# TIRTHA: Tourism Information Retrieval and Text-based Hindi Answering

Krishna Tewari[1,*], Supriya Chanda[2] and Aarya Chaturvedi[1]

[1]Indian Institute of Technology (BHU), Varanasi, INDIA
[2]Bennett University, Greater Noida, INDIA

**Abstract**

Hindi Tourism QA (HTQA) addresses the challenge of extracting precise answers from Hindi context paragraphs within the specialized tourism domain of Varanasi, where limited annotated resources and complex linguistic structures pose significant hurdles. As part of the FIRE 2025 VATIKA shared task, which focuses on Hindi-language QA, we developed and evaluated multiple QA approaches using a structured dataset consisting of context–question pairs in JSON format. Three main strategies were explored: (i) fine-tuning the multilingual mT5 model, which demonstrated reasonable language support but occasionally produced fallback answers; (ii) span-based extractive modeling using XLM-RoBERTa, enhanced with post-processing techniques to refine short-span predictions; and (iii) a zero-shot approach leveraging ChatGPT with batch-wise prompt engineering applied over 50 context–question pairs purely for comparative analysis. Evaluation was performed using BLEU (1-4), ROUGE-L, and QA-F1 metrics. While ChatGPT achieved higher metric scores, only open-source models are considered for leaderboard results; hence, the ChatGPT results are reported separately as ablation.

**Keywords**

QA, Extractive QA, XLM-RoBERTa, ChatGPT, Zero-shot Learning, Tourism

## 1. Introduction

The rich cultural and spiritual heritage of Varanasi, also known as Kashi, makes it one of the world's oldest living cities and a prominent pilgrimage destination in India. Renowned for its sacred kunds, temples, and ghats, the city attracts millions of tourists and devotees each year. However, most information about these landmarks exists in unstructured textual formats, which poses significant barriers for Hindi-speaking visitors seeking concise, accurate, and reliable knowledge.

Hindi-language QA (QA) systems offer a solution by automatically extracting precise answers from large bodies of text, enabling efficient information retrieval for end-users. A typical QA task involves processing a question $Q = (q_1, q_2, \ldots, q_T)$ in natural Hindi language and retrieving the correct answer span from a given context paragraph $C = (c_1, c_2, \ldots, c_N)$. However, several challenges complicate this process in low-resource and specialized domains. First, the lack of large-scale annotated datasets in Hindi limits supervised training of robust models [1, 2]. Second, domain-specific variability in phrasing, complex syntactic structures, and culturally grounded concepts further increase modeling difficulty [3]. Third, ambiguity in question formulation and answer granularity creates additional hurdles in achieving precise and reliable retrieval [4].

To advance research in this direction, the Forum for Information Retrieval Evaluation (FIRE) introduced the VATIKA shared task in 2025 [5], focusing on Hindi QA in the tourism domain of Varanasi. The dataset comprises structured JSON instances pairing context passages about sacred sites with corresponding questions, providing a valuable benchmark for systematic development and evaluation of QA systems.

In this work, we benchmark multiple QA approaches in this culturally rich, low-resource setting, including transformer-based fine-tuning and zero-shot prompting strategies using large language models

*Forum for Information Retrieval Evaluation, December 17-20, 2025, India*
*Corresponding author.
krishnatewari.rs.cse24@iitbhu.ac.in (K. Tewari); suplife24@gmail.com (S. Chanda); aarya.chaturvedi.mec22@itbhu.ac.in (A. Chaturvedi)
0009-0005-6599-9956 (K. Tewari); 0000-0002-6344-8772 (S. Chanda)

as comparative analysis [6]. Our study demonstrates the potential of these approaches and highlights key challenges in building robust Hindi QA systems for specialized domains, pointing toward promising directions for future research.

The rest of the paper is structured as follows: Section 2 discusses related work; Section 3 describes the dataset; Section 4 presents the proposed methodology; Section 5 reports results and analysis; and Section 6 concludes with key findings.

## 2. Related Work

The field of QA has been fundamentally reshaped by the introduction of the Transformer architecture, which enabled large pre-trained language models (PLMs) to excel across NLP tasks [7]. Early breakthroughs such as BERT established the dominant pre-train and fine-tune paradigm, learning rich contextual representations from vast text corpora to achieve state-of-the-art performance on many language tasks [8].

Two primary paradigms for QA have emerged. Extractive QA, popularized by benchmark datasets like SQuAD [1], formulates the task as span prediction over a context paragraph. Cross-lingual transformers such as XLM-RoBERTa have demonstrated strong performance in this space by enabling transfer learning across languages [9]. In contrast, Generative QA treats the task as a text-to-text problem, where models like T5 unify multiple NLP tasks into a single sequence-to-sequence framework [10].

With the advent of Large Language Models (LLMs) such as GPT-3 and GPT-4, zero-shot and few-shot prompting strategies have gained significant attention. These models perform tasks by interpreting instructions embedded in prompts, often achieving competitive results without task-specific fine-tuning [11, 12]. Zero-shot prompting has proven especially viable for low-resource settings, allowing models to generalize to unseen tasks [6, 13].

While most research in QA has focused on high-resource languages such as English, several efforts have extended QA to low-resource and cross-lingual settings. IndicQA and TyDi QA are notable benchmarks focusing on diverse Indian languages, highlighting challenges such as code-mixing, transliteration, and limited data availability [14, 15]. Transfer learning and multilingual pretraining strategies have been proposed to overcome these challenges, demonstrating that models pretrained on multilingual corpora (e.g., mBERT) show strong cross-lingual transferability [16, 17].

Domain-specific QA has also seen increasing interest. Specialized benchmarks in medical, legal, and scientific domains have revealed that generic models often struggle with domain-specific jargon and knowledge representation [18, 19, 20]. Fine-tuning on domain-specific data significantly improves performance but remains challenging in low-resource settings.

Recent studies have started exploring hybrid architectures that combine neural and symbolic methods to improve robustness and interpretability [21, 22]. Such models aim to bridge the gap between purely data-driven approaches and rule-based systems, often improving precision and reducing ambiguity in specialized applications.

Despite these advances, a direct comparative analysis of extractive, generative, and zero-shot paradigms on a low-resource, culturally specific dataset such as the VATIKA Hindi QA remains underexplored. Our work benchmarks these paradigms in a tourism domain setting, shedding light on their practical effectiveness and identifying key areas for future improvement.

## 3. Dataset

The dataset used in this study is released as part of the FIRE 2025 VATIKA Shared Task on Hindi QA. It is designed to support machine reading comprehension (MRC) and QA applications in the tourism domain of Varanasi, focusing on cultural and spiritual heritage. The dataset is provided in a structured JSON format, organized by domain → context → question-answer pairs.

Each entry is organized into three primary fields: **Context**, a factual, descriptive paragraph in Hindi (Devanagari script) detailing specific landmarks (e.g., temples, kunds, ghats), historical events, or cul-

tural rituals in Varanasi; **Question**, a fact-seeking wh-question in Hindi (e.g., "कहाँ," "कब," "कौन"), designed to be answerable based solely on the provided context; and **Answer**, the ground-truth answer, a verbatim span directly extracted from the context paragraph, enforcing an extractive span prediction task.

The dataset is pre-divided into training, validation, and test splits to ensure standardized evaluation. The training set contains 2,452 question-answer pairs, the validation set contains 273 pairs, and the blind test set contains 915 pairs. The full distribution is summarized in Table 1.

**Table 1**
**VATIKA Dataset Statistics**

| Split | Number of QA Pairs |
|---|---|
| Training | 2,452 |
| Validation | 273 |
| Test | 915 |
| **Total** | **3,640** |

The VATIKA dataset covers 10 tourism-relevant domains: Ganga Aarti, Cruise, Food Court, Public Toilet, Kund, Museum, General, Ashram, Temple, and Travel. Each domain includes detailed paragraph-level contexts followed by multiple question-answer pairs, simulating real-world information-seeking behavior in natural Hindi language.

A representative structured entry from the "kund" domain is shown below:

**Domain:** kund

**Contexts:**

- मणिकर्णिका चक्र पुष्करणीय कुंड लाल बहादुर शास्त्री अंतरराष्ट्रीय ह...
  - **QID:** 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।
  - **QID:** 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।

A qualitative review of the data highlights several key characteristics. Contexts are rich in proper nouns (e.g., place names, deity names), dates, and factual details. The questions are predominantly factoid, focusing on the retrieval of specific entities rather than complex reasoning or synthesis. Answer spans are typically short phrases directly extracted from the context. This structured and curated dataset provides a robust benchmark for evaluating extractive QA models in a specialized low-resource Hindi setting, promoting research toward domain-specific QA systems.

## 4. Methodology

We address the problem of developing a robust Hindi QA system for the Varanasi tourism domain. Formally, given a question $q$ in Hindi and a set of context paragraph $c$, the goal is to produce an answer $a$ that is fluent, factually consistent, and derived strictly from the provided context. This can be expressed as:

$$\hat{a} = \arg \max_{a \in \mathcal{A}(c)} P(a \mid q, c),$$

where $\mathcal{A}(c)$ denotes the set of plausible answer spans or sequences within the context. For extractive methods, $\mathcal{A}(c)$ is restricted to spans of text that exist verbatim in $c$, while for large language model (LLM) approaches, $\mathcal{A}(c)$ encompasses all possible text sequences that can be generated from the context.

The design of our QA system is centered around three complementary computational paradigms: generative QA using fine-tuned mT5, extractive QA using XLM-RoBERTa with post-processing, and zero-shot answer generation using a large language model. These paradigms were selected to leverage their respective strengths and flexibility for generative QA, precision and interpretability for extractive QA, and contextual fluency and completeness for LLM-based generation.

### 4.1. Generative QA with Fine-Tuned mT5

Our first approach employed the mT5-small model, a multilingual version of the Text-to-Text Transfer Transformer (T5) [10]. The T5 framework uniquely treats all NLP tasks as a text-generation problem, making it a flexible choice for generative QA. We fine-tuned the model on the official training set by providing the question and context as input, with the objective of teaching the model's decoder to generate the ground-truth answer. Despite its potential to produce fluent responses, this approach proved underwhelming. The model often defaulted to generic, uninformative answers (e.g., "उत्तर स्पष्ट नहीं है"), suggesting that the limited size of the training corpus was insufficient for robust domain adaptation. This highlighted the significant data and computational requirements of fine-tuning generative models for specialized tasks.

All experiments were conducted using the PyTorch framework and the Hugging Face Transformers library. For the generative approach, the `google/mt5-small` model was fine-tuned for 5 epochs with a batch size of 8 and a learning rate of 2e-5 using the AdamW optimizer.

### 4.2. Extractive QA with XLM-RoBERTa and Post-Processing

The extractive paradigm employs XLM-RoBERTa (XLM-R) [9], a transformer-based model pretrained for cross-lingual understanding, capable of processing Hindi text directly. The model formulates QA as a span prediction problem: given a context paragraph $c_i \in C$, it predicts a start token $s$ and an end token $e$ such that the answer is extracted as:

$$a = c_s c_{s+1} \dots c_e,$$

where $c_j$ denotes the $j$-th token of the context.

**Challenges in raw predictions:** Despite the model's accuracy at identifying relevant tokens, we observed two recurring issues: 1. *Incomplete spans:* The predicted spans were often too short, omitting critical contextual information necessary for coherent understanding. 2. *Low-confidence predictions:* In cases involving ambiguous questions or rare domain-specific vocabulary, the model occasionally generated predictions with very low confidence scores, leading to unreliable outputs.

To address these challenges, we devised a two-step post-processing pipeline that improves answer completeness and reliability:

1. **Sentence Expansion:** The predicted span $(s, e)$ is mapped back to the full sentence containing it, producing a more comprehensive answer:

$$a_{\text{expanded}} = \text{sentence\_containing}(c_s \dots c_e)$$

2. **Confidence Filtering:** Predictions with confidence below a threshold (empirically set at 0.05) that are unusually short are further analyzed. We check for the presence of domain-specific keywords (e.g., names of locations, temples, or ghats relevant to Varanasi). If keywords are missing, the answer is replaced with a standard fallback message:

$$a_{\text{final}} = \begin{cases} a_{\text{expanded}}, & \text{if confidence is high or keywords present,} \\ \text{उत्तर उपलब्ध नहीं है}, & \text{otherwise.} \end{cases}$$

**Implementation Details:** We use the `deepset/xlm-roberta-base-squad2` checkpoint. Context paragraphs are tokenized using XLM-R's SentencePiece tokenizer. The model processes inputs in batches of 16. By integrating sentence expansion and confidence filtering, this extractive pipeline produces answers that are both accurate and contextually complete while remaining interpretable.

### 4.3. Zero-Shot Prompting with a Large Language Model

As an ablation experiment, we used large language model (LLM), specifically ChatGPT / GPT-4o mini [12], in a zero-shot setting. This model was not part of the official runs due to task restrictions prohibiting closed-source systems. Unlike extractive QA, LLMs generate answers as free-form sequences of text rather than extracting spans. This approach does not require fine-tuning on domain-specific data.

For each question-context pair $(q, C)$, we construct a detailed prompt that instructs the model to answer strictly using the provided context. The prompt is formulated as:

prompt = 'Please provide answer based on the given context only'

$q$ = मणिकर्णिका चक्र पुष्करणीय कुंड लाल बहादुर शास्त्री अंतरराष्ट्रीय हवाई अड्डे (वाराणसी) से कितनी दूर है?

$C$ = मणिकर्णिका चक्र पुष्करणीय कुंड लाल बहादुर शास्त्री अंतरराष्ट्रीय ह...

**Advantages and rationale:** The zero-shot LLM approach offers several benefits:

- *Fluency:* Answers are generated in grammatically correct and natural Hindi.
- *Contextual completeness:* The model can combine information from multiple sentences to produce richer answers.
- *High performance without fine-tuning:* The model performs well in this domain, making zero-shot prompting effective.

**Implementation Details:** Prompts are submitted in batches of 50 question-context pairs via the OpenAI. Responses are parsed to extract the answer segment, discarding any additional commentary.

Results from this ablation are reported separately for reference and are excluded from official leaderboard discussion.

## 5. Results

The VATIKA 2025 Shared Task evaluated submissions on Test Data-II using three complementary families of metrics: (i) **QA-F1**, the primary measure balancing precision and recall; (ii) **BLEU-1 to BLEU-4**, assessing lexical overlap and fluency across increasing $n$-gram lengths; and (iii) **ROUGE-L**, capturing the longest common subsequence and content coverage. The official leaderboard, covering all participating teams and runs, is presented in Table 2.

IReL's submissions show a clear progression across runs. Run 1 established a baseline (QA-F1 of 0.4169, BLEU-4 of 15.4), but its precision and recall were limited. Run 2 improved moderately in QA-F1 (0.4612), indicating better overall ranking, while maintaining similar BLEU and ROUGE-L values.

Compared with other teams, IReL's Run 2 is highly competitive. Its QA-F1 of 0.4612 surpasses all runs from CSE_SVNIT, MUCS, Namaste NLP, NLP Fusion, and IIIT Surat, while also outperforming AiNauts (best QA-F1 of 0.4529). In summary, IReL demonstrated steady improvements across its two runs, culminating in Run 2, which achieved competitive performance against the best systems in the task.

**Table 2**
**Performance metrics for all teams in the FIRE 2025 VATIKA Shared Task.**

| Team | Run | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L | QA-F1 |
|---|---|---|---|---|---|---|---|
| AiNauts | Run 1 | 56.5 | 27.2 | 15.0 | 9.4 | 0.0818 | 0.4529 |
| | Run 2 | 55.8 | 33.2 | 25.5 | 19.6 | 0.0518 | 0.1069 |
| CSE_SVNIT | Run 1 | 43.7 | 22.9 | 14.8 | 10.8 | 0.0577 | 0.4329 |
| | Run 2 | 41.3 | 20.9 | 14.3 | 10.6 | 0.0552 | 0.3939 |
| | Run 3 | 39.3 | 17.5 | 10.6 | 7.6 | 0.0790 | 0.2799 |
| IIIT Surat | Run 1 | 6.8 | 0.4 | 0.3 | 0.2 | 0.0061 | 0.0618 |
| | Run 2 | 6.8 | 0.4 | 0.3 | 0.2 | 0.0061 | 0.0618 |
| | Run 3 | 6.8 | 0.4 | 0.3 | 0.2 | 0.0061 | 0.0618 |
| **IReL** | **Run 1** | **51.2** | **30.5** | **20.7** | **15.4** | **0.0587** | **0.4169** |
| | **Run 2** | **49.2** | **28.6** | **19.8** | **15.1** | **0.0594** | **0.4612** |
| MUCS | Run 1 | 36.7 | 20.2 | 13.8 | 10.1 | 0.0759 | 0.3351 |
| | Run 2 | 36.3 | 22.0 | 16.0 | 12.2 | 0.0096 | 0.0416 |
| | Run 3 | 9.4 | 0.9 | 0.6 | 0.4 | 0.0685 | 0.0582 |
| Namaste NLP | Run 1 | 59.9 | 31.7 | 21.2 | 15.7 | 0.0131 | 0.2429 |
| | Run 2 | 44.2 | 25.6 | 17.4 | 12.8 | 0.0626 | 0.3056 |
| | Run 3 | 40.6 | 22.6 | 15.0 | 10.9 | 0.0685 | 0.3519 |
| NLP Fusion | Run 1 | 26.0 | 12.2 | 6.0 | 3.5 | 0.0128 | 0.2808 |
| Scalar | Run 1 | 63.7 | 41.2 | 29.4 | 22.5 | 0.0561 | 0.5050 |
| | Run 2 | 52.3 | 25.9 | 14.0 | 8.1 | 0.0276 | 0.3937 |
| | Run 3 | 46.5 | 21.8 | 11.0 | 5.9 | 0.0192 | 0.3518 |
| VA-BO-INTERN | Run 1 | 49.5 | 27.0 | 17.5 | 12.5 | 0.0773 | 0.5663 |
| | Run 2 | 60.1 | 38.4 | 27.2 | 20.5 | 0.0777 | 0.5617 |
| | Run 3 | 60.9 | 38.7 | 27.3 | 20.6 | 0.0763 | 0.5757 |

### 5.1. Ablation: Zero-Shot Closed-Source Baseline

While Runs 1 and 2 were submitted officially, an additional ablation using a closed-source ChatGPT model (Run 3) yielded higher scores (QA-F1 of 0.5507, BLEU-1 of 61.5, BLEU-4 of 17.9 and ROGUE-L of 0.0824). These results are provided solely for diagnostic comparison and are excluded from task evaluation due to the use of proprietary models. However, this study indicates the potential of large models for low-resource Hindi QA, motivating exploration of open-source instruction-tuned counterparts in the future.

## 6. Conclusion and Future Work

The VATIKA 2025 Shared Task showed the difficulty of Hindi question answering in the tourism domain of Varanasi, where data scarcity and linguistic complexity limit system performance. Among the official open-source submissions, Run 2 achieved the best performance. An additional ablation with ChatGPT indicated the potential of large models for low-resource Hindi QA. These results confirm that careful refinement leads to better balance across lexical fluency, semantic coverage, and retrieval precision. Still, challenges remain. Systems struggle with domain-specific terms, long contexts, and ambiguous user queries. Future work should focus on fine-tuning multilingual transformers on Hindi tourism data, and using retrieval-augmented generation to improve context–answer alignment. Post-

processing can help make outputs more complete and fluent. Hybrid pipelines combining extractive accuracy with generative flexibility may further improve results. Incorporating structured knowledge of cultural sites can add robustness. Domain-adaptive evaluation and query expansion strategies may also raise coverage. Together, these directions can push Hindi QA toward more accurate, fluent, and user-friendly systems in specialized low-resource settings.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] 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.

[2] Y. Sun, D. Cheng, Z. Gan, X. Li, J. Liu, D. Zhou, Investigating transferability of pre-trained language models for neural question answering, arXiv preprint arXiv:1908.08962 (2019).

[3] S. Ruder, M. E. Peters, S. Swayamdipta, T. Wolf, Transfer learning in natural language processing, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials (2019) 15–18.

[4] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading wikipedia to answer open-domain questions, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), 2017, pp. 1870–1879.

[5] 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).

[6] Y. Liang, W. Ling, J. Yu, J. Lin, Q. Wang, J. Zhou, Zero-shot question answering by prompting pre-trained language models, arXiv preprint arXiv:2009.07118 (2020).

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186.

[9] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[11] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

[12] OpenAI, Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

[13] S. Chanda, K. Tewari, A. Mukherjee, S. Pal, Leveraging chatgpt and xlm-roberta for sarcasm detection in dravidian code-mixed languages, in: Proceedings of FIRE (Working Notes), Forum for Information Retrieval Evaluation, 2024, India, 2024. URL: https://ceur-ws.org/Vol-4054/T4-14.pdf.

[14] D. Kakwani, A. Ghosal, M. Shrivastava, S. Sitaram, V. Sastry, P. Talukdar, Indicnlp suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages, in: Findings of the Association for Computational Linguistics (EMNLP), 2020, pp. 4947–4958.

[15] C. Clerwall, D. Y. Tang, A survey of question answering in low-resource languages, ACM Computing Surveys 54 (2021) 1–34.

[16] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001. doi:10.18653/v1/P19-1493.

[17] J. Phang, X. Guo, K. Tran, K. Cho, English is enough! leveraging english data in code-switching language modeling, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 2021, pp. 2421–2435.

[18] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[19] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2898–2904. doi:10.18653/v1/2020.findings-emnlp.261.

[20] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. doi:10.18653/v1/D19-1371.

[21] S. Gupta, P. Malik, A. Jaiswal, S. Jha, R. Prasad, Neural-symbolic approaches in natural language processing: A survey, arXiv preprint arXiv:2105.06375 (2021).

[22] Z. Dai, Y. Sun, Y. Zhang, Q. Liu, A survey of knowledge-enhanced text generation, IEEE Transactions on Knowledge and Data Engineering 33 (2021) 3567–3584.