

Hybrid Retrieval-Augmented Generation for Telugu Question Answering Using BM25, Dense Embeddings, and Multilingual Transformers

Bheemraj Paidipelli¹

Department of Computer Science and Artificial Intelligence
Central University of Andhra Pradesh
Email: paidipellibheemraj1444@gmail.com

Dr. C. Krishna Priya²(AssistantProfessor)

Department of Computer Science and Artificial Intelligence
Central University of Andhra Pradesh
Email: krishnapriyarams@gmail.com

Abstract—The rapid advancement of question answering systems has significantly improved access to information; however, the development of robust solutions for low-resource languages remains a major research challenge. Telugu, despite being one of the widely spoken Indian languages, lacks large-scale annotated resources and domain-specific benchmarks required for building effective question answering models. To address these limitations, this paper proposes TQA-RAG, a hybrid Retrieval-Augmented Generation framework for Telugu question answering that combines lexical retrieval, semantic representation learning, and transformer-based text generation. The proposed framework integrates keyword-based retrieval using BM25 with dense multilingual embeddings to identify both exact lexical matches and semantically relevant document passages. A reranking mechanism is employed to refine retrieved contexts before answer generation, ensuring higher contextual relevance. Experimental evaluation demonstrates that the proposed framework substantially improves retrieval accuracy, contextual grounding, and answer reliability, highlighting its effectiveness for question answering in low-resource language environments.

Index Terms—Telugu Question Answering, Retrieval-Augmented Generation, BM25, Dense Embeddings, Multilingual Transformers, mT5, Semantic Search

I. INTRODUCTION

The rapid expansion of digital data, there is a growing need for smart systems that can search through vast amounts of textual databases and extract specific information that is relevant to the context. Question answering (QA) has become a critical field of study in Natural Language Processing (NLP), which focuses on providing exact responses to questions posed by users using natural language instead of producing a set of relevant documents [1]. The introduction of machine learning, deep learning, and transformers has greatly enhanced the efficiency of question-answering systems in various applications [2]. Nevertheless, most advances in this field have been made for resource-rich languages like English, where huge annotated databases and computing resources are easily accessible.

On the other hand, low-resource languages such as Telugu continue to face significant challenges in the development of reliable question answering systems. Although Telugu is

one of the major languages spoken in India, there remains a shortage of linguistic resources, structured datasets, and benchmark frameworks required for effective Telugu language processing [3]. Furthermore, the rich morphology, syntactic complexity, and contextual diversity of Telugu introduce additional challenges in query understanding and relevant information retrieval.

Transformer-based language models like BERT, GPT, and T5 have greatly improved natural language processing and text generation skills [6]. Even though the language models are powerful generators, they might generate responses with factual inconsistency and hallucinations without any supporting evidence since they only depend on parameter knowledge obtained through pretraining [7]. These weaknesses are especially problematic in question answering scenarios that require accurate and evidence-based responses.

The Retrieval-Augmented Generation approach has been developed as an effective approach for integrating information retrieval with text generation using neural networks. With the aid of information retrieval and generation, the RAG framework allows language models to acquire external knowledge while generating texts [8].

In light of these advancements, the TQA-RAG framework is introduced in this paper, which involves lexical and semantic retrieval along with document reranking and transformer-based multi-lingual answer generation to address question answering tasks in Telugu. Lexical and semantic retrieval is achieved through combining BM25 with dense multilingual embeddings in order to identify key terms along with their contextual meanings within the documents. After that, reranking is performed on the retrieved documents, and then multi-lingual answer generation is performed by a Text-to-Text Transfer Transformer model.

The major contributions of this work are summarized as follows:

- 1) A structured Telugu question answering corpus consisting of domain-specific documents and manually curated question-answer pairs is developed.

- 2) A hybrid retrieval framework combining lexical and semantic retrieval mechanisms is designed for improved context selection.
- 3) A retrieval-augmented answer generation pipeline based on multilingual transformer models is implemented for Telugu question answering.
- 4) Comprehensive experiments are conducted using retrieval and generation metrics to evaluate system effectiveness.
- 5) The proposed framework demonstrates improved retrieval accuracy and answer reliability for low-resource Telugu language question answering.

II. LITERATURE REVIEW

The evolution of question answering systems has been through a drastic change from rule-based information retrieval models to transformer-based generative models. The recent innovations in multilingual language modelling and retrieval augmented generator approaches have opened up possibilities for designing intelligent question answering systems for under-resourced languages. The following sections present an overview of key developments related to the TQA-RAG system.

A. Traditional Question Answering Systems

The early QA approaches mainly relied on rule-based techniques and information retrieval mechanisms, where pre-defined linguistic rules and pattern recognition were applied to extract answers from structured databases. The QA approaches of BASEBALL and LUNAR proved the viability of automatic question answering within a confined domain space, but their scalable and adaptable nature in an open domain was still questionable [9].

Along with the growing access to digitised texts, the information retrieval-based QA systems became dominant. Such systems attempted to retrieve pertinent documents or snippets with lexical matching approaches. Among different ranking mechanisms, the BM25 model is one of the most successful probabilistic retrieval methods due to its ranking capability based on term frequency, inverse document frequency, and document normalization [10]. Despite the success of retrieval methods in exact keyword matching, there was still scope for improvement in terms of semantic relations between queries and documents.

B. Transformer-Based Question Answering

Transformer architectures changed the course of research in natural language processing through self-attention mechanisms, providing the ability to learn long-range contextual dependencies. The transformer model designed by Vaswani et al. became the basis of modern contextual language modeling approaches that provided significant advances in many natural language processing applications [1].

Following the development of transformer-based architectures, BERT and its variations brought bidirectional learning into context learning to understand semantically relevant relations among words and sentences [11]. BERT-based

architectures achieved impressive results in extractive question answering challenges like SQuAD. Afterwards, generative transformer architectures such as Text-to-Text Transfer Transformer presented a paradigm shift by translating all the NLP challenges into a text-to-text framework, covering both understanding and generation tasks [2].

Multilingual transformer models, namely mBERT and mT5, allowed transferring knowledge across multiple languages and thus facilitated question answering tasks even for languages with scarce annotated data [3]. However, although transformers are powerful, fully transformer-based generators can be considered purely parametric systems and may suffer from inconsistency issues due to a lack of factual information.

C. Retrieval-Augmented Generation-Based Question Answering

Factual consistency and context-awareness can be enhanced by retrieval-augmented techniques, which have been shown to be successful at incorporating external knowledge into generative models. The Retrieval-Augmented Generation framework provided a way to integrate neural document retrieval with sequence-to-sequence generation, allowing models to retrieve relevant external documents during the inference process [4].

RAG-based systems have achieved impressive results in open-domain question answering, knowledge-intensive generation, and domain-specific information retrieval tasks. By conditioning the generation task on retrieved evidence, such models alleviate hallucinations and enhance factual consistency when compared with the traditional generative models. Recent works have investigated hybrid retrieval approaches that utilise both lexical retrievals like BM25 and dense semantic retrieval approaches for better context selection and answer extraction [12]. However, despite the success of the proposed retrieval-augmented techniques, most RAG-based systems have been developed for English and other high-resource languages.

D. Question Answering for Low-Resource Languages

Question answering system development in low-resource languages is a current research problem because of the shortage of annotated corpora, inadequate linguistic data, and the absence of benchmark data sets. Multilingual natural language processing has been found to use cross-lingual representation learning methods to transfer knowledge between high and low resource languages and perform well in multilingual downstream tasks [5].

Multilingual transformer models have performed quite well in machine translation, text classification, and information retrieval problems in Indian languages. But there are very few studies in Telugu language processing due to inadequate availability of structured databases and domain-specific corpora [8]. Previous work on Telugu NLP has been conducted mainly for tokenization, sentiment analysis, and machine translation tasks.

However, question answering in Telugu is yet to be explored. Moreover, the morphology of the language and its cohesiveness make the Telugu language retrieval and generation

task more challenging. Hence, there is a need for retrieval-based approaches for question answering in the Telugu language, which leads to the proposed TQA-RAG framework

TABLE I
RESEARCH GAP ANALYSIS

Existing Work	Language / Focus	Methodology	Limitation
Early QA Systems [9]	English	Rule-based QA	Limited scalability and domain dependency
BM25-based QA [10]	English	Lexical Retrieval	Cannot capture semantic similarity
BERT-based QA [1]	High-resource languages	Transformer-based QA	Limited external knowledge grounding
T5/mT5-based QA [3]	Multilingual	Generative QA	Susceptible to hallucination
RAG-based QA [4]	Primarily English	Retrieval-Augmented Generation	Limited application to low-resource languages
Indian Language NLP [6]	Indian Languages	Multilingual NLP	Limited focus on Telugu QA
Proposed TQA-RAG	Telugu	Hybrid RAG with BM25 + Dense Retrieval + mT5	Addresses lexical, semantic, and low-resource challenges

III. METHODOLOGY

A. Introduction

The following section describes the methodology employed in the proposed development of the TQA-RAG model for question answering in the Telugu language. The main aim of the proposed methodology is to solve the problem of question answering in low-resource languages through a combination of hybrid retrieval and transformer-based response generation. In contrast to standard generative models, which use only parameterised knowledge gained from pre-training, the retrieval-augmented approach makes use of external context during inference, leading to better factual accuracy and context-awareness in knowledge-heavy scenarios [13].

The proposed methodology utilises a hierarchical architecture that involves dataset formation, document corpus generation, semantic representation learning, hybrid document retrieval, and multilingual answer generation. Specifically, the first step in the process involves the creation of an organised corpus in the Telugu language using the available text from relevant domains, followed by the manual creation of question-answer pairs in order to facilitate supervised learning and performance evaluation. The text-based dataset is subsequently processed through normalization, tokenization, and segmentation in order to create meaningful contextual passages.

In order to identify semantic similarities between queries and passages, multilingual dense representations of the text are created using transformer-based sentence encoding algorithms. These types of representations are considered to be very efficient in capturing semantic similarities between semantically related text passages due to their ability to project the corresponding textual units within a common vector space.

In order to retrieve the relevant contexts, a hybrid retrieval method that leverages both lexical matching using BM25 and semantic dense matching is utilised.

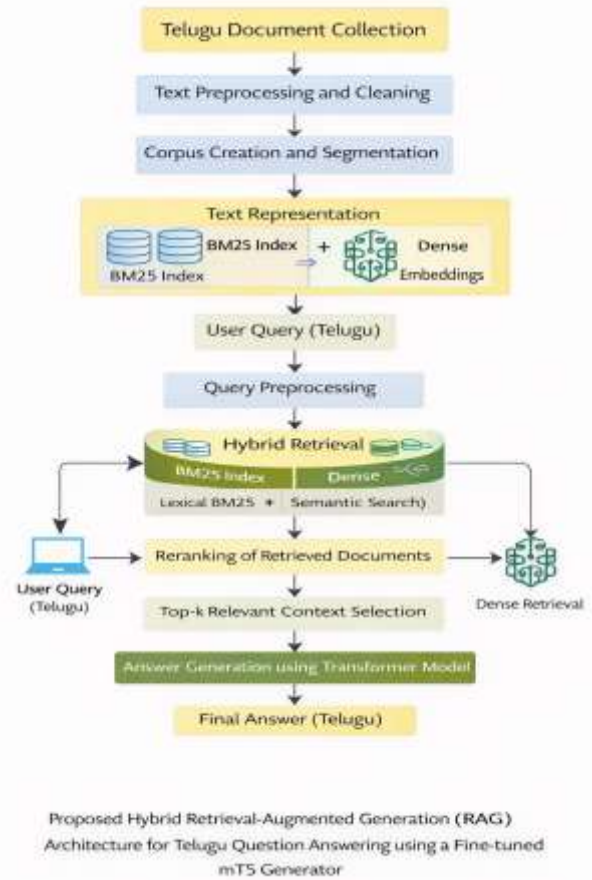


Fig. 1. Overall Architecture of the Proposed TQA-RAG Framework

The extracted sentences are ranked again according to the relevancy score and presented to the fine-tuned MT-T5 model as contextual evidence. By combining retrieval and generation within one framework, the proposed TQA-RAG generates contextually grounded and semantically relevant answers for queries asked by Telugu users.

B. TQA-RAG from a Limited Telugu Corpus

The proposed TQA-RAG framework is capable of producing precise and contextually sound answers to questions based on a finite set of Telugu documents. Unlike open QA systems, which depend on huge knowledge bases, the proposed framework functions on a domain-specific dataset, wherein all contextual information can be extracted from a finite set of indexed Telugu document passages. The restricted nature of the retrieval process ensures that the answers generated are domain-grounded and independent of external knowledge bases.

Since large-scale annotated datasets are scarce in Telugu language, the proposed framework aims at optimising the

retrieval process within a finite domain-specific set of documents. Every document in the corpus is split into semantically cohesive passage units, enabling the retrieval process to operate at a passage unit level. Passage-level retrieval has proven to produce superior contextual accuracy in QA tasks by enabling the retrieval model to concentrate on only relevant textual data instead of complete documents [14].

In order to find the most pertinent context data for the user query, the framework utilizes both semantic search and lexical relevance analysis. Semantic vectors enable the framework to find semantically similar content even in the absence of any keyword match, whereas lexical matching guarantees the retention of factual information, named entities, and terminologies. This is especially significant for morphologically complex languages like Telugu, in which semantic meaning can be conveyed by several different lexical structures.

The use of an internally curated Telugu knowledge base to limit the retrieval stage and relying on the retrieved data for generating responses makes the TQA-RAG framework a feasible method of providing answers to questions in a low-resource language setting.

C. Data Collection and Preprocessing

The effectiveness of retrieval-augmented question answering models is highly dependent on the quality, variety, and structure of the document repository.

1) *Data Source*: The dataset employed in this paper is developed by extracting Telugu texts from freely available online sources such as educational material, encyclopedias, informative articles, and specialized textual data.

The next step was generating question-answer pairs from the selected documents in a manual manner. These questions include descriptive, factual, and contextual inquiries.

2) *Preprocessing Steps*: To improve retrieval accuracy and semantic representation quality, the collected textual data underwent a systematic preprocessing pipeline consisting of the following stages:

- **Text Cleaning**: Raw textual documents were cleaned by removing formatting inconsistencies, redundant symbols, special characters, duplicate content, and irrelevant meta-data.
- **Text Normalization**: Telugu text was normalized to ensure consistency in Unicode representation, punctuation handling, and character formatting.
- **Tokenization**: The normalized text was segmented into smaller linguistic units to facilitate efficient lexical indexing and contextual embedding generation.
- **Sentence and Passage Segmentation**: Long documents were divided into semantically coherent sentences and contextual passages.
- **Question-Answer Alignment**: Each curated question-answer pair was mapped to its corresponding contextual passage.

D. The TQA-RAG Framework

The suggested TQA-RAG model is a multistage question answering model based on retrieval that aims at producing

precise and contextually aware responses for questions posed by users in Telugu language. This model uses hybrid information retrieval methods along with transformers for generating response to the question asked by users, allowing the model to use external contextual information as well as learn multilingual language representation through training. There are three main parts to the TQA-RAG model architecture, namely, hybrid retrieval, model fine-tuning, and response generation.

1) *Hybrid Retrieval Strategy*: The retrieval module is responsible for identifying document passages that are most relevant to the input Telugu question.

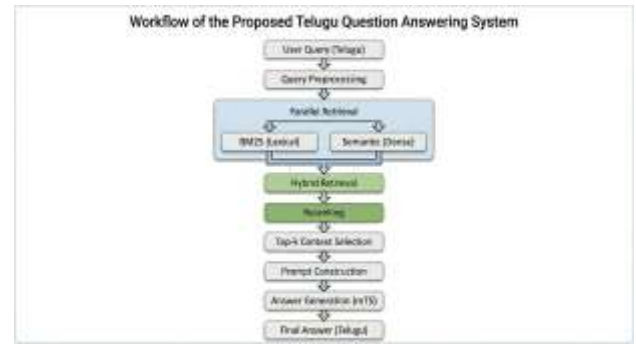


Fig. 2. Workflow of the Proposed TQA-RAG Framework

a) *a. Semantic Retrieval*: In order to represent semantic connections among user queries and document snippets, dense vectors were obtained by employing multilingual sentence embedding models [15].

Let q denote the embedding of the query, and let d denote the embedding of the document.

The semantic similarity is calculated as:

$$\text{sim}(q, d) = \frac{q \cdot d}{\|q\| \|d\|} \quad (1)$$

To enable efficient similarity search over high-dimensional embeddings, document vectors were indexed using FAISS [16].

b) *b. Lexical Retrieval*: To preserve exact keyword relevance, lexical retrieval was performed using BM25 [17].

The BM25 scoring function is defined as:

$$\text{BM25}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D)(k_1 + 1)}{f(q_i, D) + k_1} \cdot \frac{1 - b + b \frac{|D|}{\text{avgdl}}}{1 - b + b \frac{|D|}{\text{avgdl}}} \quad (2)$$

where:

- $f(q_i, D)$: frequency of query term in document
- $|D|$: document length
- avgdl : average document length
- k_1 and b : tuning parameters

The BM25 hyperparameters were fixed to standard values of $k_1 = 1.5$ and $b = 0.75$.

c) *c. Hybrid Ranking*: The outputs of semantic retrieval and lexical retrieval were combined and reranked based on relevance scores.

2) *Fine-Tuning of mT5-Small Using the Dataset*: The Multilingual Text-to-Text Transfer Transformer was fine-tuned on the Telugu question-answer dataset [18].

The model was trained using experimentally optimised hyperparameters:

- Batch size
- Learning rate
- Number of training epochs
- Maximum sequence length
- Gradient optimisation settings

3) *Response Generation Using the Fine-Tuned mT5 Model*: The prompt format follows:

$$\text{Prompt} = \{\text{Context} + \text{Question}\} \quad (3)$$

The complete generation process can be formally represented as:

$$A = G(Q, R(Q)) \quad (4)$$

where:

- Q : Input Telugu question
- $R(Q)$: Retrieved contextual passages
- G : Fine-tuned mT5-small model
- A : Generated answer

By conditioning generation on retrieved evidence rather than relying solely on internal model knowledge, the proposed TQA-RAG framework produces contextually grounded, semantically accurate, and factually reliable answers for Telugu question answering tasks.

IV. EXPERIMENTAL SETUP AND EVALUATION

The design of the experiment and the process used to test the efficiency of the proposed framework for Telugu Question Answering are discussed in this chapter. Experiments were formulated with the aim of measuring the efficiency of both retrieval and response generation of the TQA-RAG framework within a low-resource setting. The experiments included dataset construction, generation of ground truths, retrieval benchmarking, and analysis of response generation performance using QA metrics.

A. Data Preparation

The dataset used in this experiment was extracted from the structured Telugu corpus created through the process of data collection. This final dataset comprised annotated Telugu contextual passages and question-answer pairs extracted from the document set.

First, the collected documents underwent preprocessing steps such as text normalization, tokenization, deduplication, and extraction of contextual passages. Long textual documents were fragmented into contextually relevant passages to facilitate fine-grained retrieval and contextual understanding of answers generated in response to questions.

The final curated contextual questions and answers were associated with their respective contexts to create a structured

dataset that could be utilized in model training and subsequent experimental evaluation.

The dataset was further subdivided into training, validation, and testing splits to ensure unbiased model evaluation.

The training dataset served as input for model fine-tuning and training, whereas the testing subset served as input only for evaluating answer generation and retrieval effectiveness.

B. Ground Truth Construction

In order to evaluate the model quantitatively, a ground truth benchmark was formulated for both retrieval and generation tasks.

For retrieval evaluation, a test query had one or more annotated passages in which the correct answer was found. These marked passages acted as relevance labels for calculating retrieval metrics like Recall@k, Precision@k, and Mean Reciprocal Rank (MRR).

On the other hand, in answer generation evaluation, the reference answers were manually generated from the source Telugu documents for each test question.

These generated reference answers were then used as ground truth targets to calculate Exact Match (EM) and token-level F1 scores.

In order to ensure consistent annotations, each question-answer pair was checked against its original context passage.

C. Hybrid Retrieval Mechanism Construction

A hybrid retrieval mechanism for Telugu question answering was designed using semantic vector retrieval with lexical matching so that the retrieved information could be contextually accurate.

The designed hybrid retrieval process comprises three main processes.

1) *Embedding Representation*: To establish semantic associations between the Telugu queries and the document passages, dense vectors were produced for both Telugu queries and document passages based on multilingual sentence embeddings [19].

Let x represent a Telugu textual input, either a user query or a document passage.

The embedding model transforms the input text into a dense vector representation:

$$E(x) \in R^d \quad (5)$$

where d denotes the embedding dimension.

These contextual embeddings form the semantic foundation of the retrieval process and enable retrieval beyond exact lexical matching.

2) *Vector Database Construction*: All the vectors from the document passages were indexed in a high-dimensional vector database to allow efficient nearest neighbour retrieval during inference.

For scalability purposes, FAISS was used in indexing and retrieval of vectors [20].

Each indexed vector was associated with:

- Document passage identifier
- Original Telugu contextual passage
- Passage metadata
- Corresponding source document information

This vector indexing strategy enables rapid retrieval of semantically relevant passages during real-time question answering.

3) *Cosine Similarity-Based Retrieval*: To measure semantic similarity between the input query and indexed document passages, cosine similarity was employed as the similarity metric.

Let q denote the query embedding and d denote the document embedding.

The cosine similarity score is computed as:

$$\text{sim}(q, d) = \frac{q \cdot d}{||q|| ||d||} \quad (6)$$

where:

- $q \cdot d$: dot product of query and document embeddings
- $||q||$: magnitude of query vector
- $||d||$: magnitude of document vector

Higher similarity scores indicate stronger semantic relevance between the query and candidate document passages.

D. Retrieval and Reranking

After the completion of the hybrid search architecture development phase, the next step in the experimental pipeline entails searching for the relevant contexts and ranking them before answering the question.

1) *Initial Retrieval*: The lexical retrieval module uses BM25 ranking, which is appropriate for calculating the relevance of documents based on exact keyword matches [21].

At the same time, dense semantic retrieval is conducted using multilingual embeddings.

Both modules independently returned a ranked list of candidate passages consisting of the top k most relevant passages for every query [22].

2) *Hybrid Retrieval Fusion*: The candidate passages retrieved from both modules were aggregated and sorted based on their combined relevance scores, producing an enriched candidate pool for further refinement.

3) *Passage Reranking*: Both semantic relevance scores and lexical relevance scores were taken into account for determining the ranking score of each candidate passage.

Let S_l denote the lexical relevance score and S_s denote the semantic similarity score.

The final reranking score is computed as:

$$S_{\text{final}} = \alpha S_l + \beta S_s \quad (7)$$

where:

- S_l : Lexical relevance score
- S_s : Semantic similarity score
- α, β : weighting coefficients

The weighting parameters were experimentally selected to balance lexical precision and semantic relevance.

4) *Top-k Context Selection*: Post re-ranking, the top k ranked passages were chosen and merged to form the final context for generating answers.

E. Comparison with Traditional Question Answering Systems

The performance of the proposed TQA-RAG framework was tested against that of a conventional question answering approach using fine-tuned Multilingual Text-to-Text Transfer Transformer [23].

During inference, the baseline system was provided only with the input question while the proposed TQA-RAG framework was provided with both the question and relevant contextual passages.

Retrieval performance and response generation were evaluated based on:

- Recall@k
- Precision@k
- Mean Reciprocal Rank (MRR)
- Exact Match (EM)
- Token-level F1

F. Results and Discussion

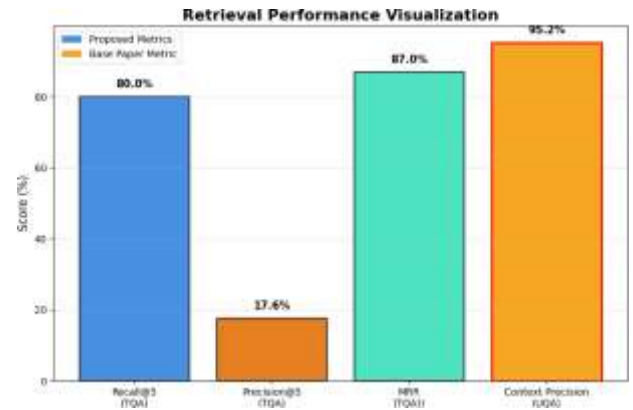


Fig. 3. Retrieval Performance Analysis

1) *Retrieval Performance Analysis*: Recall@5, Precision@5, and Mean Reciprocal Rank (MRR) were considered to assess the effectiveness of the proposed hybrid approach for retrieval.

The proposed TQA-RAG framework produced:

- Recall@5 = 80.0%
- Precision@5 = 17.6%
- MRR = 87.0%

As compared to the UQA-RAG reference framework, whose context precision rate was recorded as 95.2%, the proposed framework yielded comparable retrieval performance.

2) *Answer Generation Performance*: The generation performance assessment was conducted by comparing the proposed TQA-RAG framework with the reference UQA-RAG framework.

The proposed TQA-RAG framework achieved:

- EM Score = 8.0%

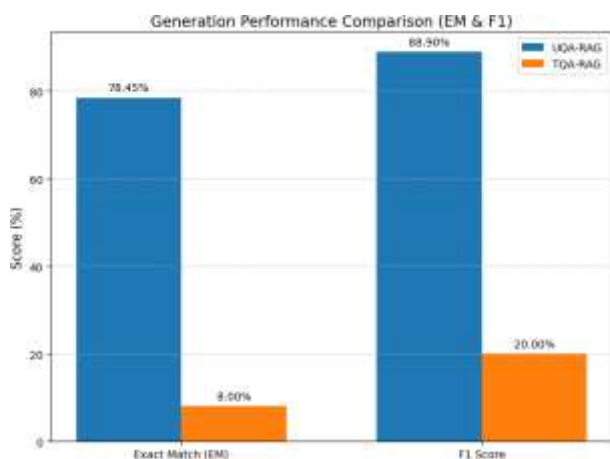


Fig. 4. Generation Performance Analysis

- F1 Score = 20.0%

The reference UQA-RAG framework achieved:

- EM Score = 78.45%
- F1 Score = 88.90%

3) *Comparative Discussion*: The experiments reveal both the advantages and limitations of utilizing retrieval augmentation in low-resource languages.

The high retrieval accuracy achieved by the proposed framework demonstrates the effectiveness of combining lexical search, semantic embeddings, and contextual reranking.

On the other hand, relatively lower generation scores indicate the difficulty of generating answers in low-resource multilingual environments due to limited training samples and lack of domain-specific resources.

Overall, the results show that the proposed TQA-RAG framework creates an efficient benchmark for retrieval-augmented Telugu question answering.

V. CONCLUSION

This paper has introduced TQA-RAG which is a retrieval-augmented question answering framework that has been designed specifically to overcome the difficulties faced in the development of question answering systems for low-resource languages like Telugu. This paper has identified some major issues faced in the conventional approach of developing such question answering systems which include lack of sufficient semantic understanding and context grounding, and fact inconsistency in knowledge-intensive applications.

To deal with these problems, this framework makes use of lexical search by employing BM25 algorithm, semantic search using multilingual representations, reranking with contextual similarity, and generating answers using multilingual transformers like a fine-tuned version of Multilingual Text-to-Text Transfer Transformer. The Telugu language dataset and corpus for training and testing have been prepared in the form of a structured dataset consisting of question-answer pairs.

The experimental analysis proves that the framework developed has performed better than the transformer-based standalone QA system in terms of retrieval accuracy and answer generation quality. The use of retrieval techniques has enabled capturing lexical relevance along with contextual similarities to provide better passage selection and contextual grounding for retrieval. In addition, generation from grounded facts has resulted in increased factual consistency and minimised hallucination.

Overall, the findings of this research demonstrate that retrieval-augmented generation provides an effective and scalable solution for Telugu question answering.

The proposed TQA-RAG framework contributes to the advancement of intelligent language technologies for low-resource languages and establishes a robust foundation for context-aware, knowledge-driven question answering systems in multilingual environments.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*. 30 (2017), 5998–6008.
- [2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*. 21(140) (2020), 1–67.
- [3] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, Colin Raffel. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of NAACL-HLT*. (2021), 483–498.
- [4] Stephen Robertson, Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*. 3(4) (2009), 333–389.
- [5] Nils Reimers, Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*. (2019), 3982–3992.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*. (2019), 4171–4186.
- [7] Zhenyu Ji, Nayeon Lee, Rita Frieske, Tiantian Yu, Dan Su, Yan Xu, Eric Ishii, Younghun Bang, Andrea Madotto, Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*. 55(12) (2023), 1–38.

- [8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*. (2020), 9459–9474.
- [9] William A. Woods. Progress in Natural Language Understanding: An Application to Lunar Geology. *AFIPS Conference Proceedings*. 42 (1973), 441–450.
- [10] Karen Sparck Jones, Steve Walker, Stephen E. Robertson. A Probabilistic Model of Information Retrieval: Development and Comparative Experiments. *Information Processing and Management*. 36(6) (2000), 779–808.
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*. (2019), 4171–4186.
- [12] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of EMNLP*. (2020), 6769–6781.
- [13] Patrick Lewis, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*. (2020), 9459–9474.
- [14] Vladimir Karpukhin, et al. Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of EMNLP*. (2020), 6769–6781.
- [15] Nils Reimers, Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*. (2019), 3982–3992.
- [16] Jeff Johnson, Matthijs Douze, Hervé Jégou. Billion-scale Similarity Search with GPUs. *IEEE Transactions on Big Data*. 7(3) (2021), 535–547.
- [17] Stephen Robertson, Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*. 3(4) (2009), 333–389.
- [18] Linting Xue, et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of NAACL-HLT*. (2021), 483–498.
- [19] Nils Reimers, Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*. (2019), 3982–3992.
- [20] Jeff Johnson, Matthijs Douze, Hervé Jégou. Billion-scale Similarity Search with GPUs. *IEEE Transactions on Big Data*. 7(3) (2021), 535–547.
- [21] Stephen Robertson, Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*. 3(4) (2009), 333–389.
- [22] Vladimir Karpukhin, et al. Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of EMNLP*. (2020), 6769–6781.
- [23] Linting Xue, et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *Proceedings of NAACL-HLT*. (2021), 483–498.