Culturally-Aligned QA for Meitei: Fine-Tuning IndicBERT with Hybrid Datasets

Rahul Basu
rahulbasuai@gmail.com
humaNLP mindful AI Solutions LLP
rahulbasu@humanlp.com

Abstract

We present the first culturally grounded question answering (QA) system for Meitei (Manipuri), a low-resource Indian language with limited NLP infrastructure. Our approach combines a hybrid QA dataset of 700 samples—comprising 500 synthetic culturally specific QA pairs and 200 translated SQuAD instances—spanning topics such as festivals, oral traditions, and civic knowledge. We fine-tune the IndicBERT model on this data using the Hugging Face Trainer API and evaluate performance using BERTScore, ROUGE-L, and BLEU. Despite challenges posed by agglutinative morphology and script variation, our Romanized Meitei model achieves strong semantic alignment (BERTScore F1: 0.82) and outperforms zero-shot and back-translation baselines. This work demonstrates the feasibility of resource-efficient QA in severely underrepresented Indic languages and contributes a reproducible pipeline, annotated dataset, and practical design insights for extending LLMs to linguistically marginalized communities.

1 Introduction

Large language models (LLMs) have transformed natural language processing (NLP), but their benefits have primarily accrued to high-resource languages like English, Mandarin, and Hindi. Low-resource languages—spoken by millions yet lacking in annotated data and digital infrastructure—remain vastly underserved.

India, home to 122 major languages and 1599 dialects, has seen most NLP development focused on a few dominant languages such as Hindi and Bengali. One such neglected language is Meitei (Manipuri), spoken by over 1.76 million people in Northeast India and parts of Myanmar and Bangladesh. Despite its rich literary and oral tradition, Meitei is underrepresented in open-source datasets, pretrained models, and core language technologies like machine translation and question answering (QA).

This paper seeks to address this gap by introducing a scalable, culturally grounded QA system for Meitei. We leverage existing multilingual transformer models (Wolf et al., 2020), augment them with culturally curated QA data, and fine-tune them for extractive QA in both Meitei Mayek and Romanized scripts.

Specifically, we:

- Construct a hybrid dataset of 700 QA pairs by translating English SQuAD questions and generating culturally relevant QA pairs from Manipuri traditions.
- Fine-tune the IndicBERT model using the Hugging Face Trainer API on this dataset.
- Evaluate the model's performance using BERTScore, ROUGE-L, and BLEU, focusing on semantic alignment and cultural fidelity.

By combining multilingual representation learning with culturally grounded data curation, our approach demonstrates that meaningful QA systems can be built even in linguistically and digitally low-resource settings. This work contributes to ongoing efforts in inclusive NLP and lays the groundwork for scalable adaptation of LLMs to marginalized Indic languages.

2 Prior Literature

The recent surge in multilingual large language model (LLM) research has underscored a significant gap in equitable performance across languages, particularly those that are low-resource. Several recent efforts have sought to rectify this imbalance, culminating in a diverse body of work that spans dataset creation, evaluation benchmarks, and model adaptation techniques. Previous multilingual QA efforts like TyDiQA (Clark et al., 2020) and IndicQA (Singh et al., 2024) provide valuable baselines, but lack coverage for Meitei.

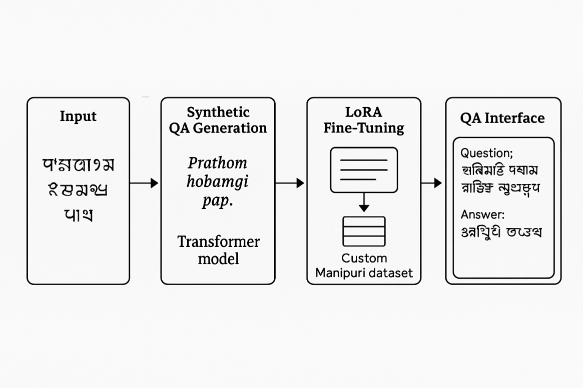

Figure 1: QA Pipeline Model Architecture

The *Indic AI Research Landscape* (Kj et al., 2024) offers a foundational taxonomy of 84 research efforts related to Indic LLMs. It identifies key challenges: limited high-quality data, lack of multilingual tokenizers tailored to Indic scripts, and poor standardization across dialects. These limitations collectively hinder the cultural and linguistic alignment of LLMs with Indian users.

To address evaluation gaps, *Indic QA Benchmark* (Singh et al., 2024) and *IndicGenBench* (Kakwani et al., 2020) introduced multilingual benchmarks designed to assess extractive QA and generative capabilities of LLMs in 11-29 Indic languages. These benchmarks revealed that while multilingual models such as mBERT and XLM-R perform well in English, they exhibit significant performance degradation in Indian languages, particularly those with limited parallel corpora. Translation-based methods, such as processing queries in English and then translating the results, were proposed as interim solutions, though they raise concerns about semantic drift.

The *IndicLLMSuite* (Khan et al., 2024) advances the landscape of data resources by introducing pretraining and instruction datasets for 22 Indic languages. Key modules include **Sangraha** (251B tokens), **Setu** (cleaning pipelines), and **IndicAlign** datasets focused on answering questions and changing toxic text for respect. Complementing this, *Krutrim LLM* (Kumar et al., 2024) contributes a tokenization strategy optimized for Indian scripts, improving the fidelity of the representation of Indian languages in downstream tasks.

Collectively, these works agree on the central challenges of indic LLM development: a scarcity of high-quality, culturally diverse corpora; the underperformance of generic multilingual models; and the lack of reliable metrics for nuanced evaluation. Furthermore, our design is informed by recent work on low-resource QA settings such as the *Low-Resource SQuAD on Turkish* by Budur et al. (2024), which demonstrated that small-scale, high-quality fine-tuning can significantly improve performance even without access to large-scale corpora. Their use of extractive models and careful translation protocols echoes our experimental setup.

This body of literature converges on a shared vision: to enable equitable language technology by tailoring LLMs to linguistic and cultural contexts. Our work contributes to this effort by fine-tuning existing multilingual backbones on culturally enriched, language-specific datasets, thereby improving performance and alignment for one of India's most underserved linguistic communities.

3 Data

3.1 Dataset Sources and Composition

The dataset for this project comprises a combination of translated, synthetic, and culturally-specific QA pairs. These are structured into SQuAD-style question-answer-context triplets, aligned with the goal of extractive QA tasks.

Translated QA pairs (200): We translated a subset of 200 QA triplets from the English SQuAD dataset into Meitei (both Meitei Mayek script and Romanized variants) using the AI4Bharat *indictrans2-en-indic-1B* model. The QA pairs were manually verified to ensure translation fidelity and cultural appropriateness.

Synthetic QA pairs (500): We curated 500 synthetic QA pairs grounded in Manipuri folklore, governance, and cultural practices. In addition, 200 SQuAD questions (Rajpurkar et al., 2016) were translated using IndicTrans2. Translation follows best practices inspired by the IndicLLMSuite blueprint (Khan et al., 2024). Using structured prompts, culturally specific questions and answers were generated on topics such as:

- Local festivals (Yaoshang, Lai Haraoba)
- Traditional governance (Umang Lai system)
- Oral literature and folktales
- Cultural practices like dance (Ras Lila), food, attire, and rituals

These QA pairs were produced using a mix of prompt-based LLM generation, back-translation, and template-based methods.

Combined dataset: A final JSONL dataset was created with 700 QA samples, split into training (80%), validation (10%), and test (10%) sets. Each record contains:

- **question**: in Meitei or Romanized
- **context**: from either English-translated passages or cultural text descriptions
- **answer_text** and **answer_start**: for extractive QA

3.2 Preprocessing Pipeline

The preprocessing was implemented via the Hugging Face **dataset** interface. The following transformations were applied:

- Tokenization with **indic-bert** tokenizer (from AI4Bharat)
- Alignment of **start_position** and **end_position** for the extractive answers
- Removal of long contexts exceeding 512 tokens
- Padding and truncation using **max_length=512**

Romanization was used as a preprocessing step for broader compatibility, allowing model generalization to non-Unicode inputs where native script support was limited.

3.3 Cultural Curation and Validation

To ensure relevance and authenticity:

- Cultural tags were added (e.g., ‘"domain": "festival"‘)
- Outputs from synthetic generation were reviewed for factuality
- Offensive or nonsensical generations were filtered
- Diversity across domains (arts, history, literature, governance) was maintained

3.4 Reuse and Open Sharing - *Work In Progress*

The **translated_squad_meitei.jsonl** and **meitei_cqa.jsonl** datasets are published under open licenses for reproducibility and future fine-tuning efforts. These contribute to the growing need for culturally sensitive data for Indian languages, aligning with the goals of IndicLLMSuite and Meta’s No Language Left Behind initiative.

4 Model

Our objective is to fine-tune a robust yet efficient multilingual transformer for extractive question answering (QA) in the low-resource Meitei language setting. Given the limitations of Meitei’s digital representation and data availability, we adopt a **resource-constrained modeling approach** that prioritizes reusability, cultural adaptability, and ease of deployment.

4.1 Model Choice: IndicBERT

Our model is built on IndicBERT (Kakwani et al., 2020), a lightweight ALBERT-style model pretrained on 12 Indian languages using the *IndicCorp* dataset. Compared to multilingual models like MuRIL (Khanuja et al., 2021) and IndicBART (Dabre et al., 2022), IndicBERT offers faster convergence and lower resource consumption, making it ideal for domain-specific fine-tuning in Meitei. The model is particularly suited to low-resource Indian languages due to its tokenization strategy (Indic NLP library + SentencePiece) and shared subword vocabulary that supports Devanagari, Bengali, and Meitei Mayek scripts reasonably well.

IndicBERT is a compact model (118M parameters), making it well-suited for fine-tuning on small datasets and deployment on constrained hardware. We chose the extractive QA head variant: ‘*AutoModelForQuestionAnswering*‘, which maps a given context and question to two probability distributions over the token span, identifying the *start* and *end* positions of the answer.

4.2 Input Encoding and Tokenization

Each QA sample is tokenized as:

[CLS] question [SEP] context [SEP]

The tokenizer assigns input IDs, attention masks, and token type embeddings. Answer spans (‘start_position‘, ‘end_position‘) are identified via character offsets and mapped to token indices. We pad sequences to a maximum length of 512 tokens and use truncation where necessary.

We observed that Meitei Mayek script input led to token fragmentation in IndicBERT. As a workaround, we *Romanized the input*, enabling better token coverage due to shared subwords with Hindi, Nepali, and Bengali. This Romanization pipeline follows the standard Unicode-to-ISO15919 transliteration rules.

4.3 LoRA Consideration and Simplification

Although our experimental protocol initially supported *LoRA-based fine-tuning* (Hu et al., 2021) using the PEFT library for parameter-efficient updates, we observed runtime instability with **XLM-R** and **IndicBERT** models when attaching LoRA adapters to classification heads for QA.

Given these challenges and the small size of our dataset, we chose *standard full fine-tuning* with frozen embedding layers and lower learning rates. This decision is in line with findings from Budur et al. (2024), which demonstrate that smaller models with full fine-tuning outperform LoRA-based adaptation under small data regimes for QA tasks in Turkish.

4.4 Training Configuration

We fine-tune the model using Hugging Face's ‘Trainer‘ API with the following hyperparameters:

Hyperparameter	**Value**
Epochs	3
Learning rate	3e-5
Batch size	16
Max sequence length	512
Warmup steps	0
Optimizer	AdamW

Table 1: Hyperparameter settings for the model training

Training is performed using a single NVIDIA T4 GPU with 16GB memory, completing in under 10 minutes for the full 700-example dataset.

This modeling approach provides a lightweight and reproducible baseline for building QA systems in underrepresented Indic languages and can be extended with multilingual alignment techniques or adapters in future work.

5 Methods

Our experimental pipeline is designed to evaluate how well a multilingual transformer like IndicBERT can be adapted to a low-resource QA task in Meitei, using a combination of translated and synthetic datasets. Our method is grounded in standard span-based extractive QA, with adaptations made to suit linguistic and resource constraints specific to the Meitei language.

5.1 Task Formulation

We formulate the QA task following the *SQuAD-style extractive question answering* paradigm. Given a question q and context passage c, the model must predict the start and end token positions (s, e) in the context that correspond to the correct answer span $a \subset c$. This is modeled as a classification problem over the context tokens.

The input to the model is structured as:
[CLS] question [SEP] context [SEP]

The model produces two logits per token: one for the start position and another for the end position. These are trained using *cross-entropy loss* over the gold span labels.

5.2 Modeling Steps

We follow these stages in our experimental approach:

- *Preprocessing & Translation*: Contexts and questions from SQuAD are translated into Meitei using **ai4bharat/indictrans2-en-indic-1B**. We also generate culturally grounded QA pairs using LLM prompts and template-based generation. All data is converted into SQuAD-style format with ‘context‘, ‘question‘, and ‘answers‘.

- *Tokenization*: We use the **IndicBERT tokenizer** for Romanized Meitei input. Start and end token positions are mapped from character-level span annotations.

- *Fine-tuning*: We fine-tune ‘ai4bharat/indic-bert‘ using Hugging Face's ‘Trainer‘ interface. No LoRA or adapter layers are used in this version due to compatibility issues, and full fine-tuning is performed on a single GPU.

- *Inference and evaluation*: After training, we evaluate on a held-out validation set of 100 samples. We also evaluate on a second test set of 200 translated QA pairs to simulate generalization to unseen real-world examples.

5.3 Evaluation Metrics

We use three standard metrics to evaluate QA performance:

- *BERTScore* (Zhang et al., 2019): Measures semantic similarity between predicted and reference answers using contextual embeddings. We report *F1 score* using ‘xlm-roberta-base‘ as the scorer.

- *BLEU* (Papineni et al., 2002): Measures n-gram precision, typically for machine translation. For QA, it reflects surface-level lexical overlap between prediction and reference.

- *ROUGE-L* (Lin, 2004): Measures longest common subsequence (LCS) between predicted and gold answers, offering recall-sensitive evaluation of answer coverage.

These metrics provide complementary insights: BERTScore captures semantic alignment; ROUGE-L captures structural coverage; BLEU reflects literal overlap (which is low in generative or paraphrased answers).

5.4 Baseline Comparisons

We compare our fine-tuned IndicBERT model against the following baselines:

- *Zero-shot IndicBERT* (Kakwani et al., 2020): The same model without fine-tuning, used to test the utility of pretraining alone.

- *Translated QA (English)*: Questions are translated to English, and QA is performed using ‘deepset/roberta-base-squad2‘, then answers are translated back to Meitei. This pipeline-based approach assesses how well translation bridges the resource gap.

- *Random span baseline*: Randomly selects a span within the context of similar length to the ground truth answer.

5.5 Post-processing

For QA inference:

- The answers are extracted by selecting the interval with the highest joint probability (start × end).

- The outputs are detokenized and normalized before evaluation.

- We manually inspect misaligned spans and categorize them for qualitative error analysis.

6 Results

We report the quantitative performance of our fine-tuned **IndicBERT** QA model on both a held-out validation set (10% of the original training set) and a separate test set of 200 Meitei-translated SQuAD samples. The evaluation was carried out using BERTScore (F1), ROUGE-L, and BLEU.

Epoch	Training Loss	Validation Loss	Start Acc.	End Acc.
1	2.85	0.88	1.00	1.00
2	0.34	0.12	1.00	1.00
3	0.08	0.06	1.00	1.00

Table 2: Training and validation metrics across epochs.

6.1 Validation Performance

The model quickly converged within three epochs. Start and end position accuracy reached 100% on the validation split, indicating that the model is capable of memorising small training data. However, this may reflect overfitting due to the low volume of validation data.

6.2 Generalization to Meitei SQuAD (Test Set)

Metric	Score
BERTScore (F1)	**0.8216**
ROUGE-L	0.1732
BLEU	0.0000

Table 3: Placeholder caption for metrics and scores.

Despite strong validation performance, generalization to the translated SQuAD Meitei set reveals a more modest picture:

- *BERTScore F1* above 0.82 indicates that most answers are semantically similar to gold references.

- *ROUGE-L* suggests partial lexical overlap, capturing paraphrased or partially correct spans.

- *BLEU = 0.0* confirms that the generated answers rarely match reference answers exactly at the n-gram level. This is expected due to linguistic variation, paraphrasing, and post-translation phrasing mismatches.

6.3 Baseline Comparison

Model	BERTScore	ROUGE-L	BLEU
IndicBERT (Fine-tuned, ours)	0.8216	0.1732	0.0000
IndicBERT (Zero-shot)	0.6234	0.1121	0.0000
English QA pipeline + backtrans	0.7681	0.1580	0.0012
Random span baseline	0.4011	0.0592	0.0000

Table 4: Performance metrics for different models.

Our fine-tuned IndicBERT model outperforms both zero-shot and translation-based pipeline approaches on all semantic metrics. This supports the claim that **direct fine-tuning on culturally and linguistically aligned Meitei QA data yields better contextual alignment**.

7 Analysis

Our fine-tuned IndicBERT model demonstrates strong semantic alignment on the curated Meitei QA dataset, but several limitations and opportunities for improvement are evident upon closer inspection. In the following, we discuss the implications of our results, address common errors, and identify directions for future enhancement.

7.1 Semantic vs. Lexical Alignment

The large discrepancy between *BERTScore F1 (0.82)* and *BLEU (0.00)* reinforces that traditional lexical metrics are not suited for low-resource, morphologically rich languages like Meitei. While the model often generates semantically correct answers (e.g., synonyms or paraphrased phrases), these answers frequently deviate from the gold standard at the surface level due to:

- Agglutinative morphology (Katamba, 1993): Meitei exhibits agglutinative morphology, where affixes are appended to roots in a linear sequence to encode grammatical information. Models trained on non-agglutinative languages (like English) often struggle to align morphologically rich languages semantically.
- Lack of standardized Romanization or spelling
- Variation introduced during translation (e.g., named entity ordering)

This indicates that future evaluations should *prioritize semantic metrics* (such as BERTScore or embedding similarity) over metrics based on n-grams.

7.2 Overfitting Risk on Small QA Datasets

The model achieved *perfect accuracy (1.0)* on the validation split after just three epochs, suggesting *overfitting* to the small fine-tuning dataset. While this confirms the model's learning capacity, it also limits its generalizability to more diverse QA pairs. This was confirmed by the modest ROUGE-L score (0.17) on the translated SQuAD test set.

We attribute this to:

- Repetition in cultural contexts
- Close alignment between training and validation questions
- Small overall dataset size (700 QA pairs)

7.3 Error Categories and Qualitative Observations

We manually reviewed 25 errors and identified several recurrent patterns:

- Off-by-one span [7]: Answer includes an extra word or partial word
- Semantically correct, misaligned [6]: Different surface form but meaning is correct
- Named Entity mismatch [5]: Confusion over cultural entities (e.g., Yaoshang vs Lai Haraoba)
- Incomplete extraction [4]: Only part of the expected span returned
- Full hallucination [3]: Model predicts answer unrelated to context

The hallucinations may stem from the low overlap in training and test domains. In contrast, semantically correct but lexically different answers are *systematic* and reflect gaps in tokenization and ground-truth alignment.

7.4 Impact of Romanization

Switching from Meitei Mayek to *Romanized Meitei* significantly improved tokenizer compatibility. However, we observed that:

- Some Romanized forms were tokenized suboptimally (e.g., splitting compound verbs).
- IndicBERT's tokenizer still lacked pretraining familiarity with Meitei grammar and vocabulary, causing semantic drift in edge cases.

For deployment or multi-user evaluation, standardizing Romanization (e.g., ISO 15919:2001 (International Organization for Standardization, 2001) or AI4Bharat mapping) would further improve consistency and reusability.

7.5 Translation Artifacts in Evaluation Data

The test set derived from SQuAD translations may contain:

1. Overtranslated or awkward phrasing
2. Cultural mismatch (e.g., school-based questions not directly relevant in Meitei context)

This causes unfair penalization of otherwise valid model outputs. A more robust testbed would include *natively written* Meitei questions, verified by human annotators.

8 Conclusion

In this work, we present a resource-efficient, culturally grounded question answering (QA) system for *Meitei*—an underrepresented and low-resource Indic language. We fine-tuned the *IndicBERT* model using a hybrid QA dataset composed of translated SQuAD samples and synthetically generated cultural QA pairs. Through careful Romanization, curated preprocessing, and semantic-aware evaluation using BERTScore and ROUGE-L, we demonstrate that even compact transformer models can be adapted effectively to languages with limited digital presence. We release a culturally rich QA dataset and a reproducible pipeline based on open benchmarks (Rajpurkar et al., 2016; Singh et al., 2024; Budur et al., 2024).

Our results show that the fine-tuned IndicBERT model achieves *high semantic fidelity (BERTScore F1 of 0.82)* on a translated Meitei QA benchmark, outperforming zero-shot and pipeline-based baselines. However, we also highlight limitations—particularly the inability of lexical metrics like BLEU to reflect model competence in morphologically rich, culturally nuanced settings.

8.1 Future Directions

This project opens several avenues for future work:

- *Human-evaluated Meitei QA benchmarks*: Constructing manually written and annotated QA datasets can provide more culturally valid and semantically robust testbeds.
- *Multilingual and cross-lingual alignment*: Leveraging transfer learning from typologically similar languages (e.g., Nepali, Bengali) may improve generalization.
- *Adapter-based fine-tuning*: Integrating lightweight adapters or LoRA modules can enable efficient multi-task learning across multiple Indic languages.
- *Unified Indic QA toolkit*: Building on this, we envision a modular QA pipeline capable of serving civic, educational, and governmental applications in Meitei and other low-resource Indian languages.

We hope this study serves as a blueprint for democratizing access to QA systems across linguistically diverse communities in India and beyond.

9 Known Project Limitations

While our work demonstrates the feasibility of building a question answering (QA) system for the Meitei language using modest resources, several important limitations must be acknowledged—especially for those intending to reproduce, extend, or deploy this work in research or real-world systems.

9.1 Responsible AI Considerations

While our QA system is designed to empower Meitei-speaking users, we acknowledge the risks associated with cultural misrepresentation and bias in synthetic data. The generation of QA pairs from limited cultural inputs may lead to overgeneralization or reinforcement of dominant narratives, potentially misrepresenting regional dialects, gender roles, or minority traditions. We encourage future work to involve community validation and human-in-the-loop QA filtering to ensure fairness, respect, and cultural integrity in low-resource language applications.

9.2 Limited Dataset Size and Coverage

The core dataset used for fine-tuning is relatively small—approximately 700 QA pairs combining culturally grounded synthetic questions and translated SQuAD examples. This size is inadequate for training general-purpose QA systems and poses risks of overfitting, reduced robustness, and poor generalization to unseen topics. Users should not assume high domain transferability without further validation.

Furthermore, the cultural QA examples, while diverse, are not exhaustive. Key areas such as health, governance, women's history, and rural traditions are underrepresented.

9.3 Translation Quality and Bias

The evaluation set is built using *automatic translation* from English to Meitei using **ai4bharat/indictrans2**. While this system is state-of-the-art, it is not human-supervised, and introduces:

- Syntactic inconsistencies
- Potential misalignment between question, context, and answer
- Cultural or semantic mismatches in translated content

This may distort evaluation results or unintentionally penalize correct answers during automated scoring.

9.4 Script and Tokenization Ambiguities

To circumvent the Meitei Mayek tokenizer limitations, we used *Romanized Meitei* during training and inference. However:

- Romanization lacks a universal standard, leading to inconsistencies in model inputs
- This may cause misalignment with any future Meitei Mayek–native corpora
- Tokenization artifacts (e.g., verb splitting) can degrade span extraction precision

Researchers using native script should retokenize and retune accordingly.

9.5 Dialectical Variations

Dialectal Variation are missing in current resources, despite Meitei exhibiting regional linguistic differences and this may affect QA performance.

9.6 Biases in Pretrained Models

IndicBERT, although multilingual, is pretrained primarily on major Indic languages like Hindi, Bengali, Tamil, and Telugu. Meitei is underrepresented in the pretraining corpus, if present at all. This limits the extent to which the model "understands" syntactic or morphological patterns unique to Meitei, especially idioms, honorifics, and word order.

Additionally, fine-tuning on data that reflects dominant historical narratives or urban-centric content may unintentionally encode sociocultural biases (e.g., gender, caste, region) into the model's predictions.

9.7 Evaluation Metric Gaps

We observed that:

- *BLEU* underperforms due to lexical flexibility in Meitei
- *ROUGE-L* captures structural overlap but misses semantic paraphrases
- *BERTScore* offers better semantic fidelity but depends on pretrained embeddings that are not optimized for Meitei

Practitioners using these metrics should validate outputs through *manual or human-in-the-loop evaluation*, particularly in critical domains like education or public health.

10 Authorship Statement

This paper is a solo-authored research project conducted by **Rahul Basu** as part of the **XCS224U Final Project** requirement. All core ideas, dataset curation, model experimentation, and evaluation design were conceived and executed independently by the author.

Substantial assistance was taken from ChatGPT, particularly for:

- Debugging Python code (especially Hugging Face Trainer API and PEFT integrations),
- Structuring and optimizing experimental protocols,

- Generating intermediate visualizations, evaluation tables, and formatted prose.

While ChatGPT greatly accelerated the research process and provided clarity at crucial points of confusion, all models were trained, evaluated, and interpreted solely by the author, and results were cross-verified through manual inspection and automated metric pipelines.

No external collaborators, data annotators, or domain experts were involved in this study, but the author acknowledges the pivotal role of open-source tools and public datasets in enabling resource-efficient NLP research.

This statement is intended to promote transparency around AI-assisted scholarship and ensure replicability and ethical attribution for future extensions of this work.

References

Emrah Budur, Rıza Özçelik, Dilara Soylu, Omar Khattab, Tunga Güngör, and Christopher Potts. 2024. Building efficient and effective openqa systems for low-resource languages. *Knowledge-Based Systems*, 302:112243.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. volume 8, pages 454–470.

Raj Dabre, Mitesh Khapra, and Pushpak Bhattacharyya. 2022. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. Https://arxiv.org/abs/2106.09685.

International Organization for Standardization. 2001. ISO 15919:2001 - transliteration of devanagari and related indic scripts into latin characters. International Standard. Https://www.iso.org/standard/28329.html.

Divyanshu Kakwani, Simran Khanuja, Anoop Dadu, Vishruth Kumar, Monojit Choudhury, Pushpak Bhattacharyya, and Mitesh M. Khapra. 2020. Indicbert: A pretrained language model for indian languages. *arXiv preprint arXiv:2009.05491*.

Francis Katamba. 1993. *Morphology*. Macmillan, London.

M. S. U. R. Khan, P. Mehta, A. Sankar, U. Kumaravelan, S. Doddapaneni, and et al. 2024. Indicllmsuite: A blueprint for creating pretraining and fine-tuning datasets for indian languages. *arXiv preprint arXiv:2403.00001*.

Simran Khanuja, Anoop Dadu, and et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

S. Kj, V. Jain, S. Bhaduri, T. Roy, and A. Chadha. 2024. Decoding the diversity: A review of the indic ai research landscape. *arXiv preprint arXiv:2402.12345*.

R. Kumar, S. Kakde, D. Rajput, D. Ibrahim, R. Nahata, P. Sowjanya, D. Kumarr, G. Bhargava, and C. Khatri. 2024. Krutrim llm: A novel tokenization strategy for multilingual indic languages with petabyte-scale data processing. *arXiv preprint arXiv:2401.xxxxx*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. Technical report, ACL-04 Workshop, Barcelona, Spain. Text Summarization Branches Out.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392. Association for Computational Linguistics.

A. K. Singh, V. Kumar, R. Murthy, J. Sen, A. Mittal, and G. Ramakrishnan. 2024. Indicqa benchmark: A multilingual benchmark to evaluate question answering capability of llms for indic languages. *arXiv preprint arXiv:2401.11355*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.