MUCS@ - Question Answering in Hindi for Tourism: Evaluation of Transformer-Based Approaches on VATIKA

Rachana Nagaraju*,†, Hosahalli Lakshmaiah Shashirekha†

Department of Computer Science, Mangalore University, Mangalore, Karnataka, India

Abstract

Question Answer (QA) systems play a vital role in the field of Natural Language Processing (NLP) as they are designed to automatically generate precise answers to user queries expressed in natural language. In the tourism domain, QA systems are especially significant as they assist travelers by providing reliable and context-aware multilingual information, thereby enhancing visitor's experience overall. With the growing demand for intelligent information retrieval, such systems contribute directly promoting cultural heritage, supporting sustainable tourism, and improving accessibility of knowledge for domestic and international tourists. In view of these objectives, **VATIKA: Varanasi Tourism in Question Answer System** shared task emphasizes on domain-specific QA for tourism in Varanasi (Kashi), one of the world's oldest living cities and a prominent cultural-spiritual hub. The dataset spans 10 tourism-relevant domains such as Ganga Aarti, Cruise, Temples, Ashrams, Food Courts, and Museums, in Hindi and the system must accurately answer factual, navigational, and experiential queries. This task is crucial for enabling a smoother, enriched, and hassle-free tourism experience. In this paper, we — team **MUCS** — describe a transformer-based QA pipeline for fine-tuning `MuRIL`, a pre-trained multilingual model, for extractive QA. We explored three fine-tuning strategies: Hugging Face Trainer, a Custom AdamW Trainer, and a Simplified Trainer variant. On Test-A set, the Hugging Face Trainer achieved F1 score of 0.4972, BLEU score of 0.3529, and ROUGE-L score of 0.5239; the Custom AdamW Trainer approach obtained F1 score of 0.5003, BLEU score of 0.3454, and ROUGE-L score of 0.5300; while the Simplified Trainer produced F1 score of 0.4510, BLEU score of 0.3175, and ROUGE-L core of 0.5095. On the more challenging Test-B set, Hugging Face Trainer delivered the best overall results with an F1 score of 0.3351, BLEU score of 0.2214, and ROUGE-L score of 0.3621, compared to AdamW's 0.0416, 0.2810, and 0.2024, and Simplified Trainer's 0.0582, 0.1956, and 0.2165, F1-score, BLUE score and ROUGE-L score, respectively. These results highlight the effectiveness of the Hugging Face Trainer's fine-tuning strategy in capturing contextual semantics and maintaining robustness across diverse tourism-related queries in Hindi.

Keywords

Question Answer, Tourism, Hindi, Transformer Models, Information Retrieval, Sustainable Tourism

1. Introduction

Tourism plays a vital role in economic development worldwide, contributing to income generation, employment opportunities, and the preservation of cultural heritage. In India, tourism not only boosts regional pride and cultural exchange but also strengthens sustainable development by encouraging infrastructure growth and global awareness. Among Indian cities, Varanasi (also known as Kashi or Banaras) holds a unique position as one of the world's oldest living cities and as a spiritual and cultural hub. Millions of domestic and international tourists visit Varanasi every year, seeking spiritual enrichment, cultural experiences, and historical exploration. Hence, providing tourists with timely, accurate, and multilingual information has become increasingly important for improving their overall travel experience.

QA systems have emerged as a core application in the field of NLP. They are designed to automatically return precise answers to natural language queries posed by users, leveraging structured knowledge bases or unstructured documents. Unlike traditional search engines, which return a ranked list of documents, QA systems directly address the user's information need, thereby reducing cognitive

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India

*Corresponding author.

†These authors contributed equally.

rachananagaraju20@gmail.com (R. Nagaraju); hlsrekha@mangaloreuniversity.ac.in (H. L. Shashirekha)

0000-0002-9421-8566 (H. L. Shashirekha)

load and making information retrieval more user-friendly [1, 2]. The significance of QA becomes particularly evident in specialized domains such as healthcare [3], education [4], and tourism [5], where domain-specific queries require reliable and contextualized answers.

Most QA systems focus on high-resource languages like English, leaving low-resource languages unexplored in this direction. Many of the Indian languages are low-resourced due to limited annotated datasets, diverse scripts, and lack of robust language models tailored to regional nuances. Added to this, many Indian languages have multiple dialects and informal usage patterns, which complicate accurate understanding and response generation. Further, code-mixed queries and script variations add their share of challenges. Overall, the ecosystem is still evolving, with promising research but with significant gaps in resources and infrastructure.

VATIKA: Varanasi Tourism in Question Answer System shared task at Forum for Information Retrieval Evaluation (FIRE) 2025 invites researchers to develop models to address the challenges of QA systems in Hindi - a low-resource Indian language in tourism sector. The dataset is provided in structured JSON format, organized by *domain* → *context* → *QAs*. Each QA pair includes a unique ID, question in Hindi, and its corresponding answer as shown in Figures 1 and 2. VATIKA dataset is curated to cover ten important tourism–relevant domains of Varanasi, including Ganga Aarti, Cruise, Temples, Ashrams, Kunds, Museums, Food Courts, Travel Agencies, Public Toilets, and General Information. It consists of Hindi language contexts written in Devanagari script, paired with realistic QA pairs simulating actual tourist queries. This makes VATIKA one of the first Hindi QA datasets targeting a domain-specific real-world application in tourism. VATIKA shared task highlights the importance of specialized QA systems in tourism, which not only supports individual travelers, but also contributes significantly to cultural promotion and sustainable tourism growth.

In this paper, we - team **MUCS** describe the transformer-based QA models submitted for VATIKA shared task to answer tourism-related queries in Hindi. We experimented fine-tuning `MuRIL` - a pretrained multilingual model, using Hugging Face Trainer, Custom AdamW Trainer, and Simplified Trainer Variant, for extractive QA. The models are evaluated based on standard QA metrics - F1 score, BLEU, and ROUGE-L. Our model, fine-tuned with Hugging Face Trainer, emerged as the best-performing system, achieving **F1 score of 0.3351**, **ROUGE-L score of 0.2625**, and **BLEU score of 0.2214**, demonstrating the effectiveness of our pipeline in capturing semantic alignment between questions and contexts in Hindi. Our code is available on GitHub[1] to reproduce the results and explore further. The importance of this task lies not only in advancing research on Hindi QA systems but also in its direct applicability to real-world tourism. By enabling intelligent, accurate, and accessible information delivery to tourists, such systems can enhance cultural promotion, improve visitor satisfaction, and foster sustainable tourism growth in cities like Varanasi. This paper presents our approach to the shared task, detailing our methodology, experimental setup, and results, followed by an analysis of system performance, and discussion on future directions.

The subsequent sections of this paper details the related works (Section 2), methodology (Section 3), experiments, results, and implications of our approach (Section 4), declaration on generative AI (Section 5) followed by conclusion and future work (Section 6).

2. Related Works

Research in multilingual and Indic-language QA has seen rapid growth in recent years, particularly with the advancement of transformer-based architectures. Singh et al. [6] explored multilingual QA approaches for Indic languages, benchmarking transformer models such as mBERT, XLM-R, IndicBERT, and MuRIL. Their experiments reported F1 scores ranging between 58–72% across Indic languages, with IndicBERT showing superior results in low-resource contexts. Clark et al. [7] introduced TyDi QA, a multilingual benchmark that has been widely used for training and evaluating QA models in diverse languages, including Hindi. They reported average F1 score of 65% for high-resource languages and 45% for low-resource ones, demonstrating the challenges in handling morphologically rich languages

[1] https://github.com/rachanabn20/VATIKA-Varanasi-Tourism-in-Question-Answer-System

```
{
 "domains": [
  {
   "domain": "kund",
   "contexts": [
    {
     "context": "मणिकर्णिका चक्र पुष्करणीय कुंड लाल बहादुर शास्त्री अंतरराष्ट्रीय ह...",
     "qas": [
      {
       "id": "kund_1467",
       "question": "मणिकर्णिका चक्र पुष्करणीय कुंड लाल बहादुर शास्त्री अंतरराष्ट्रीय हवाई अड्डे (वाराणसी) से कितनी दूर है?",
       "answer": "मणिकर्णिका चक्र पुष्करणीय कुंड लाल बहादुर शास्त्री अंतरराष्ट्रीय हवाई अड्डे (वाराणसी) से 25.8 किलोमीटर दूर है।"
      }
      {
       "id": "kund_1468",
       "question": "मणिकर्णिका चक्र पुष्करणीय कुंड लाल बहादुर शास्त्री अंतरराष्ट्रीय हवाई अड्डे के पास से कैसे पहुँचा जा सकता है?",
       "answer": "मणिकर्णिका चक्र पुष्करणीय कुंड लाल बहादुर शास्त्री अंतरराष्ट्रीय हवाई अड्डे से यह दूरी टैक्सी या अन्य निजी परिवहन के माध्यम से तय की जा सकती है।"
      }
     ]
    }
   ]
  }
 ]
```


Figure 1: Sample Train Data

such as Hindi. Artetxe et al. [8] introduced the XQuAD benchmark, which consists of 1,190 QA pairs translated into ten languages, to evaluate zero-shot cross-lingual transfer. Their experiments showed that multilingual models such as mBERT, when fine-tuned on English, achieved nontrivial transfer performance across target languages. Subsequent studies reported that XLM-R outperformed mBERT by 7–10 points in both Exact Match score and F1 score across several languages, reaching around 70% F1 score for Hindi and Spanish, while performance dropped substantially for low-resource languages.

Li et al. [9] focused on domain-specific QA, proposing a tourism knowledge graph-based system. Their evaluations on tourism datasets achieved an F1 score of 81% and BLEU-4 of 26, demonstrating the efficiency of integrating structured knowledge with neural architectures for domain QA tasks. Contractor et al. [10] investigated QA for tourism-related queries, emphasizing user-generated reviews and geo-entity retrieval. Their neural QA system reported precision of 73% and recall of 68%, with an overall F1 score of 70%, highlighting the utility of combining entity retrieval with neural encoders. Nguyen et al. [11] proposed a tourism-oriented QA framework for conversational systems. Using transformer-based architectures, they demonstrated BLEU-4 scores of 28.5, ROUGE-L of 61.2, and F1 of 76% on a Vietnamese tourism corpus, showing strong applicability of QA models in tourism dialogue systems. Lee et al. [12] advanced QA systems by incorporating verification mechanisms to ensure answer reliability. Their evaluations across multilingual QA tasks showed F1 improvements of 4–6 points over standard transformers, reporting final F1 scores between 68–74% depending on the dataset, and emphasized trustworthiness in QA responses.

In summary, recent works show significant progress in QA systems in different languages and specific applications to tourism. However, only few studies have explored Hindi-centric QA systems for tourism, emphasizing the importance of contributions like VATIKA in bridging this research gap.

```
{
 "domains": [
  {
   "domain": "kund",
   "contexts": [
    {
     "context": "पांडव कुंड में दर्शन करने के लिए कोई निर्धारित समय सीमा नहीं है। फिर भी, श्रद्धालुओं के लि...",
     "qas": [
      {
       "id": "kund_1256",
       "question": "क्या नारद कुंड में दर्शन के लिए सुबह और शाम का समय सर्वोत्तम माना जाता है?",
       "answer": " "
      }
     ]
    }
   ]
  }
 ]
}
```


Figure 2: Sample Test Data

3. Methodology

The proposed methodology employs fine-tuning MuRIL - a multilingual pretrained model, for QA on the VATIKA dataset. To provide clarity on the fine-tuning process, we describe the end-to-end pipeline in detail. Three different training strategies are implemented to explore the impact of fine-tuning choices. The end-to-end pipeline for fine-tuning MuRIL model on VATIKA is illustrated in Figure 3.

3.1. Dataset Preparation

VATIKA dataset contains multiple domains with contexts paired with corresponding QA pairs. The JSON files are parsed such that each instance aligns a context with its corresponding question and gold-standard answer. These are then converted into Hugging Face `Dataset` objects for training, validation, and evaluation.

3.2. Pre-processing

MuRIL[2] tokenizer is applied to both questions and contexts. The following preprocessing steps ensure compatibility with extractive QA fine-tuning:

- Questions and contexts are tokenized with a maximum sequence length of 512 tokens.
- A sliding window with a stride of 128 tokens is applied to cover long contexts.
- Character-level answer spans are mapped to token indices to obtain start and end positions for supervision.
- Sequences are padded to a fixed length with attention masks for batching.

[2]https://huggingface.co/google/muril-base-cased

Table 1
Hyperparameter configuration for fine-tuning

Hyperparameter	Value
Model	`google/muril-base-cased`
Learning Rate	3×10^{-5}
Batch Size (per device)	8
Number of Epochs	2–3
Weight Decay	0.01
Max Sequence Length	512
Stride	128
Optimizer	AdamW

These steps collectively prepare the input data in a structured format suitable for training the QA model.

3.3. Fine-tuning Strategies

Fine-tuning a pretrained model for QA involves adapting that model to understand and extract answers from the context passages based on given questions. During fine-tuning, MuRIL model is extended with a QA-specific output head, consisting of a linear layer that predicts the start and end positions of the answer span within the context. The input is structured as a concatenation of the question and context, separated by special tokens, and tokenized using MuRIL's native tokenizer to preserve linguistic nuances in Hindi. While no modifications is made to the core architecture of MuRIL, the task-specific output layer is randomly initialized and trained from scratch. This ensures MuRIL's multilingual representations could be leveraged with learning task-specific parameters for the Hindi QA task.

We fine-tuned MuRIL model for Hindi QA using three distinct strategies to evaluate training efficiency and model performance, and the strategies are explained below:

1. **Hugging Face Trainer** uses `Trainer` API, which manages batching, forward and backward passes, loss computation, gradient updates, and optimizer scheduling automatically. Training runs with `AdamW` optimizer under this framework, while monitoring training loss across epochs.
2. **Custom AdamW Training** - is a manual training loop implemented without the high-level `Trainer`. Each epoch involves:
   - Forward pass of the model on a mini-batch
   - Loss computation using the predicted and gold answer spans
   - Backpropagation with `loss.backward()`
   - Parameter updates with the `AdamW` optimizer

   This setup allows explicit control over gradient accumulation, optimizer steps, and evaluation checkpoints.
3. **Simplified Trainer** is a reduced version of the `Trainer`, focusing exclusively on fine-tuning with training data. Unlike the full setup, this variant omits additional evaluation or logging steps during training, serving as a lightweight baseline for comparison.

In the three approaches, the fine-tuning process is supervised using question–context pairs from the VATIKA dataset. Each input pair is tokenized and aligned with annotated answer spans, enabling the model to learn semantic correspondence between questions and context passages. This design allowed us to evaluate the effect of different optimization strategies on MuRIL's ability to generalize in low-resource Hindi QA setting. The specific hyperparameters used in our experiments are summarized in Table 1.

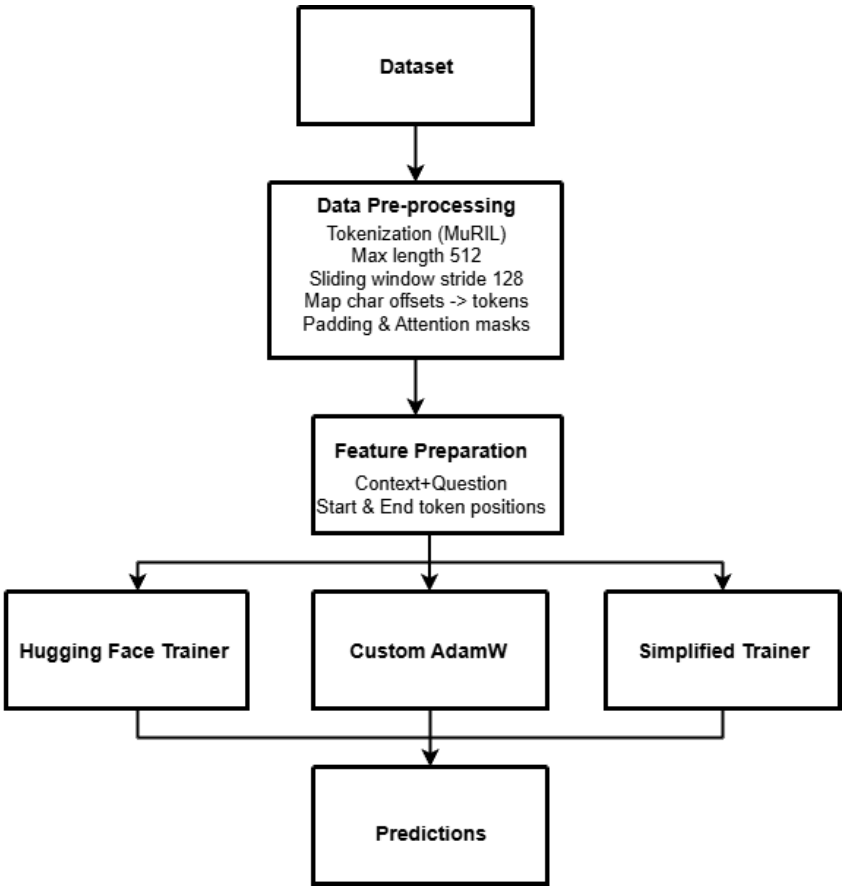

Figure 3: Overview of the proposed QA system

Table 2
Statistics of the VATIKA dataset

Dataset	Contexts	QA Pairs
Train	5,244	13,092
Validation	1,134	2,798
Test-A	1,143	2,902
Test-B	430	1,196

4. Experiments and Results

The experiments are designed to evaluate the performance of transformer-based models on the VATIKA dataset in Hindi. The goal is to assess effectiveness of the models to handle natural language queries related to tourist information and services in Hindi, ensuring a smooth and informative experience for users. VATIKA dataset is a domain-specific QA dataset comprising of contexts and QA pairs written in Hindi, covering multiple tourism-related domains such as Ganga Aarti, temples, cruises, museums, and public services. The dataset is divided into training, validation, and two test sets (Test-A and Test-B), and the statistics of datasets [13] are provided in Table 2.

4.1. Results

We evaluated the models on Validation and Test sets using multiple metrics: F1 Score, BLEU, and ROUGE-L, and the results for three fine-tuning strategies are presented in Table 3. The evaluation focused on the ability of the models to predict accurate and fluent answers aligned with the gold standard annotations. The results indicate that Hugging Face Trainer-based fine-tuning achieved the best performance, with the highest F1 and ROUGE-L scores, demonstrating strong alignment with the

Table 3
Performances of the proposed QA models

Fine-tuning Strategy	Metric	Validation	Test-A	Test-B
Hugging Face Trainer	F1	0.5104	0.4972	**0.3351**
	BLEU	0.3667	0.3529	0.2214
	ROUGE-L	0.5364	0.5239	**0.3621**
AdamW training	F1	0.5071	**0.5003**	0.0416
	BLEU	0.3499	0.3454	0.2810
	ROUGE-L	0.5372	**0.5300**	0.2024
Simplified Trainer	F1	0.4625	0.4510	0.0582
	BLEU	0.3280	0.3175	0.1956
	ROUGE-L	0.5210	0.5095	0.2165

gold standard answers. Custom AdamW Training and Simplified Trainer, fine-tuning strategies, while showing some competitiveness in BLEU scores, lagged behind in overall performance. This suggests that Hugging Face Trainer-based fine-tuning provided a better balance between exactness and fluency, making it the most effective configuration for the task.

The findings highlight the challenges of QA in Hindi, particularly in the tourism domain where answers may be diverse, context-specific, and phrased differently across contexts. However, the promising results of Hugging Face Trainer-based fine-tuning underscore the feasibility of building robust QA systems tailored for tourism applications in Hindi.

5. Declaration on Generative AI

Generative Artificial Intelligence (GenAI) tools are used in the preparation of this paper exclusively for language refinement, grammar correction, and LaTeX formatting assistance. GenAI is not used for generating research ideas, experiments, datasets, results, or conclusions. All core research activities, including data preprocessing, model training, evaluation, and interpretation, are performed entirely by the research team.

6. Conclusion and Future Work

This work presented a Hindi QA system for the tourism domain of Varanasi using VATIKA dataset, submitted by our team MUCS. QA system is developed by implementing three strategies - Hugging Face Trainer-based setup, Custom AdamW Trainer and a Simplified Trainer, to fine-tune pretrained `MuRIL` model, for extractive QA. Each fine-tuning strategy followed the same dataset preparation and pre-processing steps but differed in model optimization and training. Among the three models submitted by our team MUCS, Hugging Face Trainer-based fine-tuning achieved the best results with F1 score of 0.3351, BLEU score of 0.2214, and ROUGE-L score of 0.3621. These results confirm the feasibility of building robust domain-specific QA systems in Indian languages, where linguistic diversity and complex query styles pose unique challenges. Comparatively, Custom AdamW Trainer and Simplified Trainer fine-tuning strategies delivered lower F1 and ROUGE-L scores, underscoring the effectiveness of Hugging Face Trainer-based fine-tuning as the most reliable configuration. This work demonstrates the role of domain-adapted QA systems in enhancing the accessibility of tourism information, thereby enriching visitor experiences in culturally significant cities. Future work will aim to further improve contextual understanding, extend the system's applicability across diverse queries, and explore practical deployment strategies in real-world tourism scenarios.

References

[1] L. Hirschman, R. Gaizauskas, Natural Language Processing And Question Answering, Natural Language Engineering 7 (2001) 275–300.

[2] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ Questions for Machine Comprehension of Text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016, pp. 2383–2392.

[3] D. Weissenborn, G. Wiese, L. Seiffe, Making Neural Qa as Simple as Possible but Not Simpler, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL), 2017, pp. 271–280.

[4] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 2369–2380.

[5] J. Jiang, Information Extraction And Question Answering: Emerging Directions, in: Proceedings of the 2003 Conference on Computational Linguistics and Intelligent Text Processing, Springer, 2003, pp. 473–483.

[6] A. Singh, R. Kumar, P. Bansal, IndicQA: Multilingual Question Answering for Indic Languages, Journal of Natural Language Engineering 30 (2024) 145–162.

[7] J. H. Clark, E. Choi, M. Collins, D. Garrette, T. Kwiatkowski, J. Palomaki, S. P. Smith, TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4544–4560.

[8] M. Artetxe, S. Ruder, Xquad: A Cross-lingual Question Answering Dataset, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 253–261.

[9] W. Li, H. Zhang, M. Chen, TourismKB-QA: A Knowledge Graph Based Question Answering Framework for Tourism, Information Processing & Management 59 (2022) 103097.

[10] D. Contractor, S. Gupta, A. Sharma, TourismQA: Neural Question Answering for Tourism Information Retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 1265–1268.

[11] T. Nguyen, H. Le, M. Vo, SaigonTourism-QA: Transformer Based Conversational Question Answering for Tourism, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2025, pp. 2112–2125.

[12] H. Lee, J. Park, D. Kim, Verified QA: Enhancing Answer Reliability in Multilingual Question Answering, Transactions of the Association for Computational Linguistics 13 (2025) 122–138.

[13] P. Gatla, Anushka, N. Kanwar, G. Sahoo, R. K. Mundotiya, Tourism Question Answer System in Indian Language using Domain-Adapted Foundation Models, arXiv preprint (2025).