Towards Indian Intelligent Tourism Assistance: Design and Evaluation of the VATIKA QA Dataset

Praveen Gatla[1,*], Anushka[2], Nabanita Sadhukhan[3] and Rajesh Kumar Mundotiya[3]

[1]Department of Linguistics, Faculty of Arts, Banaras Hindu University, Varanasi, India
[2]Department of Humanistic Studies, Indian Institute of Technology (BHU), Varanasi, India
[3]Department of Computer Science and Engineering, Indian Institute of Technology Bhilai, India

Abstract

The VATIKA-2025 shared task aims to advance research in Indic language knowledge augmentation, focusing on generating context-aware answers grounded in culturally rich narratives. Designed as a benchmarking challenge for Indian language technologies, the task-VATIKA provides participants with a carefully curated dataset and evaluates system performance through established NLG and QA metrics, including BLEU, ROUGE, and QA-F1. A total of ten teams participated in the task, of which eight submitted working notes detailing their methodologies. Submissions demonstrated substantial variation in system performance, reflecting diverse modeling strategies such as fine-tuned language models, prompted LLMs, and ensemble-based approaches. The best-performing systems: VA-BO-INTERN (Run-3), IReL (Run-3), and Scaler (Run-1), achieved QA-F1 scores of 0.5757, 0.5507, and 0.5050, respectively, showing strong competency in generating high-quality, semantically aligned responses. This overview paper presents the task design, datasets, evaluation methodology, and a detailed comparative analysis of all team submissions to provide insights into current progress and future directions for Indic knowledge-grounded NLP research.

Keywords

Question-Answer, Tourism, Hindi, Benchmark.

1. Introduction

Varanasi, often described as the spiritual capital of India, has immense historical, cultural, and religious significance. Every corner of the city tells a story, whether it is the sight of pilgrims taking ritual baths in the ganga river (ghats), the sound of temple bells echoing through narrow lanes, or the smell of street food mingling with the chants of evening aarti. For first-time visitors, these experiences can be profoundly moving yet simultaneously overwhelming, raising questions about the significance of rituals, the history of sacred sites, or navigating the city's complex spiritual geography.

In this context, intelligent systems tailored for tourism can serve as valuable companions, providing accurate, contextual, and easily understandable information in a language that resonates with users. Considering this, the VATIKA 2025 shared task was conceived to explore the development of question answering systems specifically for Varanasi's tourism domain, with a focus on Hindi as the primary language. This resource enables participants not only to benchmark their systems but also to engage with the challenges that arise from working with low-resource languages in culturally rich contexts.

By bringing together researchers, VATIKA 2025 highlights the role of language technologies in making Indian cultural heritage more accessible. It reminds us that beyond metrics and models, the ultimate goal is to create systems that enrich visitors' journeys (yatra), preserve the stories of a timeless city, and foster innovation in the growing field of domain-specific question answer system.

Forum for Information Retrieval Evaluation, December 17-20, 2025, Varanasi, India
*Corresponding author.
praveengatla@bhu.ac.in (P. Gatla); anushka.rs.hss25@iitbhu.ac.in (Anushka); nabanitas@iitbhilai.ac.in (N. Sadhukhan); rmundotiya@iitbhilai.ac.in (R. K. Mundotiya)
0000-0002-8042-8685 (P. Gatla); 0009-0000-6030-5875 (Anushka); 0009-0002-1351-7175 (N. Sadhukhan); 0000-0002-0096-2440 (R. K. Mundotiya)

```
{
  "domains": [
    {
      "domain": "kund",
      "contexts": [
        {
          "context": "भागीरथ कुंड पं. दीन दयाल उपाध्याय रेलवे स्टेशन से
            14.1 किलोमीटर दूर है। स्टेशन से कुंड तक पहुँचने के लिए टैक्सी
            , कैब, या बस सेवाओं का उपयोग किया जा सकता है। यह स्टेशन पूर्व
            में मुगलसराय के नाम से जाना जाता था और भारत के प्रमुख रेल
            जंक्शनों में से एक है। यहाँ से भागीरथ कुंड तक की यात्रा
            वाराणसी की ऐतिहासिक गलियों और घाटों के दृश्य प्रदान करती है।
            इस यात्रा में भक्तों को वाराणसी की सांस्कृतिक विरासत का अनुभव
            मिलता है, जो इस धार्मिक स्थल के महत्व को और भी बढ़ा देता है।"
        },
        {
          "id": "kund_636",
          "question": "भागीरथ कुंड पं. दीन दयाल उपाध्याय रेलवे स्टेशन से
            कितना किलोमीटर दूर है?",
          "answer": "भागीरथ कुंड पं. दीन दयाल उपाध्याय रेलवे स्टेशन से 14
            .1 किलोमीटर दूर है।"
        },
        {
          "id": "kund_637",
          "question": "भागीरथ कुंड पं. दीन दयाल उपाध्याय रेलवे स्टेशन से
            कैसे पहुँच सकते हैं?",
          "answer": "पं. दीन दयाल उपाध्याय रेलवे स्टेशन से टैक्सी, कैब,
            या बस सेवाओं का उपयोग करके भागीरथ कुंड पहुँचा जा सकता है।"
        }
      ]
    }
  ]
}
```


Figure 1: Hindi Sample Data

2. VATIKA Task Description

The VATIKA 2025 shared task focuses on building a QA system to assist tourists in navigating Varanasi, with Hindi as the main language of interaction. Its aim is to design and evaluate systems that respond to visitors' questions, like the timings of the Ganga Aarti, directions to a temple or museum, or the nearest food court. By grounding the task in such authentic needs, VATIKA connects computational research to the lived realities of tourism.

A VATIKA dataset- a part of the Manually Created Hindi Question Answer Dataset (MCHQAD) [1]- extended to reflect real-world scenarios rather than artificial templates. Covering domains such as ghats, temples, ashrams, museums, food, travel agencies, and general guidance, the dataset captures the variety and richness of tourist queries. Emphasizing Hindi addresses the needs of domestic travelers while also filling a gap in resources, which are often dominated by English or culturally detached.

As shown in Figure 1, the dataset is provided in a structured JSON format organized hierarchically as domain → context → QAs. The VATIKA dataset spans ten domains: Ganga Aarti, Cruise, Food Court, Public Toilet, Kund, Museum, Travel Agencies, Ashram, Temple, and General Queries. Each domain contains context passages, natural Hindi questions, and their corresponding answers. The dataset is released in four splits—Train, Validation, Test-A, and Test-B. The statistics of these splits, along with

domain-wise distributions of contexts and QA pairs, are presented in Table 1.

Split	Type	aarti	ashram	cruise	foodcourt	general_qna	kund	museum	temple	toilet	travel	Total
Train	Context	10	1089	11	7	37	310	335	1818	6	1621	5244
	QA	23	2303	36	8	47	1055	446	5592	7	3575	13092
Validation	Context	3	234	3	3	9	68	73	391	2	348	1134
	QA	4	501	7	4	12	227	97	1168	3	775	2798
Test-A	Context	2	233	2	1	7	70	71	401	1	355	1143
	QA	7	492	8	4	16	233	105	1255	2	780	2902
Test-B	Context	15	64	16	11	53	42	40	100	9	80	430
	QA	40	172	48	25	133	124	115	299	18	222	1196

Table 1
Dataset distribution across sub-domains and splits.

3. Methodology and Results

Team	Submission	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	QA-F1	Rank
CSE_SVNIT	Run-1	43.7	22.9	14.8	10.8	0.0577	0.0260	0.0577	**0.4329**	6
	Run-2	41.3	20.9	14.3	10.6	0.0552	0.0226	0.0552	0.3939	
	Run-3	39.3	17.5	10.6	7.6	0.0790	0.0452	0.0790	0.2799	
AiNauts	Run-1	56.5	27.2	15.0	9.4	0.0818	0.0454	0.0818	**0.4529**	5
	Run-2	55.8	33.2	25.5	19.6	0.0518	0.0272	0.0518	0.1069	
NLP_Fusion	Run-1	26.0	12.2	6.0	3.5	0.0128	0.0020	0.0128	**0.2808**	9
IReL	Run-1	51.2	30.5	20.7	15.4	0.0587	0.0235	0.0587	0.4169	2
	Run-2	49.2	28.6	19.8	15.1	0.0594	0.0340	0.0594	0.4612	
	Run-3	61.5	36.4	24.5	17.9	0.0824	0.0467	0.0824	**0.5507**	
Scalar	Run-1	63.7	41.2	29.4	22.5	0.0561	0.0260	0.0561	**0.5050**	3
	Run-2	52.3	25.9	14.0	8.1	0.0288	0.0154	0.0276	0.3937	
	Run-3	46.5	21.8	11.0	5.9	0.0201	0.0094	0.0192	0.3518	
Namaste NLP	Run-1	59.9	31.7	21.2	15.7	0.0131	0.0038	0.0131	0.2429	7
	Run-2	44.2	25.6	17.4	12.8	0.0626	0.0360	0.0626	0.3056	
	Run-3	40.6	22.6	15.0	10.9	0.0685	0.0392	0.0685	**0.3519**	
MUCS	Run-1	36.7	20.2	13.8	10.1	0.0759	0.0438	0.0759	**0.3351**	8
	Run-2	36.3	22.0	16.0	12.2	0.0096	0.0078	0.0096	0.0416	
	Run-3	9.4	0.9	0.6	0.4	0.0685	0.0411	0.0685	0.0582	
IIIT Surat	Run-1	6.8	0.4	0.3	0.2	0.0061	0.0013	0.0061	**0.0618**	10
	Run-2	6.8	0.4	0.3	0.2	0.0061	0.0013	0.0061	0.0618	
	Run-3	6.8	0.4	0.3	0.2	0.0061	0.0013	0.0061	0.0618	
VA-BO-INTERN	Run-1	49.5	27.0	17.5	12.5	0.0773	0.0424	0.0773	0.5663	1
	Run-2	60.1	38.4	27.2	20.5	0.0777	0.0415	0.0777	0.5617	
	Run-3	60.9	38.7	27.3	20.6	0.0763	0.0411	0.0763	**0.5757**	
Ghumelu	Run-1	13.5	3.3	1.0	0.5	0.0633	0.0301	0.0631	**0.4679**	4

Table 2
Model performance results across teams and submissions on Test-B.

As shown in Table 2, a total of ten teams viz. VA-BO-INTERN, IReL, Scalar, AiNauts, CSE_SVNIT, NLP_Fusion, MUCS, IIIT_Surat participated in the VATIKA task, out of which eight teams submitted working notes. The system ranking for VATIKA is determined based on the QA-F1 score, with VA-BO-INTERN (Run-3), IReL (Run-3), and Scaler (Run-1) achieving the first, second, and third positions, respectively. The methodologies adopted by each team and their corresponding results are summarized in this section.

IIIT SURAT [2] employed a retriever-reader framework centered on a pre-trained `IndicBERT` model. To ensure input consistency, Hindi text normalization was performed using the indic-nlp-library, followed by the alignment of character-level answer boundaries to token-level indices. The model was fine-tuned for the extractive QA task via the `AutoModelForQuestionAnswering` architecture, with Hugging Face Trainer API's optimizer. For inference, the system integrates FAISS-based semantic search to retrieve relevant contexts. The model subsequently predicts the optimal start and end token spans, which are decoded into surface text, supplemented by a fallback mechanism for low-confidence queries. They demonstrated consistent performance across all three submitted runs. Each run achieves a BLEU-4 score of 0.2, with minimal variation in the associated metrics, indicating highly stable model behavior. The corresponding F1 scores are uniformly low, with Run-1, Run-2, and Run-3 all registering 0.0061 for the primary F1 measure, reflecting limited accuracy in the predicted outputs.

NLP_Fusion [3] has fine-tuned the `mT5-small` model on the provided data. They submitted a single run that achieved a BLEU-4 score of 3.5, indicating limited fluency and n-gram overlap with the reference texts. The F1 score of approximately 0.28 reflects moderate answer accuracy but suggests room for improvement.

VA-BO-INTERN [4] investigated the efficacy of synthetic data augmentation for Long-Form Question Answering (LFQA) using Small Language Models (SLMs). The team employed large teacher models—specifically `Llama-3.1-70B` and `Phi-4-14B` to generate synthetic QA pairs via few-shot prompting on training contexts. Three fine-tuning strategies were evaluated: a baseline `Llama-3.1-8B` trained solely on gold data ($M1$); a continued fine-tuning approach ($M2$) where $M1$ was further trained on `Phi-4-14B` synthetic data; and a multi-source strategy ($M3$) training on a composite dataset of real instances plus synthetic samples from both teacher models. To address script-specific challenges, the tokenizer was optimized for Hindi character handling. VA-BO-INTERN exhibited a clear and consistent improvement across their three runs, with BLEU-4 scores increasing from 12.5 in Run-1 to 20.6 in Run-3, indicating enhanced fluency and n-gram alignment with reference texts. Their F1 scores also remain strong and stable, peaking at 0.5757 in the final run, which reflects accurate and reliable answer prediction.

Scaler [5] proposed a hybrid encoder-decoder framework designed to decouple understanding and generation. The system utilizes `l3cube-pune/hindi-bert-v2` as an encoder for Hindi text representation, connected via a linear projection layer to a decoder (ai4bharat/IndicBART) for natural language generation. This end-to-end architecture is further augmented with a NER module to explicitly identify entity spans within the context, enhancing interpretability. The Scalar team exhibited a gradual decline in performance across their three runs. BLEU scores consistently decreased, with BLEU-4 dropping from 22.5 in Run-1 to 5.9 in Run-3, indicating a reduction in n-gram overlap and fluency with the reference texts. The QA-F1 score also declined notably, from 0.5050 in Run-1 to 0.3518 in Run-3, suggesting a decrease in the accuracy and reliability of answer prediction.

IReL [6] explored a multi-paradigm approach, implementing three distinct strategies: (1) a generative method fine-tuning mT5 for multilingual adaptability; (2) a span-based extractive approach utilizing XLM-RoBERTa, supplemented by post-processing heuristics to refine short-span predictions; and (3) a zero-shot baseline leveraging ChatGPT with batch-wise prompt engineering to establish a comparative benchmark against the supervised models. Across the three IReL submissions, Run-3 achieved the strongest overall performance, outperforming the other systems on all BLEU and ROUGE metrics as well as QA-F1. Specifically, it obtained the highest BLEU-1 (61.5), BLEU-2 (36.4), BLEU-3 (24.5), and BLEU-4 (17.9) scores, indicating superior n-gram precision. This trend was consistent in the ROUGE measures, where Run-03 yielded the highest ROUGE-1 (0.0824), ROUGE-2 (0.0467), and ROUGE-L (0.0824) scores, reflecting better recall-oriented text overlap. Furthermore, it achieved the QA-F1 score (0.5507), indicating stronger relevance and accuracy of the answers.

CSE_SVNIT [7] focused on static embedding architectures to model semantic similarity. The approach leveraged pre-trained FastText embeddings to generate 300-dimensional vectors, aggregated into sentence-level representations. These vectors were utilized in two configurations: unsupervised retrieval via cosine similarity to identify relevant contexts and a supervised ridge regression model for answer span prediction. Additionally, Word2Vec embeddings were employed to encode dense semantic vectors, providing a comparative basis for context alignment tasks. They showed a declining trend in BLEU-4 scores across their three runs, dropping from 10.8 in Run-1 to 7.6 in Run-3, indicating a reduction in n-gram overlap and fluency with reference texts. Similarly, their F1 scores decrease from 0.4329 in Run-1 to 0.2799 in Run-3, reflecting a decline in answer accuracy and consistency. Despite this, Run-3 shows a slight increase in precision and recall metrics, suggesting some improvement in specific aspects of model output quality.

AiNauts [8] concentrated on fine-tuning large pre-trained multilingual models, specifically mBART-50 and mT5-small. The preprocessing pipeline involved concatenating the question and context into a single sequence, truncated to a maximum length of 512 tokens. The models were optimized to leverage their encoder-decoder attention mechanisms for extracting and generating answers from the provided Hindi contexts. Between the two AiNauts submissions, Run-1 demonstrated stronger performance across most evaluation metrics, particularly in ROUGE and QA-F1. Although Run-2 achieved higher BLEU-2 (33.2), BLEU-3 (25.5), and BLEU-4 (19.6) scores, indicating improved multi-gram precision. Moreover, Run-1 achieved a markedly higher QA-F1 score (0.4529) compared to Run-2 (0.1069), suggesting considerably better answer accuracy and semantic alignment.

MUCS [9] fine-tunes the MuRIL model for the dataset using a structured pipeline consisting of dataset preparation, preprocessing, and multiple training strategies. Preprocessing employs the MuRIL tokenizer with sequence-length constraints, sliding windows for long contexts, token-level mapping of answer spans, and padding with attention masks. Fine-tuning adds a QA-specific linear output layer to MuRIL to predict answer span,i.e., start and end positions, while the base architecture remains unchanged. Three training strategies are examined: (1) the Hugging Face Trainer, which automates optimization and training workflows; (2) a custom AdamW training loop that provides explicit control over model updates; and (3) a simplified Trainer variant that performs minimal fine-tuning without evaluation or logging. This setup enables comparison of training efficiency and performance across different fine-tuning approaches. Among the three MUCS submissions, Run-1 delivered the most balanced and overall strongest performance. It achieved the highest BLEU-1 (36.7), BLEU-3 (13.8), and BLEU-4 (10.1) scores, along with the ROUGE-1 (0.0759), ROUGE-2 (0.0438), and ROUGE-L (0.0759) values, indicating superior lexical overlap and recall-driven text similarity. Run-2 showed marginal improvements over Run-1 only in BLEU-2 (22.0 vs. 20.2) and had higher BLEU-3 and BLEU-4 than Run-3, but its ROUGE and QA-F1 scores were substantially lower, with QA-F1 dropping to 0.0416. Run-3 exhibited the weakest performance overall, particularly on BLEU metrics, where scores fell below 1 for BLEU-2 through BLEU-4; however, its ROUGE scores remained moderately comparable to the other systems.

4. Conclusion

The VATIKA-2025 shared task provided a comprehensive platform for evaluating knowledge-grounded answer generation systems in Indic languages. The diversity of participating teams and methodologies highlights the growing interest in culturally anchored NLP tasks and the rapid evolution of models capable of reasoning over narrative contexts. The evaluation results show that systems leveraging larger pre-trained language models or hybrid architectures consistently outperformed traditional baselines, achieving higher BLEU, ROUGE, and QA-F1 scores. Among all participants, VA-BO-INTERN (Run-3) attained the highest QA-F1 score of 0.5757, followed by IReL (Run-3) and Scaler (Run-1), demonstrating strong capability in producing contextually relevant and semantically accurate responses. At the same time, several submissions with lower performance highlight ongoing challenges in handling long contexts, maintaining semantic consistency, and generating fluent responses in Indic languages. Overall, VATIKA-2025 offers valuable insights into current system strengths and limitations, establishes new performance benchmarks, and provides clear directions for future research, particularly in enhancing reasoning abilities, cultural grounding, and cross-lingual generalization in Indian-language NLP systems.

Declaration on Generative AI

As we wrote the paper, we only employed a generative AI assistant in a limited way to facilitate the writing process. The AI was mostly used to help refine the language, help structure sections, and maintain consistency in LaTeX format.

Acknowledgment

We thank Banaras Hindu University, Varanasi for providing the grant as a part of Transdisciplinary Research Grant, Institute of Eminence. We also thank the annotators Shreya Pandey, Bhaskar Singh, Aman Gupta, Himesh Jee Amar, Abhilasha Gupta, and others for extending their hand to create the VATIKA dataset. We thank Supriya Chauhan, Iram Ali Ahmad, Jyoti Kumari for proofreading the dataset. We also thank Jagdeesan T, Suresh S. for the academic collaboration during the Transdisciplinary grant, Institute of Eminence at BHU.

References

[1] 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).

[2] R. Kumar, S. C. Jaiswal, D. Bhatia, Varanasi tourism in question answer system track: Iiit surat @ fire'25 shared task, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, Varanasi, India, 2025.

[3] A. Hegde, S. Coelho, A. M. Shetty, M. Z. Taljeh, Hindi tourism qa system: Low-resource question answering using mt5-small, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, Varanasi, India, 2025.

[4] S. Majhi, P. Bhattacharya, Va-bo-intern: Adapting small language models to low-resource domains: A case study in hindi tourism qa, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, Varanasi, India, 2025.

[5] P. K. R. N. Subbannagari, A. S. Velidi, A. K. Madasamy, Vatika-qa: A hybrid bert-indicbart approach for hindi question answering in tourism domain, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, Varanasi, India, 2025.

[6] K. Tewari, S. Chanda, A. Chaturvedi, Tirtha: Tourism information retrieval and text-based hindi answering, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, Varanasi, India, 2025.

[7] A. Jariwala, S. S. Sahu, Svnit_cse: Building a question answering system for hindi using word-embedding, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, Varanasi, India, 2025.

[8] H. Mishra, N. Yadav, N. K. Tagore, R. K. Kumar, Vatika: A hindi machine reading comprehension approach for varanasi tourism question answering using mt5, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, Varanasi, India, 2025.

[9] R. Nagaraju, H. L. Shashirekha, Mucs@: Question answering in hindi for tourism: Evaluation of transformer-based approaches on vatika, in: Working Notes of FIRE 2025 - Forum for Information Retrieval Evaluation, CEUR Workshop Proceedings, CEUR-WS.org, Varanasi, India, 2025.