps to control the adaptation-capacity-efficiency trade-off. Very low rank constrains the dimension of the subspace in which task-specific knowledge can be adapted. In contrast, higher rank allows learning more complex adaptations but leads to a higher number of trainable parameters. Nowadays, there has been increasing interest in studying properties of LoRA rank that go far beyond simple adjustment of the number of trainable parameters. Kim et al. have looked at possible alternative adaptation structures and found out that different adaptation branches can adapt to different kinds of knowledge from pretrained representations [37]. This finding means that the architecture of the adaptation mechanism could play a significant role in downstream performance beyond just increasing the number of trainable parameters. There have been some attempts to improve rank selection lately. Cao et al. suggested Efficient LoRA and showed that regular LoRA may retain some inefficient ranks in the adaptation subspace. They developed an algorithm that finds more useful rank components and discards the rest and show that the effective rank of the adaptation mechanism influences efficiency and performance [38] . LoRA is a multi-scale low-rank adaptation mechanism that uses multiple low-dimensional adaptation spaces instead of one [39] . Layer-wise adaptive LoRA is also an example showing that a unified setting for the adaptation process may not be optimal at all times. According to Gu et al., different layers may have different properties of adaptation [20]. Therefore, there is both theoretical and practical support for running rank-sensitivity experiments as opposed to choosing one LoRA rank without any comparison. The present research takes a similar approach in a relatively simplified way. In other words, the researcher does not propose any novel adaptive rank algorithm but uses the following LoRA settings, r = 4, r = 8, and r = 16, within the same task of Brahui conversation. As a result, it will be possible to find out if an increase in adaptation rank positively influences response generation performance.

2.6 Tokenization and Representations for Low-Resource Languages

Another aspect to consider when discussing low-resource languages is tokenization. In situations where the target language does not appear within the vocabulary or language ID set of the pretrained model, tokens of the target language can be broken down into subwords that will not make sense linguistically compared to subwords of other high-resource languages. For this low-resource machine translation, researchers have explored character-based tokenization. Rathod et al. have done research on character-based encoding for Hindi neural machine translation. The work done by them showed improvement over their baseline settings on a Hindi-English dataset. Their findings show that character-based representations can provide useful information in cases where

subword and word-level representations are not sufficient [40].

The same authors examined the use of semantically weighted back-translation in low-resource and morphologically diverse languages. In this work, the focus is on the role of linguistic properties and the generation of training examples in the case of languages with morphological diversity, which increases their vocabulary size and decreases the value of a small parallel corpus[41] , [42] . Regarding the chatbot under development, this question is especially pertinent as the Brahui language has certain orthographic features, which are inherent to it and the Perso-Arabic script used by Brahui. For this reason, in the current research, it is decided to maintain the same model configuration for all experiments and not to implement any extra language-specific embedding fusion technique.

2.7 Augmenting Resources and Effective Adaptation

A related important research area in NLP for low-resource languages concerns augmenting limited data resources. Bhuvaneswari and Varalakshmi examined an incremental NMT-SMT method for low-resource English-Tamil translation where machine-translated monolingual data was selected and fed into training phases. The results prove that proper selection of auxiliary data may contribute to improved translation when parallel corpora are not large [43]. Also, large language models inspired another perspective on data reduction required for adaptation. Zhu et al. conducted research on demonstration selection for in-context machine translation and found out that choosing the right demonstration may improve the translation capability and efficiency of machine translation. These results corroborate the general notion about the importance of example quality and selection besides model architecture [44].

These approaches can be considered supplementary to the current study. In the current study, the authors intentionally create an experiment where the same conversational dataset of the Brahui language is used for comparison between mBART-50 fine-tuning and LoRA settings. Thus, augmentation, retrieval, and demonstration-based adaptation are deferred to further research.

2.8 Research Gap and Positioning of the Current Study

Several key results are found in the reviewed literature. First, there is potential for substantial gains in the performance of low-resource languages due to multilingual pretraining and transfer learning, but such gains depend on language properties, data availability, and the relation between the target language and languages covered by the pretrained model [21]. Second, conversational systems need to be evaluated not only based on lexeme matching but also based on the semantic relevance and context appropriateness of responses [28], [31]. Third, parameter-efficient adaptation presents an appealing option compared to fine-tuning, especially for resource-constrained settings, but PEFT techniques do not outperform fine-tuning consistently [29]. Finally, current studies show that the choice of configuration and rank of low-rank adaptation may have a significant impact on model behavior [37] . Nevertheless, there is a clear lack of work on response generation in the Brahui language using multilingual pretrained sequence-to-sequence models. In low-resource research, a large portion of the work done has been related to machine translation and other classical NLP tasks rather than conversational resources and adaptation techniques for Brahui. Even more, there are no adequate empirical works comparing fine-tuning to LoRA-based adaptation for Brahui conversational responses and the effect of different ranks of LoRA adaptation.

To fill this gap, an empirical study with a controlled experimental design is conducted in the current paper. It uses transfer learning with the mBART-50 pretrained multilingual model and tests both full model fine-tuning and LoRA-based fine-tuning on the same Brahui conversational dataset using rank ablation with r = 4, 8, and 16. Thus, the paper does not introduce any embedding fusion or other representation technique. The contribution of this paper is

empirical analysis of multilingual transfer, full model fine-tuning, LoRA parameter-efficient adaptation, and LoRA rank selection for low-resource Brahui conversational AI. This is also the basis for methodology used in the next section that discusses Brahui conversational dataset, data preparation, preprocessing, train-validation-test split, and problem statement.

3. Methodology

This research work analyses the adaptation of a multilingual sequence-to-sequence architecture for generating responses in conversations using the low-resource language, Brahui. There are four phases in the proposed methodology, which include data collection, problem definition, mBART-50 transfer learning, and parameter-efficient adaptation using LoRA. The same dataset split is applied to all the models under consideration.

3.1 Dataset Preparation

The conversation dataset includes 1,113 question-and-answer pairs in Brahui. Each sample comprises the question asked by the Brahui user and the Brahui reply to it. Thus, the data set is created as a supervised sequence-to-sequence learning data set, in which the input sequence is the question and the output sequence is the response. The data set is represented in a structured form with two columns, wherein the first column has Brahui questions, and the other column has Brahui answers. The final data set is split into training, validation, and test sets with a random seed. The splitting process resulted in 890 training samples, 111 validation samples, and 112 test samples.

International Journal of Advanced Research

Table 1 Dataset composition and partitioning

Dataset component	Number of pairs	Percentage
Training	890	80.0%
Validation	111	10.0%
Test	112	10.1%
Total	**1,113**	**100%**

Table 1 shows that the dataset has been segmented into three subsets; the training set has 80%, the validation set has 10%, and the testing set has 10.1%.

3.1.1 Text Processing

Both textual variables were first transformed into strings and then passed to the tokenizer used in Transformer. The dataset was tokenized with the help of the default mBART-50 tokenizer, which has the maximum length of a sequence of 128 tokens. Sequences longer than that were truncated. The same preprocessing step was performed for the training, validation, and test subsets. Urdu (ur_PK) language was chosen as an experimental language setting, as Brahui and Urdu use the Perso-Arabic script. Nevertheless, the same script representation does not guarantee proper subword tokenization of the Brahui language. Thus, the use of the mentioned language setting should be viewed as an experimental approach, but not as native support of Brahui by mBART-50.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

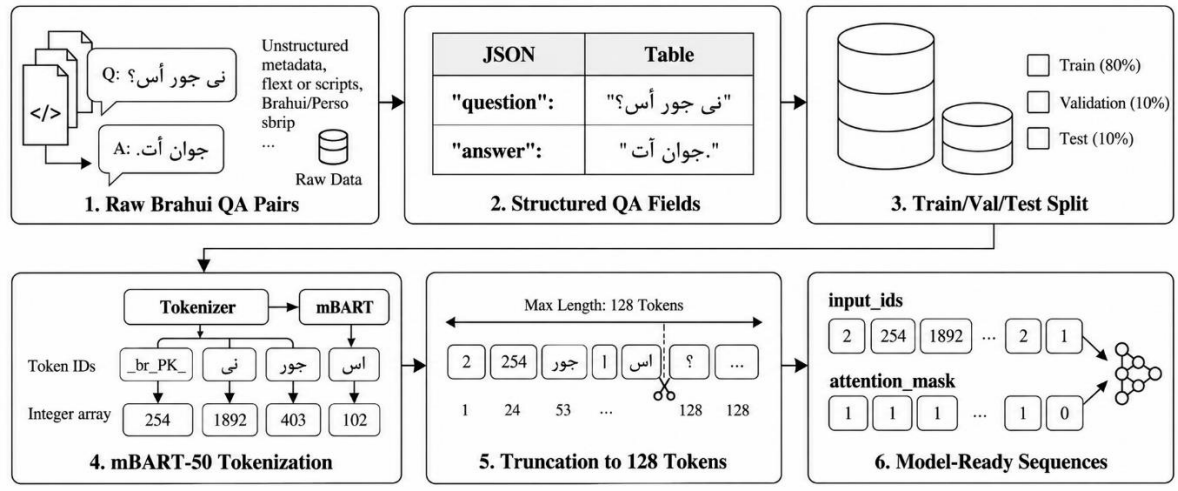


Figure 1 Brahui Question-Answer Dataset Preprocessing Pipeline

Figure 1 depicts the Brahui question-answer pairs before arranging them in a structured JSON format with questions and answers in it. These datasets are then split into three sets, with 80% being used for training, 10 % for the validation, and 10 % for the testing. After that, the process of tokenization using mBART-50 transforms the data from text format to numeric ID format. Every sequence is padded or truncated to have a maximum length of 128.

3.1.2 Dataset Splitting

The splitting process used a constant seed of 42. The first split resulted in 80% of the samples in the training dataset, and 20% went to the temporary dataset. The second split divided the temporary dataset into equal halves of validation and test datasets to get the final 890/111/112 split. This is necessary as the goal of the current research is to examine not only the transfer learning of mBART-50 but also the impact of LoRA rank on the controlled dataset.

3.2 Problem Formulation

The dialogue brief is framed as conditional response generation. Given the dataset as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$$

Where xi represents a Brahui question and yi represents a corresponding Brahui answer, and N=1113.

Given an input question

$$x = (x_1, x_2, \dots, x_m),$$

the model generates a response

$$y = (y_1, y_2, \dots, y_n).$$

The objective is to estimate the conditional probability

$$P(y \mid x) = \prod_{t=1}^{n} P\,(y_t, \mid y_{<t} x),$$

where $y_{<t}$ represents the previous target tokens generated.

This model learns by minimizing the negative log-likelihood of the sequence:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{n} \log P_\theta\,(y_t, \mid y_{<t} x).$$

Here, θ represents the parameters to be learned. For fine-tuning, the pretrained mBART-50 parameters are updated. For LoRA, the pretrained parameters are kept frozen while only the low-rank adaptation parameters are tuned.

Thus, the central methodological comparison is:

Full Fine-Tuning vs. LoRA$(r, =, 4r = 8r = 16)$.

It is a direct answer to the research question of whether parameter-efficient adaptation could be used to specialize a multilingual pretrained model to Brahui conversation and the effect of adaptation rank.

3.3 mBART-50 Model Architecture

mBART-50, a multilingual encoder-decoder model pre-trained via multilingual denoising, is employed for the proposed framework. The mBART model is specifically aimed at transferring multilingual representations to downstream tasks of sequence generation and has shown effectiveness for low-resource translations. The mBART-50 model uses the encoder-decoder model structure. Although the encoder processes the input queries into context embeddings, the autoregressive decoder processes the outputs sequentially. For a given Brahui language query x, the encoder generates

$$H = Encoder_{\theta_e}(x),$$

and the decoder estimates the probability of each response token:

$$P(y_t, | \, y_{<t} H) = Decoder_{\theta_d}(y_{<t}, H).$$

The final response is generated from the decoder output.

3.4 Full mBART-50 Fine-Tuning

The full fine-tuning setup serves as the main transfer learning setup against which the parameter-efficient setups are compared.

In the setup, the parameters of the pretrained mBART-50 model are tuned during training:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta; \mathcal{D}_{train}).$$

This means that both the encoder and the decoder can be trained on the task of talking in the Brahui language.

Full fine-tuning was done for 10 training epochs. The model that emerged from this process serves two functions:

1. First, it sets the bar on how well one can do when the whole pre-trained model is fine-tuned on the Brahui dataset.
2. Second, it is the benchmark against which one can compare LoRA to see if it performs competitively with a much smaller number of trainable parameters.

3.5 LoRA-based mBART-50 Adaptation

LoRA is proposed as the parameter-efficient component of the proposed system. LoRA replaces the pre-trained weight matrix W by freezing W and learning a low-rank update of W:

$$W' = W + \Delta W$$

where

$$\Delta W = BA,$$

with

$$B \in \mathbb{R}^{d \times r}, \; A \in \mathbb{R}^{r \times k},$$

and $r \ll \min(d, k)$.

The forward transformation can therefore be represented as

$$h' = Wx + BAx.$$

The pretrained matrix W remains static, and optimization is performed on A and B using the Brahui conversational dataset. In this way, the number of parameters to be trained is minimized without compromising the pretrained multilingual model as the backbone model. The formulation of LoRA keeps the pretrained weights fixed and adds trainable low-rank matrices.

3.6 Model and Training Configuration

The mBART-50 + LoRA system is used to generate responses to Brahui questions. Various ranks of LoRA are investigated using specified training parameters, data splits, and hardware capabilities.

Table 2 Model and Training Configuration

Parameter	Configuration
Backbone model	mBART-50
Adaptation method	LoRA
Task	Brahui question-to-response generation
Dataset split (train/validation/test)	890 / 111 / 112
Maximum sequence length	128
LoRA ranks	4, 8, 16
Target module	q_proj, v_proj
LoRA dropouts	0.05
Learning rate	6×10^{-5}
Batch size	1
Gradient accumulation step	2
Weight decay	0.01
Number of epochs	10
Precision	FP16 when CUDA is available
Hardware	(NVIDIA)Tesla T4 GPU (16 GB VRAM)
Model Training time	Approximately 8 hours

International Journal of Advanced Research

Table 2 demonstrates that the backbone architecture for Brahui question-to-response generation using mBART-50 has been used with LoRA ranks of 4, 8, and 16 on the query and value projection modules. A total of 890 instances have been reserved for training, 111 for validation, and 112 for testing. The maximum length of the sequence is 128. The training process involved the use of the following parameters: the learning rate is 6×10^{-5}, the batch size is 1, the gradient accumulation is 2, the weight decay is 0.01, the dropout is 0.05, and the number of epochs is 10. FP16 precision was activated if CUDA is available.

3.7 LoRA Rank Ablation

The primary methodology used in this research is the LoRA rank ablation. There were three different ranks analyzed:

$$r \in \{4, 8, 16\}.$$

The rank defines the dimensionality of the low-rank update and hence the capacity available for task-specific learning.

Table 3 LoRA rank configurations

Configuration	Rank r	Alpha α	Target modules
LoRA-4	4	16	q-proj, v-proj
LoRA-8	8	32	q-proj, v-proj
LoRA-16	16	32	q-proj, v-proj

Three setups for the LoRA approach are shown in Table 3, with adapter ranks of 4, 8, and 16 to investigate the influence of the adapter rank on model performance. The setup uses qproj and vproj for the projection layers with a dropout value of 0.05. The calculation of the LoRA update is done as follows:

$$W' = W + \frac{\alpha}{r} BA$$

Where (W) stands for the initial weight matrix, (W') stands for the new weight matrix, (α) denotes the scaling factor, (r) denotes the LoRA rank, and

(A) and (B) denote low-rank matrices. The purpose of conducting these experiments is to find an appropriate rank that offers the right balance between all the above aspects.

3.8 Overall Experimental Architecture

The entire process methodology may be presented as one continuous pipeline, not as individual model drawings.

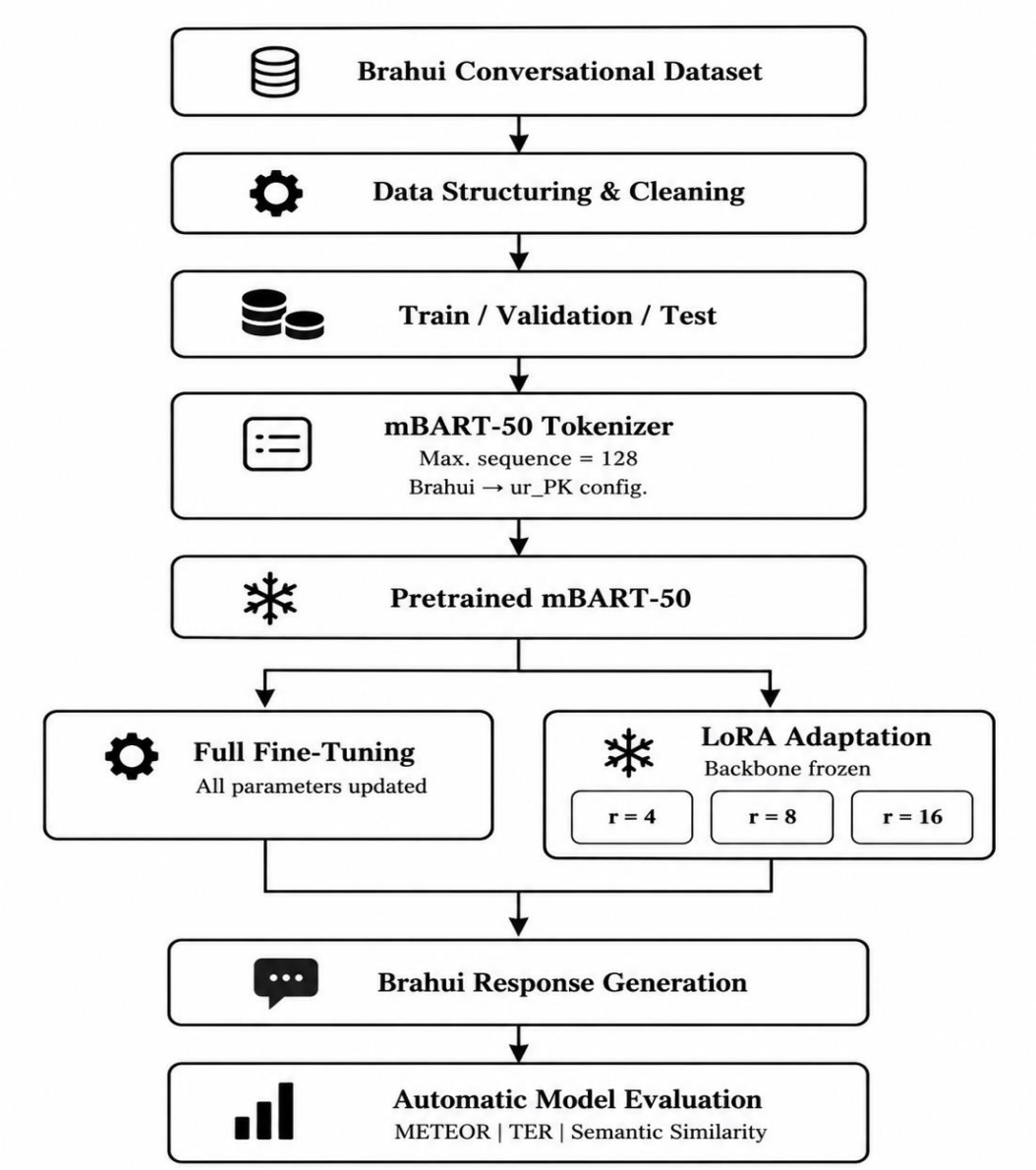


Figure 2 Overall methodology of the proposed mBART-50 and LoRA-based Brahui conversational system

Figure 2: The proposed model is based on a pipeline for Brahui response generation during conversations that begins with 1,113 pairs of questions and answers that have been cleaned, formatted, and split into training, validation, and testing sets. Texts are tokenized via the mBART-50 model with a maximum sequence length of 128 tokens. Two methods are considered: the first one, full fine-tuning method involves updating all model parameters, while the second one, LoRA

adaptation, involves freezing the backbone of the model and training low-rank adapters with ranks of 4, 8, and 16. The output of these two methods includes Brahui responses that will be compared via METEOR, TER, and semantic similarity measures.

3.9 Response Generation

Inference involves the presentation of a question in Brahui from the unseen test set to the model. The encoder forms contextual representations while the decoder outputs the answer through an autoregressive generation process. For each test question x, the generated output is

$$\hat{y} = \arg\max_{y} P_{\theta}(y \mid x).$$

The output response yˆ is then compared with the ground truth Brahui response y using METEOR, TER, semantic similarity, and perplexity. Lexical match, translation edit rate, semantic similarity, and language model perplexity are measured through these metrics.

4. Results and Discussion

4.1. Overall Performance Comparison

This section highlights the performance of the full mBART-50 fine-tuning along with LoRA with rank 4, 8, and 16 on the generation of the Brahui language. Evaluation of the models has been done through metrics such as METEOR, Translation Edit Rate (TER), and semantic similarity. In this regard, a high value of METEOR and semantic similarity denotes higher performance.

Table 4 Overall Performance Comparison of mBART-50 and LoRA Configurations

Model configuration	METEOR	TER	Semantic similarity	Overall performance
Full mBART-50 fine-tuning	78%	23%	89%	Strong
mBART-50 + LoRA, rank 4	64%	39%	85%	Moderate
mBART-50 + LoRA, rank 8	**80%**	**20%**	**89%**	**Best**
mBART-50 + LoRA, rank 16	63%	44%	82%	Weakest

The performance of the mBART-50 model with full fine-tuning and different LoRA ranks of 4, 8, and 16 is shown in Table 4. The performance of LoRA rank 8 is the best with respect to METEOR score (80), TER (20), and 89% semantic similarity. Full fine-tuning shows good performance as well, but the performance of ranks 4 and 16 is lower compared to the other two approaches.

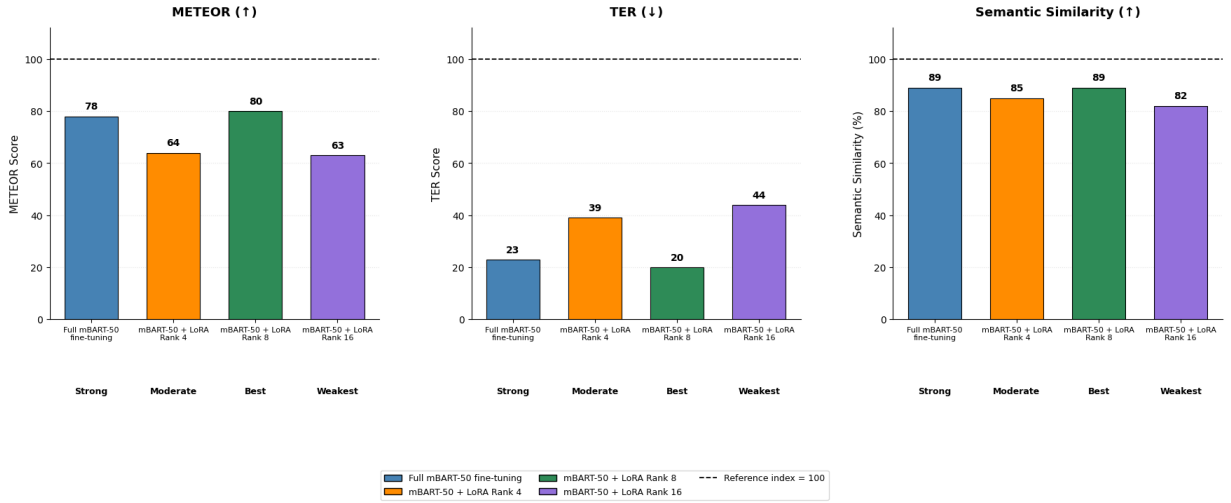


Figure 3 Overall Performance Comparison of mBART-50 and LoRA Configurations

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

Figure 3 below gives a comparison of fine-tuning of mBART-50 with LoRA of ranks 4, 8, and 16 on the basis of METEOR, TER, and semantic similarity. The best performance was exhibited by LoRA of rank 8 since it has the highest METEOR score, which is 80, the lowest TER score, which is 20, and 89% semantic similarity.

4.2. Ablation Study: Effect of LoRA Rank

In order to explore the impact of adaptation ranks on performance, three LoRA ranks were compared under the same architecture of mBART-50.

Table 5 Ablation Study of LoRA Rank

Model configuration	LoRA rank	METEOR	TER	Semantic similarity
Full mBART-50 fine-tuning	-	78%	23%	89%
mBART-50 + LoRA	4	64%	39%	85%
mBART-50 + LoRA	8	80%	20%	89%
mBART-50 + LoRA	16	63%	44%	82%

Table 5 shows a comparison between full fine-tuning of mBART-50 and LoRA of rank 4, 8, and 16 with METEOR, TER, and semantic similarity. The highest performance is shown by LoRA rank 8, which has a METEOR score of 80, a TER score of 20, and 89% semantic similarity.

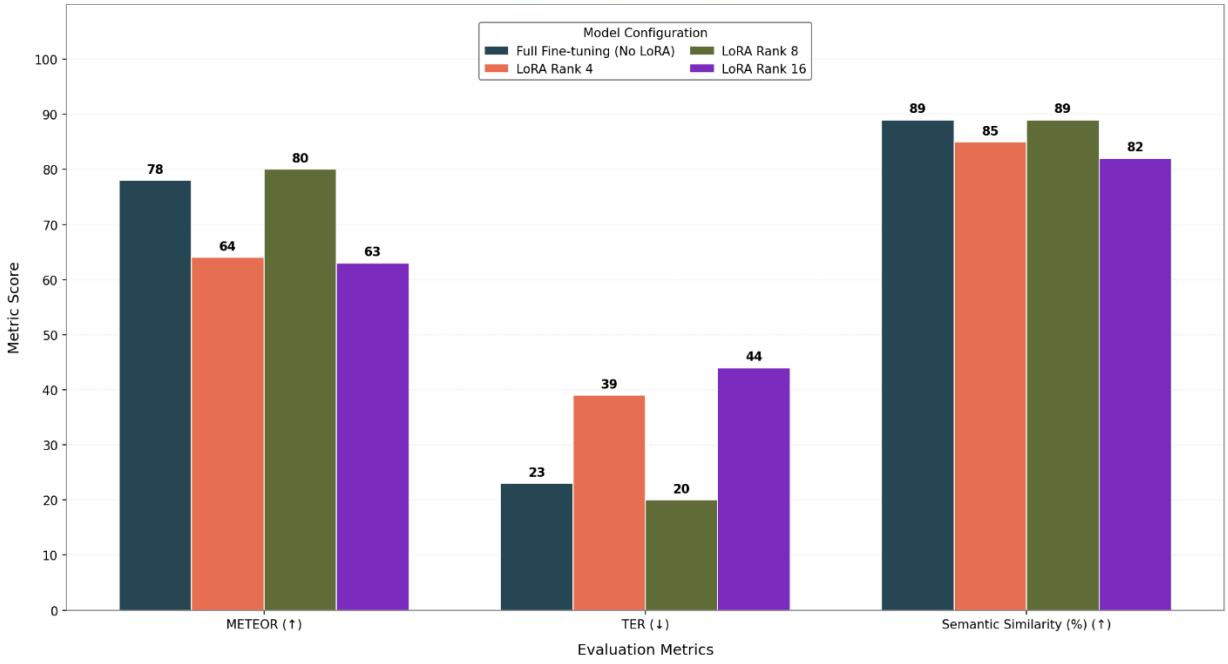


Figure 4 Ablation Study of LoRA Rank

Figure 4 shows an ablation study that tests how various LoRA Ranks perform compared to full fine-tuning on METEOR, TER, and Semantic Similarity Metrics, and LoRA Rank 8 performs on par with full fine-tuning.

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

4.3. Training and Validation Loss Analysis

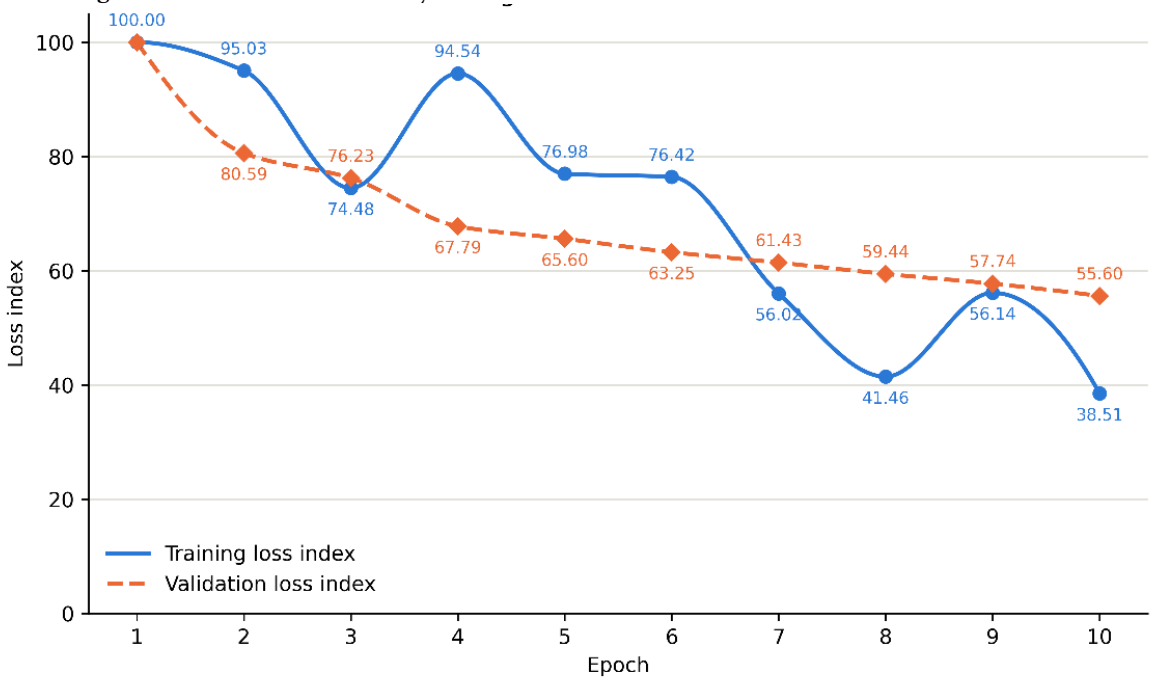


Figure 5: Training vs Validation Loss Index of the model

Figure 5, illustrates the Training and Validation Loss Index for 10 epochs, standardized such that epoch 1 is equal to 100. The training loss index is illustrated by the blue line, and the validation loss index is illustrated by the orange dashed line, where the values are provided next to their corresponding points on the graph. While both indices decrease overall, the validation loss index decreases more steadily to 55.60 (a reduction of 44%) and achieves a majority of its improvements in the first four epochs, where the decrease then slows to about two units per epoch after that. The training loss index, on the other hand, decreases more quickly to 38.51 (a reduction of 61%) but fluctuates somewhat, showing an increase in epochs 4 (94.54) and 9 (56.14). Starting from epoch 7, training loss is always lower than validation loss by about 17 units at epoch 10, indicating early overfitting.

4.4. Error Analysis

Error analysis is crucial for recognizing what kind of errors can be found in Brahui output. There were five error categories that were chosen for comparison: semantic errors, grammatical errors, lack of information, repetition errors, and irrelevancy.

Table 6 Error Analysis Categories and Reported Percentages

Error category	Full fine-tuning	LoRA rank = 4	LoRA rank = 8	LoRA rank = 16
Semantic errors	8%	15%	7%	18%
Grammatical errors	10%	18%	9%	20%
Missing information	12%	20%	10%	22%
Repetition errors	5%	9%	4%	11%

Irrelevant responses	6%	12%	5%	15%

According to Table 6, it can be observed that the optimal LoRA model architecture is Rank 8, as it gives the minimum error rates for all five different types, even beating full fine-tuning. In contrast, low capacity (Rank 4) as well as high capacity (Rank 16) give higher error rates.

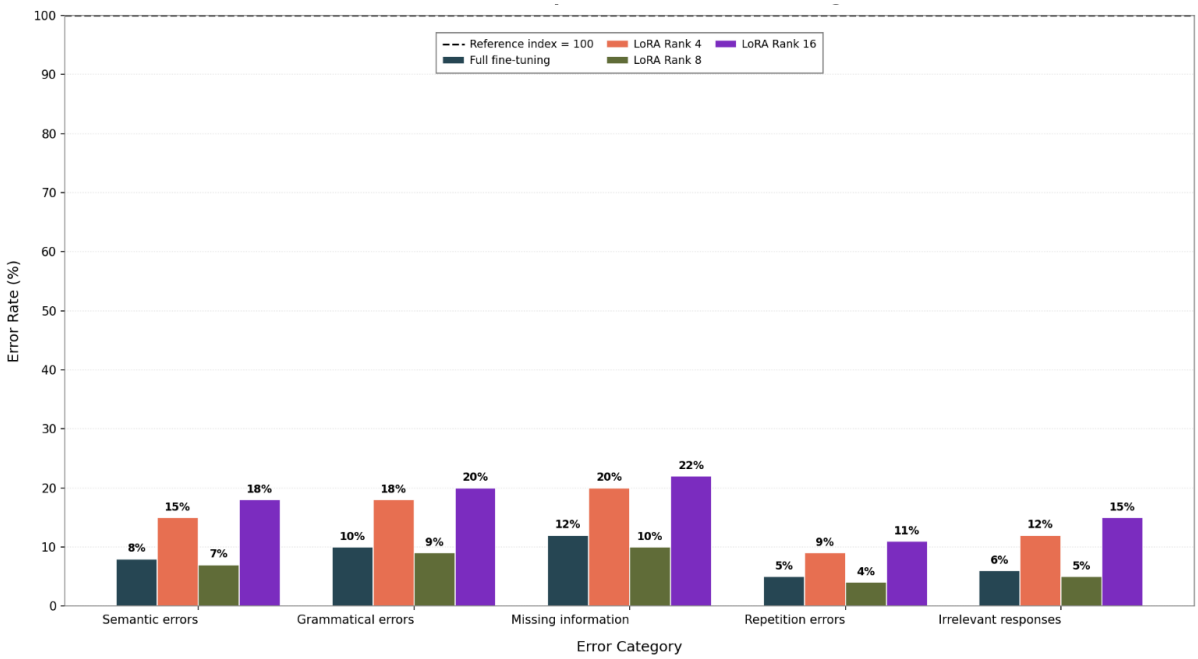


Figure 6 Translation Error Comparison Across Model Configurations

In Figure 6, the error rate of translations is compared among different model setups, where LoRA Rank 8 obtains the lowest error rates in all categories, even performing slightly better than full fine-tuning, while LoRA Rank 16 yields the highest error rates in all categories.

4.5. Implications for Low-Resource Brahui Text Generation

The results indicate that parameter-efficient adaptation with LoRA can be a successful strategy for generating text in the Brahui language. Specifically, rank 8 generated the most impressive automatic evaluation metrics results while having a semantic similarity score close to mBART-50 fine-tuning. This can be applied to the development of NLP models for languages with a limited amount of training data and computational resources. LoRA technique trains a language model with additional low-rank trainable parameters instead of adjusting all existing parameters of a model. The obtained results prove that this adaptation technique can generate high-quality output for the selected task. Nevertheless, the current comparison does not include information about training speed, memory usage, or the number of trainable parameters. Hence, there is no direct proof of the computational advantages of LoRA in this experiment. These parameters must be measured to prove the efficiency of LoRA beyond its output quality. Furthermore, automatic evaluation metrics do not provide a guarantee that the results are correct or even grammatically natural. Human evaluation from Brahui speakers will provide additional information regarding the conversational benefits of the selected metric improvement.

4.6. Limitations and Future Work

There are several limitations that have to be taken into account while analyzing the obtained results. The presented values allow making a descriptive comparison of the results; the initial values before rounding should be used for further calculations. Secondly, the current research does not include confidence intervals, statistical significance tests, or experiments using different random seeds. Thus, the improvement obtained cannot be regarded as evidence of statistically significant superiority. Future experiments should use various random seeds and proper statistical testing of the differences between full fine-tuning and LoRA settings. Thirdly, the dataset includes only 1,113 Brahui question-answer pairs, which might be insufficient for ensuring good generalization of the models. Further experiments should increase the amount of Brahui examples in the dataset and conduct evaluation of the models on independently collected data of the Brahui language.

Fourthly, the experiments were conducted with the use of the default mBART-50 tokenizer with the experimental language configuration Urdu (ur_PK), as there is no specific language option for Brahui in the selected configuration. However, sharing the script representation does not mean that the tokenization of Brahui text will be optimal. Further research should include analysis of subword fragmentation, average and median number of tokens for questions and answers, and the percentage of truncated sequences due to the 128-token limitation. Alternative preprocessing methods and tokenization approaches also need to be studied to understand if they can help in improving the representation of Brahui text. Finally, further research should include exploration of various LoRA ranks, consistent comparative experiments of the models, and measurement of computational needs, such as training time and memory consumption.

5. Conclusion

This paper presents the experimental validation of multilingual model mBART-50 fine-tuning and LoRA fine-tuning at LoRA ranks 4, 8, and 16, on a Brahui dataset of 1,113 question-answer pairs for conversation response generation. This research shows that multilingual transfer learning is feasible with Brahui as well, despite the fact that Brahui is not among the languages of the model. Full fine-tuning yielded METEOR scores of 78%, TER of 23%, and semantic similarity of 89%. At rank 8, LoRA yields the best results (METEOR score of 80%, TER of 20%, and 89% semantic similarity), with the backbone frozen and a tiny number of parameters updated. Ranks 4 and 16 produce worse results, which demonstrates that in the low-resource setting, increasing LoRA rank may not bring better quality. All results should be considered preliminary, as the dataset size is rather limited, and the experiments were run once, without significance testing. Future research would benefit from running experiments with different seeds, using a bigger Brahui dataset, analyzing tokenization, and human evaluation by native speakers.

REFERENCES

[1] L. Papa, P. Russo, I. Amerini, and L. Zhou, "A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 12, pp. 7682–7700, 2024, doi: 10.1109/TPAMI.2024.3392941.

[2] Z. Al Nazi, M. R. Hossain, and F. Al Mamun, "Evaluation of open and closed-source LLMs for low-resource language with zero-shot, few-shot, and chain-of-thought prompting," Natural Language Processing Journal, vol. 10, 2025, doi: 10.1016/j.nlp.2024.100124.

[3] B. Haddow, R. Bawden, A. V. M. Barone, J. Helcl, and A. Birch, "Survey of Low-Resource Machine Translation," Computational Linguistics, vol. 48, no. 3, pp. 673–732, 2022, doi: 10.1162/coli_a_00446.

[4] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural machine translation for low-resource languages: A survey," ACM Comput. Surv., vol. 55, no. 11, p. 229, 2023, doi: 10.1145/3567592.

[5] T. O. Tafa et al., "Machine translation performance for low-resource languages: A systematic literature review," IEEE Access, vol. 13, pp. 72486–72505, 2025, doi: 10.1109/ACCESS.2025.3562918.

[6] J. A. Mengal and M. Baloch, "The problems and future of Brahui language in 21th century," Al-Burz, vol. 10, no. 1, pp. 1–9, 2018, doi: 10.54781/abz.v10i1.78.

[7] A. Q. S. Brahui, "A research and linguistically study of verb in Brahui grammar," Al-Burz, vol. 11, no. 1, pp. 1–13, 2019, doi: 10.54781/abz.v11i1.40.

[8] Y. A. Rodeni, "Translation in Brahui literature," Al-Burz, vol. 3, no. 1, pp. 46–73, 2011, doi: 10.54781/abz.v3i1.174.

[9] A. Karakanta, J. Dehdari, and J. van Genabith, "Neural machine translation for low-resource languages without parallel corpora," Machine Translation, vol. 32, no. 1–2, pp. 167–189, Jun. 2018, doi: 10.1007/s10590-017-9203-5.

[10] R. Rubino, B. Marie, R. Dabre, A. Fujita, M. Utiyama, and E. Sumita, "Extremely low-resource neural machine translation for Asian languages," Machine Translation, vol. 34, pp. 347–382, 2020, doi: 10.1007/s10590-020-09258-6.

[11] J. Zhang et al., "Neural machine translation for low-resource languages from a Chinese-centric perspective: A survey," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 6, pp. 1–60, 2024, doi: 10.1145/3665244.

[12] A. Qorbani, R. Ramezani, A. Baraani-Dastjerdi, and A. Kazemi, "Multilingual neural machine translation for low-resource languages by twinning important nodes," Neurocomputing, vol. 634, 2025, doi: 10.1016/j.neucom.2025.129890.

[13] S. Casola, I. Lauriola, and A. Lavelli, "Pre-trained transformers: An empirical comparison," Machine Learning with Applications, vol. 9, 2022, doi: 10.1016/j.mlwa.2022.100334.

[14] L. Wang et al., "Parameter-efficient fine-tuning in large language models: A survey of methodologies," Artif. Intell. Rev., vol. 58, 2025, doi: 10.1007/s10462-025-11236-4.

[15] S. Pratap, A. R. Aranha, D. Kumar, G. Malhotra, A. P. N. Iyer, and S. S. S., "The fine art of fine-tuning: A structured review of advanced LLM fine-tuning techniques," Natural Language Processing Journal, vol. 11, 2025, doi: 10.1016/j.nlp.2025.100144.

[16] Y. Mao et al., "A survey on LoRA of large language models," Front. Comput. Sci., vol. 19, 2025, doi: 10.1007/s11704-024-40663-9.

[17] I. Sel and D. Hanbay, "Efficient adaptation: Enhancing multilingual models for low-resource language translation," Mathematics, vol. 12, no. 19, 2024, doi: 10.3390/math12193149.

[18] X. Liang, Y.-M. J. Khaw, S.-Y. Liew, T.-P. Tan, and D. Qin, "Toward low-resource languages machine translation: A language-specific fine-tuning with LoRA for specialized large language models," IEEE Access, vol. 13, pp. 46616–46626, 2025, doi: 10.1109/ACCESS.2025.3549795.

[19] W. Zhu, Y. Li, C. Liu, and Y. Zhao, "FLAT: Fusing layer representations for more efficient transfer learning in NLP," Neural Networks, vol. 179, 2024, doi: 10.1016/j.neunet.2024.106631.

[20] J. Gu, J. Yuan, J. Cai, X. Zhou, and L. Fan, "La-LoRA: Parameter-efficient fine-tuning with layer-wise adaptive low-rank adaptation," Neural Networks, vol. 194, 2025, doi: 10.1016/j.neunet.2025.108095.

[21] S. M. Singh and T. D. Singh, "Low resource machine translation of English–Manipuri: A semi-supervised approach," Expert Syst. Appl., vol. 209, 2022, doi: 10.1016/j.eswa.2022.118187.

[22] N. Ahmed, M. A. Khuhro, and A. A. Laghari, "Bridging the Dravidian Language Gap: A Back Translation Enhanced Transformer Framework for Brahui-English Neural Machine Translation," Comput. Intell., vol. 42, no. 6, Dec. 2026, doi: 10.1111/coin.70305.

[23] A. V Hujon, T. D. Singh, and K. Amitab, "Neural machine translation systems for English to Khasi: A case study of an Austroasiatic language," Expert Syst. Appl., vol. 238, 2024, doi: 10.1016/j.eswa.2023.121813.

[24] L. S. Meetei, T. D. Singh, and S. Bandyopadhyay, "An empirical study of a novel multimodal dataset for low-resource machine translation," Knowl. Inf. Syst., vol. 66, no. 11, pp. 7031–7055, 2024, doi: 10.1007/s10115-024-02087-6.

[25] L. S. Meetei, T. D. Singh, and S. Bandyopadhyay, "Exploiting multiple correlated modalities can enhance low-resource machine translation quality," Multimed. Tools Appl., vol. 83, pp. 13137–13157, 2024, doi: 10.1007/s11042-023-15721-2.

[26] C. Mi and S. Xie, "Language relatedness evaluation for multilingual neural machine translation," Neurocomputing, vol. 570, 2024, doi: 10.1016/j.neucom.2023.127115.

[27] M. Munaf, H. Afzal, K. Mahmood, and N. Iltaf, "Low resource summarization using pre-trained language models," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 23, no. 10, pp. 1–19, 2024, doi: 10.1145/3675780.

[28] Y. Wu, H. Li, L. Zhang, C. Dong, Q. Huang, and S. Wan, "Joint intent detection model for task-oriented human-computer dialogue system using asynchronous training," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 5, pp. 1–17, 2022, doi: 10.1145/3558096.

[29] Y. Mo, J. Yoo, and S. Kang, "Parameter-efficient fine-tuning method for task-oriented dialogue systems," Mathematics, vol. 11, no. 14, 2023, doi: 10.3390/math11143048.

[30] N. Ahmed, M. Ahmed Khuhro, M. Ali Dootio, and S. Pakistan, "BUILDING AND ANALYZING A BRAHUI TEXT CORPUS: APPLYING DTM AND TF-IDF TECHNIQUES," Jan. 2024. Accessed: Mar. 05, 2026. [Online]. Available: https://www.researchgate.net/profile/Naseer-Ahmed-44/publication/384143614_BUILDING_AND_ANALYZING_A_BRAHUI_TEXT_CORPUS_APPLYING_DTM_AND_TF-IDF_TECHNIQUES

[31] B. Li, H. Fei, F. Su, F. Li, and D. Ji, "Integrating discourse features and response assessment for advancing empathetic dialogue," Inf. Process. Manag., vol. 61, no. 5, 2024, doi: 10.1016/j.ipm.2024.103803.

[32] S. Wu, J. Chen, J. Yu, and W. Zhou, "Chat with one voice: Mitigating semantic disparity across languages for multilingual open-domain dialogue response generation systems," Expert Syst. Appl., vol. 298, 2026, doi: 10.1016/j.eswa.2025.129521.

[33] N. Ding et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," Nat. Mach. Intell., vol. 5, no. 3, pp. 220–235, 2023, doi: 10.1038/s42256-023-00626-4.

[34] N. Ahmed, M. Gul Bizanjo, A. Khan, S. Gull, and S. khaliq, "Analysis of Textual Feedback of Students for Course Evaluation in Universities Through Machine Learning Algorithms. Analysis of Textual Feedback of Students for Course Evaluation in Universities Through Machine Learning Algorithms," 2024.

[35] N. Ahmed, M. A. Khouro, A. Khan, M. Dawood, M. A. Dootio, and N. U. Jan, "Student textual feedback sentiment analysis using machine learning techniques to improve the quality of education," Pakistan Journal of Engineering, Technology & Science, vol. 11, no. 2, pp. 32–40, Dec. 2023, doi: 10.22555/pjets.v11i2.1039.

[36] A. K. Biswas, M. Fahim, M. T. H. Fuad, A. M. R. Mazumder, A. A. Ali, and A. K. M. Mahbubur Rahman, "Training-free layer selection for partial fine-tuning of language models," Inf. Sci. (N. Y)., vol. 740, 2026, doi: 10.1016/j.ins.2026.123204.

[37] S. Kim, H. Yang, Y. Kim, Y. Hong, and E. Park, "Hydra: Multi-head low-rank adaptation for parameter efficient fine-tuning," Neural Networks, vol. 178, 2024, doi: 10.1016/j.neunet.2024.106414.

[38] S. Cao, Y. Zhang, X. Zhang, J. Wang, and X. Zhou, "EfficientLoRA: Rethinking the efficiency of low-rank adaptation in pre-trained language models," Neural Networks, vol. 200, 2026, doi: 10.1016/j.neunet.2026.108853.

[39] J.-C. Zhang, Y.-J. Xiong, C.-M. Xia, D.-H. Zhu, and H.-J. Zhan, "LoRA2: Multi-scale low-rank approximations for fine-tuning large language models," Neurocomputing, vol. 650, 2025, doi: 10.1016/j.neucom.2025.130859.

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

[41] S. Gull, N. Ahmed, and S. Ahmed Sheikh, "Liberal Journal of Language & Literature Review Liberal Journal of Language & Literature Review Liberal Journal of Language & Literature Review Deep Learning-Driven OCR System for Brahui Printed Text: Bridging the Digital Gap in Low-Resource Language Processing Liberal Journal of Language & Literature Review", [Online]. Available: https://llrjournal.com/index.php/11

[42] S. Chauhan, S. Saxena, and P. Daniel, "Improved unsupervised neural machine translation with semantically weighted back translation for morphologically rich and low resource languages," Neural Process. Lett., vol. 54, pp. 1707–1726, 2022, doi: 10.1007/s11063-021-10702-8.

[43] K. Bhuvaneswari and M. Varalakshmi, "Efficient incremental training using a novel NMT-SMT hybrid framework for translation of low-resource languages," Front. Artif. Intell., vol. 7, 2024, doi: 10.3389/frai.2024.1381290.

[44] S. Zhu, L. Pan, and D. Xiong, "FEDS-ICL: Enhancing translation ability and efficiency of large language model by optimizing demonstration selection," Inf. Process. Manag., vol. 61, no. 5, 2024, doi: 10.1016/j.ipm.2024.103825.