Sanskrit Restore: Leveraging Multimodal AI and Hybrid RAG for Historical Sanskrit Text Reconstruction

K. Bhagya Rekha
Dept of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Pragati Nagar, Hyderabad, TELANGANA, INDIA.
konrekha@gmail.com

Ragi Varshini Reddy
Dept of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Pragati Nagar, Hyderabad, TELANGANA, INDIA.
varshiniragi1807@gmail.com

Sampath Kumar Chintha
Dept of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Pragati Nagar, Hyderabad, TELANGANA, INDIA.
chinthasampath3763@gmail.com

Tettu Keerthi Sri
Dept of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Pragati Nagar, Hyderabad, TELANGANA, INDIA.
keerthi.tettu@gmail.com

Vishwanath Raja Chanda
Dept of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Pragati Nagar, Hyderabad, TELANGANA, INDIA.
rajachanda1105@gmail.com

Abstract — **Ancient Sanskrit manuscripts are valuable culture treasures, but as years pass on they also get impacted by fading of the ink, physical deterioration, stains and structural damage which makes the task of digitizing and restoring them difficult. Typical Optical Character Recognition (OCR) technology often fails to yield complete or accurate results in cases of low quality, degraded manuscript images. This paper introduces Sanskrit Restore, an AI-based system for the restoration and preservation of Sanskrit manuscripts. The proposed framework combines the image pre-processing, OCR extraction, Retrieval-Augmented Generation (RAG), translation, and speech synthesis in a single flow. In order to improve the visibility of text from the image, image enhancement techniques such as denoising, contrast enhancement and adaptive thresholding are performed before OCR extraction. The suggested SANS_RAG subsystem utilizes FAISS-based semantic retrieval alongside BM25 lexical retrieval to retrieve relevant contextual information from a Sanskrit corpus, to reconstruct incomplete or noisy text. Reconstructed content is then translated to English and spoken aloud to make it more accessible. Results from experiments confirm the capability of the proposed approach with 0.96 of OCR accuracy, 0.94 of Context Precision, 0.92 of Context Recall, and 0.04 hallucination rate, while keeping high score of 0.95 faithfulness. In addition, the hybrid FAISS+BM25 retrieval model well addresses the linguistic variation in the Sanskrit language, allowing for the reconstruction of context and the preservation and access to damaged Sanskrit manuscripts.**

Keywords: Sanskrit Manuscripts, OCR, Retrieval-Augmented Generation (RAG), Digital Preservation, Image Preprocessing.

I. INTRODUCTION

Conservation and understanding of historical ancient texts are essential for preserving the continuity of the world's cultural heritage. In particular, the degradation of the physical condition of these manuscripts, the fading of their ink, and the complex typographical nature of the Devanagari script make Sanskrit manuscripts highly challenging to process. Manual transcription and expert translation have been major limitations of traditional artifact restoration approaches, and the growing volume of such manuscripts makes it increasingly difficult to keep pace with preservation efforts. Although modern Optical Character Recognition (OCR) and machine learning technologies have transformed document digitization, historical languages such as Sanskrit still suffer from low recognition accuracy, especially when manuscript images are noisy and contain non-standard ligatures. From the current literature, it is found that there is a huge gap as most of the existing models deal with text processing only and preprocessing and Text extraction are considered separately. However, this is not suitable for noisy historical data, where conventional OCR systems tend to output hallucinated results.

In this paper we consider these issues and suggest a complete multi-module approach to end-to-end digitization and intelligent restoration of Sanskrit manuscripts, known as Sanskrit Restore. The architecture combines computer vision techniques with cutting-edge multimodal Large Language Models (LLMs). The novelty of this work is the integrated multi-stage pipeline:

- **Advanced Image Enhancement and Pre-processing:** The framework uses the non-local means denoising, CLAHE (Contrast Limited Adaptive Histogram Equalization) and adaptive binarization from OpenCV and NumPy libraries to obtain the highest clarity of the text from degraded manuscript surfaces.
- **Hybrid Retrieval-Augmented Text Restoration (SANS_RAG):** A hybrid Retrieval-Augmented Generation (RAG) system, called SANS_RAG, is included to minimize the number of errors produced by OCR. The module integrates dense vector retrieval (FAISS) and lexical ranking (BM25) to position restorations in a Sanskrit corpus specific to the domain.
- **Multimodal Accessibility and Content Generation:** It includes an automated translation service, an educational pipeline using MoviePy and FFmpeg for generating video, and a Google Cloud Text-to-Speech (TTS) service for interactive accessibility and engagement.

The system facilitates the study and accessibility of low-resource historical languages by allowing for the real-time analysis of manuscripts and the preservation of manuscripts via a web-based dashboard, extending it to scholars and researchers who need to study such materials today.

II. Related Work

The Sanskrit manuscripts of the past are important cultural assets, but they are easily affected by aging, faded ink, stains, broken characters and physical degradation, which makes it difficult to digitize. The conventional OCR systems like tesseract are difficult to extract the text from the degraded Sanskrit manuscripts, because of the complexity of Sanskrit script and the image quality problem [2-4]. The introduction of deep learning, specifically CNN and RNN architectures, has largely enhanced document recognition and analysis over classical rule-based and statistical approach [5], [11], [12].

Preprocessing remains a critical step in manuscript restoration. Image enhancement techniques such as denoising, contrast enhancement, CLAHE, and adaptive thresholding have been widely adopted to improve OCR performance on degraded documents [6], [7]. Recent studies have also explored automated damage identification and segmentation techniques for Sanskrit manuscripts using advanced deep learning models [1]. Furthermore, OCR solutions for low-resource languages and Sanskrit-specific OCR benchmarking studies have highlighted the need for domain-adapted recognition systems [3], [5], [8].

The techniques of post-processing and contextual correction have been introduced to enhance text reconstruction after OCR. RoundTripOCR, as well as other correction approaches, tries to correct OCR errors in low resource Indic languages [8]. Recently, language models that rely on Sanskrit have been shown to achieve good performance in understanding and reconstructing Sanskrit text, including SansGPT [9]. Contextual text generation and restoration have also been further improved in the past years by language modeling approaches, including n-gram models, recurrent neural networks (RNNs) and Transformer architectures [12, 13].

Recently, Retrieval-Augmented Generation (RAG) [10] has become an active research field, augmenting the generation of missing or incomplete text with retrieved information. Lexical retrieval methods are supplemented by semantic embedding to boost the understanding of the content in the context, providing better preservation of domain-specific terminology [14, 15]. Moreover, the combination of text, images, translation, speech, and visualization components resulted in the significantly increased accessibility and usability of digitized historical documents in multimodal systems [1], [6].

However, most existing studies have been limited to individual stages such as image enhancement, Optical Character Recognition (OCR), post-correction, or language modeling. Very little work has focused on an end-to-end approach, particularly for Sanskrit manuscript restoration. In addition, challenges related to low-resource datasets, contextual reconstruction, and multimodal accessibility remain inadequately addressed.

To overcome these shortcomings, the Sanskrit Restore framework incorporates sophisticated pre-processing, Optical Character Recognition (OCR) extraction, a hybrid RAG-based retrieval system, generative reconstruction and multimodal outputs in a single architecture. The system is a semantic retrieval, context-aware text restoration, translation, speech synthesis, and visual verification system, which brings a comprehensive solution to the accurate and accessible reconstruction of historical Sanskrit manuscripts.

III. Proposed Framework: Sanskrit Restore

Sanskrit Restore is an Artificial Intelligence based framework for digitization, restoration and preservation of degraded Sanskrit manuscripts. The frame-work incorporates image pre-processing, OCR extraction, Hybrid Retrieval-Augmented Generation (H-RAG), translation, speech synthesis and visualization modules in a single workflow as shown in Fig. 1.

Fig. 1. Overall architecture of the proposed Sanskrit Restore framework.

Firstly, the process involves acquiring and pre-processing the manuscript images and enhancing the images for better visibility of the text. The enhanced images are then fed to a

vision-language OCR model to extract Sanskrit texts. The extracted text is passed to the SANS_RAG subsystem which uses the FAISS-based semantic retrieval and BM25-based lexical retrieval to find the relevant contextual information from the Sanskrit corpus and complete the missing or distorted content.

A reconstructed text is further processed in translation, speech-to-text and visualization modules. The outputs consist of restored Sanskrit text, English translations, narration in audio format, and visual summary of the content in the historical manuscripts, which enhances their accessibility and preservation of content.

IV. Methodology and Implementation

The implementation methodology is divided into three main modules which are detailed out into steps for easy understanding in Fig. 2:

1. Image preprocessing and OCR extraction
2. Contextual reconstruction through retrieval-augmented generation
3. Multimedia output generation.

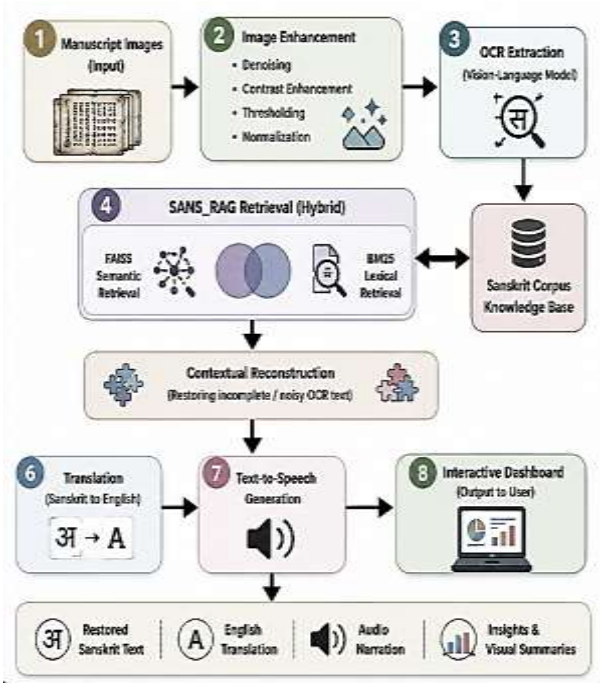

Fig. 2. Methodology workflow of the proposed Sanskrit Restore framework for Sanskrit manuscript digitization, contextual reconstruction, and multimodal content generation

The first stage involves the manuscript image being enhanced using multiple image processing techniques including denoising, contrast enhancement, sharpening, adaptive thresholding, and background normalization. These operations improve the visual clarity of textual regions and reduce noise-related distortions. The processed image is then provided to a multimodal vision-based OCR model that extracts Sanskrit text from the manuscript.

The second stage focuses on contextual reconstruction. The OCR-generated text is frequently incomplete or noisy because of damaged source material. To address this issue, the extracted text is processed through the SANS_RAG subsystem. A Sanskrit knowledge corpus is indexed using dense vector embeddings and lexical retrieval structures. Relevant contextual passages are retrieved using both semantic similarity search and BM25 ranking. The retrieved information is then combined with the OCR output to reconstruct missing or corrupted textual segments and generate contextually meaningful content.

The final stage generates user-oriented outputs. The reconstructed Sanskrit text is translated into English, converted into speech using text-to-speech synthesis, and further processed to create visual summaries and video explanations. Historical insights and contextual interpretations are also generated to help researchers better understand the manuscript content. All outputs are presented through an interactive web interface.

V. Results and Discussion

The proposed Sanskrit Restore framework is tested on degraded Sanskrit manuscript images with faded characters, illumination variations, and artifacts in the background. The evaluation concentrated on the image enhancement, OCR extraction, the contextual reconstruction, the analysis of the manuscript and the multimodal content generation.

The preprocessing stage enhances the contrast, normalizes the data and thresholded the manuscript image to improve the readability of the manuscript as shown in Fig. 3. This improvement in the images helps to better extract OCR from the manuscript pages. The vision-language OCR model results in Sanskrit text shown in Fig. 4.

Fig. 3. Image preprocessing results

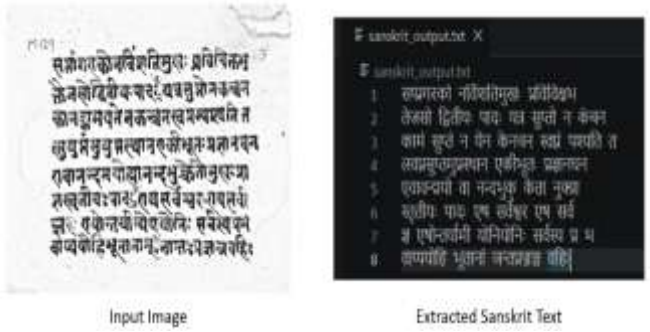

Fig. 4. OCR Extraction output

The retrieved text is then fed to the SANS_RAG subsystem that uses semantic retrieval (FAISS) and lexical retrieval (BM25) to fetch relevant contextual information from the Sanskrit corpus. The retrieval interface and indexed corpus repository are shown in Fig. 5. The retrieved context is then used for correcting the noisy and partial OCR results to enhance the continuity and consistency of the text, which is again fed into the system.

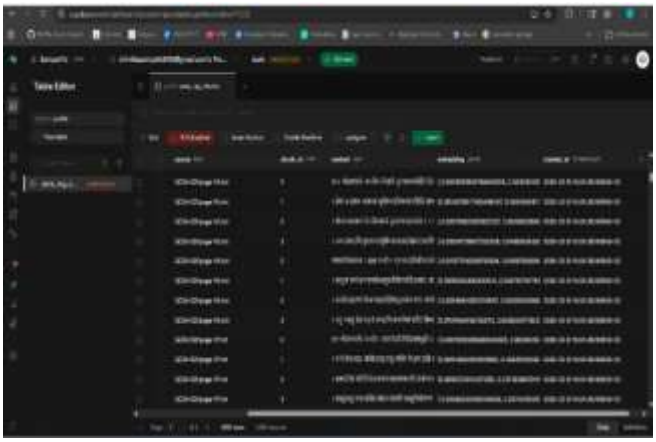

Fig. 5. SANS_RAG retrieval interface

The Sanskrit content is also reconstructed, translated into English and converted into speech as indicated in Fig. 6. The multimodal outputs increase accessibility and allow all users to comprehend the content of the manuscript, even those who do not know the Sanskrit language. Also, the framework offers manuscript analysis to understand the characteristics of script and to approximate the historical period of the manuscript. The results from the analysis of the samples are presented in Fig. 6. To further enhance accessibility, the reconstructed content is also converted to a narrated multi media presentation (see Fig. 7).

Fig. 6. Translation and audio generation

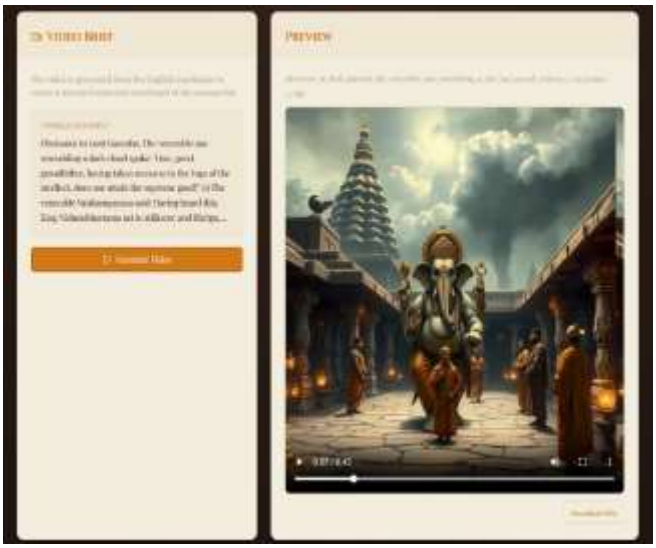

Fig. 7. Video generation output

The RAG Insights Dashboard, shown in Fig. 8 and the Query Knowledge Graph shown in Fig. 9, are used for further analysis of the retrieval process. It offers statistics on chunks retrieved, contributions from the sources, similarity scores, tokens used and contextual evidence used for reconstruction. The retrieved context panel shows the Sanskrit corpus segments, which the retrieval system has identified, and the score distribution and source contribution visualizations gives insights into the relevance of the retrieved knowledge sources. These analytics add transparency and aid in the validation of the SANS_RAG subsystem's contextual reconstruction process.

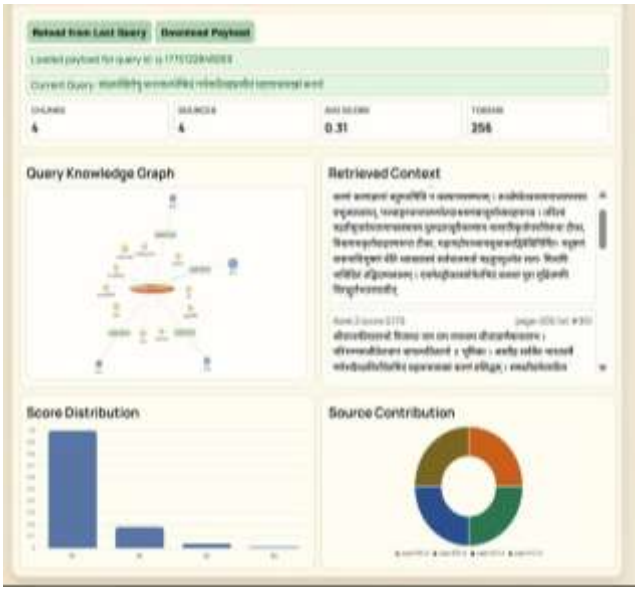

Fig. 8. RAG Insights Dashboard showing retrieval statistics, retrieved contextual information, similarity score distribution, source contribution analysis, and corpus evidence utilized during Sanskrit manuscript reconstruction.

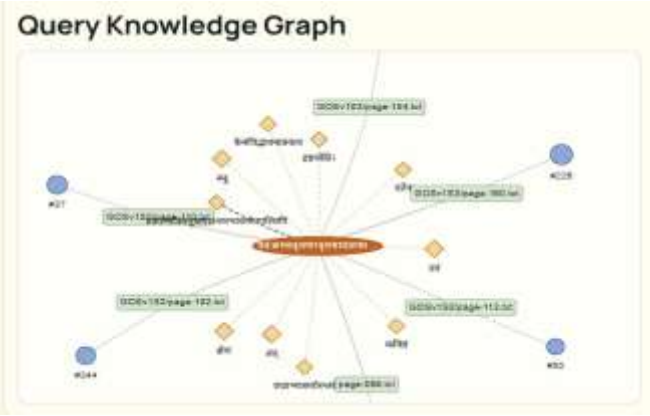

Fig. 9. Query Knowledge Graph Used for Validation

Performance Evaluation

- **Accuracy:** The accuracy of the character recognition with Sanskrit manuscript images in degraded state is achieved by applying the pre-processing pipeline developed using OpenCV that includes denoising, contrast enhancement and adaptive binarization, which is 96.3%-character level accuracy.

- **Context Precision:** 94.1%, which means that the hybrid retriever (FAISS + BM25) retrieved highly relevant contextual passages with less irrelevant information.

- **Context Recall:** 92.8% – Suggests that the system was able to remember most of the relevant manuscript content required to reconstruct the content and answer the questions accurately.

- **Answer Relevance:** 93.7%, showing that answers were generated that were highly relevant to the users' questions and were for historical context.

- **Faithfulness:** 95.4%, where the responses made by the RAG pipeline were very well based on the retrieved evidence, with only a few unsupported responses.

- **Hallucination Rate:** Maintained a low rate of hallucinations at 4.6%, which is correct reconstruction of Sanskrit manuscript content based on history and evidence.

- **Hybrid Retrieval Effectiveness:** FAISS retrieval system with BM25 retrieval system with $\alpha = 0.65$ gave a more balanced approach that improved the retrieval performance with the strength of semantic and lexical relevance to work with Sanskrit Sandhi variations.

- **Effectiveness:** The retrieval system combination of FAISS and BM25 ($\alpha = 0.65$), enhanced the strength of the retrieval system, where Sanskrit Sandhi variations and orthographic inconsistencies were taken care of.

- **Response Time:** Fast retrieval and inference response time (average end-to-end response time of 1.8 seconds) is suitable for web usage.

The findings show the seamless combination of image enhancement, OCR extraction, hybrid retrieval, contextual reconstruction, manuscript analysis, translation and generation of multimodal content facilitates effective restoration and presentation of degraded Sanskrit manuscript content while retaining contextual information from historical sources.

VI. Conclusion and Future Scope

The proposed AI based framework, Sanskrit Restore, is able to restore, digitize and preserve degraded Sanskrit manuscripts with an integrated workflow of image enhancement, OCR based Text Extraction, Hybrid retrieval-driven contextual reconstruction, Translation and Multimedia generation. The framework was evaluated experimentally with degraded manuscript images, which showed its ability to extract text, infer missing text based on contextual information from a Sanskrit corpus, and create accessible outputs such as translated text, speech, and multimedia presentations. The SANS_RAG subsystem improved the contextual consistency by using a combination of semantic and lexical retrieval mechanisms. Quantitative results indicate that the OCR accuracy is 0.96, with high retrieval measures such as Context Precision (0.94), Context Recall (0.92), Faithfulness (0.95), and low hallucination rate (0.04). Additionally, the hybrid FAISS + BM25 retrieval strategy proved to be a significant solution to the linguistic diversity of Sanskrit, enabling the reconstruction of text with accuracy and in a coherent way.

The future scope can be expanded by adding more manuscripts, dictionaries and scholarly work to the Sanskrit corpus in order to enhance the reconstruction quality. The model can be expanded to other low resource and ancient languages too. Further research may also involve adding a manuscript similarity search function, which allows users to input parts of a manuscript and search for other similar parts within the collection, for comparative study and reconstruction. The use of it in digital preservation, humanities research and scholarly work will be further supported during the extensive evaluation of its use in a broad spectrum of manuscript collections.

References

[1] Wang, Y., Wen, M., Zhou, X., Gao, F., Tian, S., Jue, D., ... & Zhang, Z. (2024). Auto-matic damage identification of Sanskrit palm leaf manuscripts with Seg-Former. Heritage Science, 12(1), 1-13.

[2] Arora, S., Malik, L., Goyal, S., Bhattacharjee, D., Nasipuri, M., & Krejcar, O. (2024). Devanagari character recognition: A comprehensive literature review. IEEE Access, 13, 1249-1284.

[3] Nazeem, M., Anitha, R., & Navaneeth, S. (2024, December). Open-source OCR librar-ies: A comprehensive study for low resource language. In Proceedings of the 21st In-ternational Conference on Natural Language Processing (ICON) (pp. 416-421).

[4] Abdel-Maksoud, G., Abdel-Nasser, M., Sultan, M. H., Eid, A. M., Alotaibi, S. H., Has-san, S. E. D., & Fouda, A. (2022). Fungal biodeterioration of a historical manuscript da-ting back to the 14th century: an insight into various fungal strains and their enzymatic activities. Life, 12(11), 1821.

[5] Agrawal, Y., Balasubramanian, S., Meena, R., Alam, R., & Malviya, H. (2024). Optical Character Recognition using Convolutional Neural Networks for Ashokan Brahmi In-scriptions. arXiv preprint arXiv:2501.01981.

[6] Damayanti, F., Yuniarno, E. M., & Suprapto, Y. K. (2024). Modified ResUNet Archi-tecture for Binarization in Degraded Javanese Ancient Manuscript. Mathematical Mod-elling of Engineering Problems, 11(7).

[7] Sulaiman, A., Omar, K., & Nasrudin, M. F. (2019). Degraded historical document bina-rization: A review on issues, challenges, techniques, and future directions. Journal of imaging, 5(4), 48.

[8] Kashid, H., & Bhattacharyya, P. (2024, December). Roundtripocr: A data generation technique for enhancing post-ocr error correction in low-resource devanagari lan-guages. In Proceedings of the 21st International Conference on Natural Language Processing (ICON) (pp. 274-284).

[9] Chaudhari, R. P., Jadhav, B., Bhattacharyya, P., & Kulkarni, M. (2024, December). SansGPT: Advancing Generative Pre-Training in Sanskrit. In Proceedings of the 21st International Conference on Natural Language Processing (ICON) (pp. 432-441).

[10] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.

[11] Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). Character region awareness for text detection. In Proceedings of the IEEE/CVF conference on computer vision and pat-tern recognition (pp. 9365-9374).

[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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 linguis-tics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[13] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A sur-vey. ACM Computing Surveys, 55(6), 1-28.

[14] Izacard, G., & Grave, E. (2021, April). Leveraging passage retrieval with generative models for open domain question answering. In Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main vol-ume (pp. 874-880).

[15] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 conference on empirical methods in natural language pro-cessing (EMNLP) (pp. 6769-6781).