BHRE-RAG: A Benchmark and Retrieval-Augmented Framework for Advancing Comprehension-Based Question Answering in Bangla

Md Saiyem Raiyan * and Nayeema Ferdous *

Posted Date: 23 January 2026

doi: 10.20944/preprints202601.1821.v1

Keywords: Bangla question answering; retrieval-augmented generation; large language models; low resource languages; BanglaRQA; zero-shot learning; few-shot learning

 doi:10.20944/preprints202601.1821.v1

Article

BHRE-RAG: A Benchmark and Retrieval-Augmented Framework for Advancing Comprehension-Based Question Answering in Bangla

Md Saiyem Raiyan * and Nayeema Ferdou *

Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh
* Correspondence: saiyem.raiyan@northsouth.edu (M.S.R.); nayeemaferdous987@gmail.com (N.F.)

Abstract

Large language models excel in English but struggle with low-resource languages such as Bengali due to limited training data and complex linguistic structures. This paper represents a groundbreaking novel system that applies state-of-the-art natural language processing techniques to examine subject-specific chapters and generate questions with corresponding solutions of different lengths and addresses this gap through two key contributions. First, we introduce the Bangla Holistic Reasoning Evaluation (BHRE), a comprehensive zero-shot and few-shot assessment of LLMs (GPT-4, Llama-3.1, Mixtral-8x, Qwen2.5, Mistral, Gemma) on the challenging BanglaRQA dataset. Second, we propose a Retrieval Augmented Generation (RAG) framework with BHRE that enhances LLM performance by retrieving precise and contextual evidence before generating answers. Using the BanglaRQA question-answering dataset, comprising 3,000 context passages and 14,889 question-answer pairs, we benchmarked these LLMs using EM and F1 metrics against BanglaT5, a fine-tuned state-of-the-art model. Our results show that Llama-3 emerged as the top performing model in F1 and EM, while our RAG-based approach elevates its performance much better than the other two approaches, surpassing the previous fine-tuned SOTA (BanglaT5). This work demonstrates that prompt engineering techniques on LLMs can rival fine-tuned models, achieving top-notch quality answers even without fine-tuning, and the effectiveness of RAG systems for low-resource languages, and provides a reproducible framework for future research by enhancing the capability of language models.

Keywords: Bangla question answering; retrieval-augmented generation; large language models; low resource languages; BanglaRQA; zero-shot learning; few-shot learning

1. Introduction

Reading comprehension QA tasks have long served as benchmarks for natural language understanding. Datasets such as SQuAD [1] and HotpotQA [2] have driven significant progress in English. However, low-resource languages such as Bengali remain under-served, despite having over 237 million native speakers across Bangladesh and India [3]. Challenges including inefficient tokenization, limited high-quality training data, and biases in machine-translated corpora hinder model performance. Although techniques such as Chain-of-Thought prompting improve reasoning, they often struggle with factual question answering [4]. To address this gap, the BanglaRQA dataset [5] introduced 14,889 question–answer pairs across four categories. Baseline experiments using fine-tuned models such as BanglaT5 [6] and mT5 [7] achieved moderate success, but revealed weaknesses in multi-span and causal reasoning.

To overcome these challenges, we present a novel system that advances the state of understanding of bangla language by applying cutting-edge NLP techniques to subject-specific chapters for generating questions and their detailed answers. Specifically, we introduce the Bangla Holistic Reasoning Evaluation (BHRE), a zero shot benchmark designed to assess the performance of state-of-the-art LLMs on the BanglaRQA dataset.The BanglaRQA dataset presents a formidable challenge due

 doi:10.20944/preprints202601.1821.v1

to the complex semantics and various types of questions in Bengali, making it an ideal benchmark for evaluating language models.The primary objective of this research aims to implement an effective Bangla Holistic Reasoning Evaluation (BHRE), an evaluation framework benchmarking multiple advanced LLMs, including GPT-4-Turbo[8], Meta-Llama-3.1-8[9], Qwen2.5-72B-Instruct[10], Mixtral-8x22B-Instruct-v0.1[11], Google-Gemma-7B[12], Mistral-7B-v0.1[13] in the challenging BanglaRQA dataset to reveal nuanced model capabilities and limitations. The BHRE evaluation focuses on two critical metrics:The Exact Match(EM) and the F1 score, which are crucial for assessing the accuracy and quality of model-generated answers. A zero-shot and few-shot learning framework was developed to assess the performance of Large Language Models (LLMs) in specific contexts, such as Bangla. However, some LLMs, like Mistral-7B and Qwen2.5-72B, may not accurately represent Bangla due to their English design. To address this issue, we implement the Retrieval-Augmented Generation (RAG) system using a direct zero-shot method, ensuring the effectiveness of the models in addressing specific contexts.

Retrieval-Augmented Generation (RAG) frameworks, which combine information retrieval with generative models [14], have demonstrated significant improvements in performance for low-resource languages by integrating contextual evidence before answer generation[15]. By incorporating RAG with BHRE, we demonstrate an enhanced performance of LLMs, as evidenced by the improvement in both EM and F1 scores, surpassing previously fine-tuned models such as BanglaT5. Our experimental results show that, in a few-shot setting, the Llama-3 model achieved good F1 and EM scores. After applying the RAG framework, we achieved more improvement on our metrics. These findings affirm that prompt engineering combined with retrieval augmented generation unlocks powerful capabilities in LLMs for low-resource languages without requiring extensive fine-tuning, presenting a reproducible framework for further research.

2. Literature Review

Research on Large Language Models (LLMs) has expanded across numerous domains, including question answering (QA) systems, which have been applied to languages like Bangla. Despite notable achievements in deploying LLMs for diverse tasks, the literature specifically addressing LLM-based Bangla QA systems remains limited. Nevertheless, broader investigations into the role of LLMs in QA and natural language processing highlight several key developments and challenges relevant to Bangla QA. While LLMs have demonstrated strong performance across a variety of biomedical tasks [16], they have also faced criticism for exhibiting unfairness in broader NLP applications [17].

Rony et al.[18] created BanglaQuAD as an open-domain question-answering dataset for the Bengali language, with the objective of enhancing the development and assessment of question-answering systems in this language. The dataset comprises 30,808 question-answer pairs derived from Bengali Wikipedia pages, encompassing a range of question categories, including factual, descriptive, and explanatory. Veturi et al.[19] presented an innovative approach to question answering by integrating RAG with LLMs to enhance the prediction of contextually relevant responses. This method improves answer relevance and reduces hallucinated information compared to traditional BERT-based models. Faieaz et al.[20] utilized the Bangla T5 base model, a transformer-based architecture designed specifically for the Bangla language, to generate contextually appropriate questions from Bangla text.

Recent efforts to develop large language models (LLMs) for Bangla remain relatively limited compared to high-resource languages. Raihan et al.[3] introduced TigerLLM, a family of Bangla LLMs trained with two major resources: Bangla-TextBook Corpus and Bangla-Instruct Dataset, consistently outperforming GPT-3.5. Ipa et al.[21] introduced the TraSe architecture, a novel framework designed to enhance Retrieval-Augmented Generation (RAG) for Bangla. The framework uses a small dataset of 200 question-answer pairs and translative prompts to improve the accuracy of the retrieval and contextual generation. The experimental results show that TraSe outperforms the baseline RAG approaches, achieving accuracy up to 63%. Shafayat et al.[22] created BEnQA: The bilingual Bengali-English QA and reasoning dataset contains 5,161 multiple choice questions from Bangladesh's national curricu-

 doi:10.20944/preprints202601.1821.v1

lum for grades 8, 10, and 12, categorizing them as factual, application-based, and reasoning-based. The study on Answer-Agnostic Question Generation (AQG) Fahad et al. [23] in Bangla proposed an automatic question generation method where questions are generated directly from a passage without requiring predefined answers. To achieve this, three pre-trained models: BanglaT5, mT5-base, and BanglaGPT2 were fine-tuned. Roy et al. [24] explored the potential of fine-tuning a pre-trained language model, BERT-Bangla, for a closed-domain QA task,using a dataset of 2,500 question-answer pairs from the Khulna University of Engineering and Technology (KUET) website. Hasan et al.[25] evaluated Bangla-Bayanno, a large-scale Bengali Visual Question Answering (VQA) dataset designed to support the development of AI models capable of understanding and answering questions about images in Bengali. With over 52,000 image-question-answer pairs, it aims to bridge the gap in VQA resources for Bengali. The authors refined translations using Large Language Models to ensure grammatically correct and culturally relevant Bengali text.

3. Methodology

The primary goal of this study is to examine how effectively large language models (LLMs) generate accurate single-span and multi-span answers from a given context. We follow a structured pipeline built around the BanglaRQA dataset and evaluate performance across multiple LLMs. This section describes our question–answering approach.

3.1. Dataset Overview

BanglaRQA is the first large-scale benchmark for Bangla reading comprehension and question answering. Sourced mainly from Bangla Wikipedia, it spans topics such as politics, history, culture, science, entertainment, and current affairs. The dataset contains 14,889 question–answer pairs, including 3,631 unanswerable questions, with an average context length of about 215 words. Table 1 presents an overview of the dataset. It is structured around three core components:

- **Context Passages:** The foundation comprises 3,000 text passages from Bangla Wikipedia, carefully selected for broad domain coverage, and cleaned to maintain quality and consistency by removing hyperlinks, citations, and non-Bangla text.
- **Question Types:** The passages generated 14,889 questions, categorized into four types to address diverse challenges.
  - Factoid: Questions that ask for specific facts, often beginning with interrogatives like ”কী (What), "কে (Who)," "কখন (When)," "কোথায় (Where)," or "কোনটি (Which)"
  - Causal: Questions that inquire about reasons or methods, typically starting with "কেন (Why)" or "কিভাবে (How)"
  - Confirmation: Questions that can be answered definitively with either "হ্যাঁ (Yes)" or "না (No)" often requiring inference.
  - List: This type of question contains keywords like কী কী/কোনগুলো (What are), কারা কারা (Who are), etc.

Table 1. BanglaRQA Dataset Stats

Context passages	3000
Question-answer pairs	14889
Average word count	215
Factoid (69.8%)	10388
Casual (9.6%)	1433
Confirmation (10.3%)	1531
List (10.3%)	1537

 doi:10.20944/preprints202601.1821.v1

Questions that require enumerating multiple items or facts from the passage. Crucially, 3,631 of these questions are unanswerable, meaning the context passage does not contain the information required to answer them.

- **Answer Types:** Answer collection was done by human annotators with 4 categories of answers respected to the questions. For answerable questions, the answers fall into three formats:
  - Single-Span: A single, contiguous segment of text extracted directly from the passage.
  - Multiple Spans: An answer composed of several text segments from different parts of the passage, combined and separated by semicolons.
  - Yes/No: A direct binary response of either"হ্যাঁ (Yes)" or "না (No)".

Figure 1 shows the sample Bangla RQA dataset which demonstrates the corresponding questions and answers.

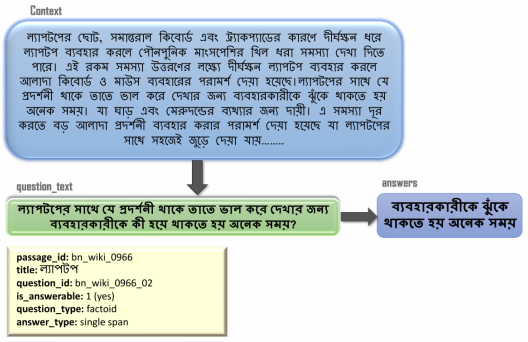

Figure 1. Sample BanglaRQA dataset. *English translation: The figure shows a Bangla context passage followed by a question and its corresponding answer, illustrating how questions are constructed and how answers are extracted directly from the context.*

3.2. Models

We evaluated six LLMs for assessment that are categorized on various models:

- **GPT-4-Turbo:** Developed by OpenAI, this model is an advanced GPT iteration that leverages deep contextual analysis to generate precise, intent-aware responses across diverse linguistic tasks.
- **Meta-Llama-3.1-8B:** This compact model from Meta features 8 billion parameters. Key specifications include 40 attention heads, an expanded vocabulary size of 128,000 tokens.
- **Qwen2.5-72B-Instruct:** Alibaba's substantial language model incorporates 72 billion parameters specifically tuned for instruction-based tasks.
- **Mixtral-8x22B-Instruct-v0.1:** Mixtral uses eight specialized components with 22 billion parameters, handling diverse queries through parallel processing pathways, and ensuring quality response generation through distinct capabilities for question analysis.
- **Google-Gemma-7B:** Google's 7-billion parameter model provides balanced performance in language understanding and generation tasks. The specialization in domain-specific sectors is made possible by its more compact architecture compared to larger models.
- **Mistral-7B-v0.1:** This foundational model from Mistral AI (7 billion parameters) serves as a capable base for various language processing applications.

3.3. Prompt Design and Evaluation

We designed a zero-shot and few-shot learning framework to benchmark the performance of Large Language Models. Using this approach, we define how well these models work for domain-specific tasks without fine-tuning for question-answering across different question types. We noticed that some LLMs cannot clearly hold context in Bangla, as they are primarily designed for English. As a result, we instructed them to generate responses in Bangla. However, some models (e.g., Mistral-7B, Qwen2.5-72B) sometimes produce out-of-context answers, especially for list-type or multi-span

 doi:10.20944/preprints202601.1821.v1

questions, either omitting or mixing list items. This makes it difficult to determine the correct list answer. To address this, we designed our prompt to ensure accurate and concise Bangla responses. We also implemented a zero-shot RAG system to strengthen evidence. Finally, we used the following organized prompt to obtain precise outputs from the LLMs:

Prompt (zero-shot): "প্রদত্ত প্রসঙ্গটি পড়ুন এবং শুধুমাত্র প্রসঙ্গের ভিত্তিতে প্রশ্নের উত্তর দিন। যদি উত্তর প্রসঙ্গে না থাকে, 'unanswerable' লিখুন। (English: "Read the given context and answer the question based solely on the context. If the answer is not in the context, write 'unanswerable'.") [INPUT:]"

English Translation: Read the given context and answer the question based solely on the context. If the answer is not present in the context, write "unanswerable".

Prompt (few-shot): few-shot = """ প্রসঙ্গ থেকে প্রশ্নের উত্তর দিন। উত্তর শুধুমাত্র প্রসঙ্গ থেকে নিতে হবে। উদাহরণ ১ (Factoid): প্রসঙ্গ: বাংলাদেশের রাজধানী ঢাকা। ঢাকা দেশের সবচেয়ে বড় শহর। প্রশ্ন: বাংলাদেশের রাজধানী কোথায়? উত্তর: ঢাকা

উদাহরণ ২ (Causal): প্রসঙ্গ: ১৯৭১ সালে বাংলাদেশের মুক্তিযুদ্ধ সংঘটিত হয়। দীর্ঘ বৈষম্য ও রাজনৈতিক দমননীতি এর কারণ ছিল। প্রশ্ন: বাংলাদেশের মুক্তিযুদ্ধ কেন হয়েছিল? উত্তর: দীর্ঘ বৈষম্য ও রাজনৈতিক দমননীতি

উদাহরণ ৩ (Confirmation / হ্যাঁ-না): প্রসঙ্গ: সুন্দরবন বিশ্বের সবচেয়ে বড় ম্যানগ্রোভ বন। প্রশ্ন: সুন্দরবন কি বিশ্বের সবচেয়ে বড় ম্যানগ্রোভ বন? উত্তর: হ্যাঁ

উদাহরণ ৪ (List): প্রসঙ্গ: বাংলাদেশের প্রধান নদীগুলি হলো পদ্মা, মেঘনা এবং যমুনা। প্রশ্ন: বাংলাদেশের প্রধান নদীগুলির নাম কী? উত্তর: পদ্মা ; মেঘনা ; যমুনা

উদাহরণ ৫ (Unanswerable): প্রসঙ্গ: রবীন্দ্রনাথ ঠাকুর ১৯১৩ সালে নোবেল পুরস্কার অর্জন করেন। প্রশ্ন: রবীন্দ্রনাথ ঠাকুর কোথায় সমাধিস্থ হয়েছেন? উত্তর: উত্তর পাওয়া যায়নি

এখন উত্তর দিন:

```
prompt = f"""{few-shot} প্রসঙ্গ: {context} প্রশ্ন: {question} উত্তর:"""
```

English Translation of Few-Shot Examples:

Example 1 (Factoid): Context: The capital of Bangladesh is Dhaka. Question: Where is the capital of Bangladesh? Answer: Dhaka

Example 2 (Causal): Context: The Liberation War of Bangladesh occurred in 1971 due to political oppression. Question: Why did the Liberation War of Bangladesh occur? Answer: Political oppression

Example 3 (Confirmation): Context: The Sundarbans is the largest mangrove forest in the world. Question: Is the Sundarbans the largest mangrove forest in the world? Answer: Yes

Example 4 (List): Context: The major rivers of Bangladesh are Padma, Meghna, and Jamuna. Question: What are the major rivers of Bangladesh? Answer: Padma; Meghna; Jamuna

Example 5 (Unanswerable): Context: Rabindranath Tagore won the Nobel Prize in 1913. Question: Where was Rabindranath Tagore buried? Answer: Unanswerable

3.4. Retrieval-Augmented Generation (RAG) Framework:

Here, we found that by LLMs, our proposed model (BHRE) lacks contextual answers for each type of question, especially long questions, because it can not directly hold the accurate points. Also, for multi-span questions, basic LLMs may not be able to retrieve each list type of question. To address the limitations of zero-shot performance, we developed a Retrieval-Augmented Generation (RAG) pipeline, which shows in Figure 2 This combines specific question or context search with text generation to produce accurate, fact-based responses [26] .Instead of relying solely on pre-trained knowledge, it first retrieves relevant information from external sources like databases or the web, then processes this data to generate well-supported answers. RAG with Langchain and ChromaDB [27] makes the output extraordinary and contextual. Langchain is used with the LLMs and the sentence transformer model. Langchain is an open-source framework designed to enhance the capabilities of LLMs. ChromaDB is used for creating vector databases.

We checked the model with a Hugging Face pipeline, using a query about the meaning, and split the context text into chunks. Each context passage was split into smaller chunks (e.g., sentences or

 doi:10.20944/preprints202601.1821.v1

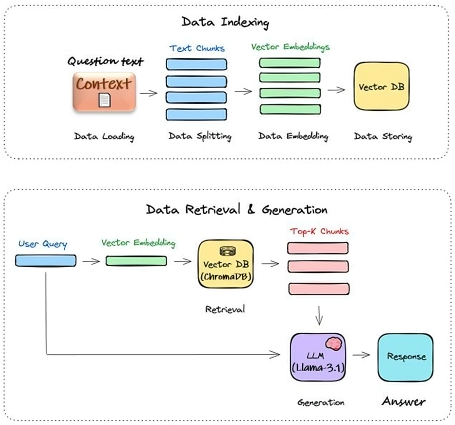

Figure 2. RAG pipeline .

short paragraphs). After that, creating embeddings and storing them in the vector store, initialize the question-answering (Q/A) chain, and test the RAG by query. A multilingual dense retriever (e.g., Sentence-BERT) encoded the question and chunks into a shared vector space. The most relevant chunk was selected based on semantic similarity. The score-generating mechanism works according to the Top-n chunks cosine similarity. If the similarity is close to 1 (e.g., 0.7, 0.8), then the context gives very accurate answers that are "answerable". If the similarity score between the question and the top chunk was below a threshold, the system returned "unanswerable." The overall mean of the similarity scores is then calculated. The cosine similarity between two vectors u and v is given by:

$$\text{Cosine Similarity } (\mu) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|} \tag{1}$$

The LLM was prompted with the retrieved chunk instead of the full context. The prompt was modified to be like a zero-shot method. As a result, the approach easily retrieves single span answers and multi span answers for each type of question, like factoid, casual, confirmation, list, or unanswerable factoid questions.

4. Results and Discussion

4.1. Performance Evaluation (BHRE and RAG)

We present the performance evaluation of various LLMs on the test-set BanglaRQA question-answering, comparing the results on zero-shot, few-shot, and our RAG system against the current SOTA fine-tuned model, BanglaT5, as shown in Table 2. For the RAG system, the mean cosine similarity (μ) on the test set.

Table 2. Performance comparison of zero-shot, few-shot LLMs, zero-shot RAG, and the fine-tuned state-of-the-art model on BanglaRQA dataset. Top results are highlighted in bold.

Model Name	Zero-shot		Few-shot		$\text{RAG}_{\text{zero-shot}}$		Cosine Similarity (μ)
	EM	F1	EM	F1	EM	F1	
GPT-4-Turbo	48.33	65.31	55.23	67.31	**65.53**	74.30	0.70
Meta-Llama-3.1-8B	46.54	67.33	**65.14**	**78.20**	**70.83**	**83.19**	0.79
Qwen2.5-72B-Instruct	39.52	45.28	52.52	64.56	**64.12**	74.00	0.64
Mixtral-8x22B-Instruct-v0.1	50.73	63.51	58.00	68.25	**65.25**	**81.11**	0.77
Google-Gemma-7B	42.33	50.57	47.91	54.87	55.33	73.51	0.66
Mistral-7B-v0.1	41.35	47.44	51.00	59.11	58.00	69.23	0.57
BanglaT5	**62.42**	**78.11**					

We evaluate the performance of the models on the BanglaRQA dataset, which is shown in Figure 3. Here, we used Exact match (EM) and F1 scores to calculate the accuracy of our work. The evaluation indicates that, without fine-tuning, LLMs can work well if they use prompt engineering techniques

 doi:10.20944/preprints202601.1821.v1

such as zero-shot and few-shot methods. The zero-shot prompt engineering techniques show poor performance in all metrics.

As our models cannot perform well with the zero-shot method, unfortunately, but for few-shot, the Meta-Llama-3.1-8B EM score performs remarkably well compared to the state-of-the-art model, BanglaT5, which is a fine-tuned model on BanglaRQA. The introduction of a few-shot example substantially improved model performance across all architectures. Meta-Llama-3.1-8B exhibited the most significant gain, improving from 46.54% to 65.14% EM (+18.6 points) and from 67.33% to 78.20% F1 (+10.87 points). This improvement pattern indicates that,in context,learning effectively compensates for the lack of task-specific fine-tuning in Bengali QA tasks.

On the other hand, we further optimized those models. For that, we used our RAG technique to get more contextual answers with high accuracy. After our implementation, the system states that GPT-4-Turbo, Meta-Llama-3.1-8B, Qwen2.5-72B-Instruct, and Mixtral-8x22B-Instruct-v0.1 outperform BanglaT5, achieving the highest scores. Meta-Llama-3.1-8B achieved the highest overall scores (70.83% EM, 83.19% F1), representing a 7.41-point EM and 5.08-point F1 improvement over the fine-tuned BanglaT5 benchmark. GPT-4-Turbo and Mixtral-8x22B-Instruct-v0.1 also showed strong RAG performance, achieving 65.53% EM, 65.25% EM, and 81.11% F1, respectively.

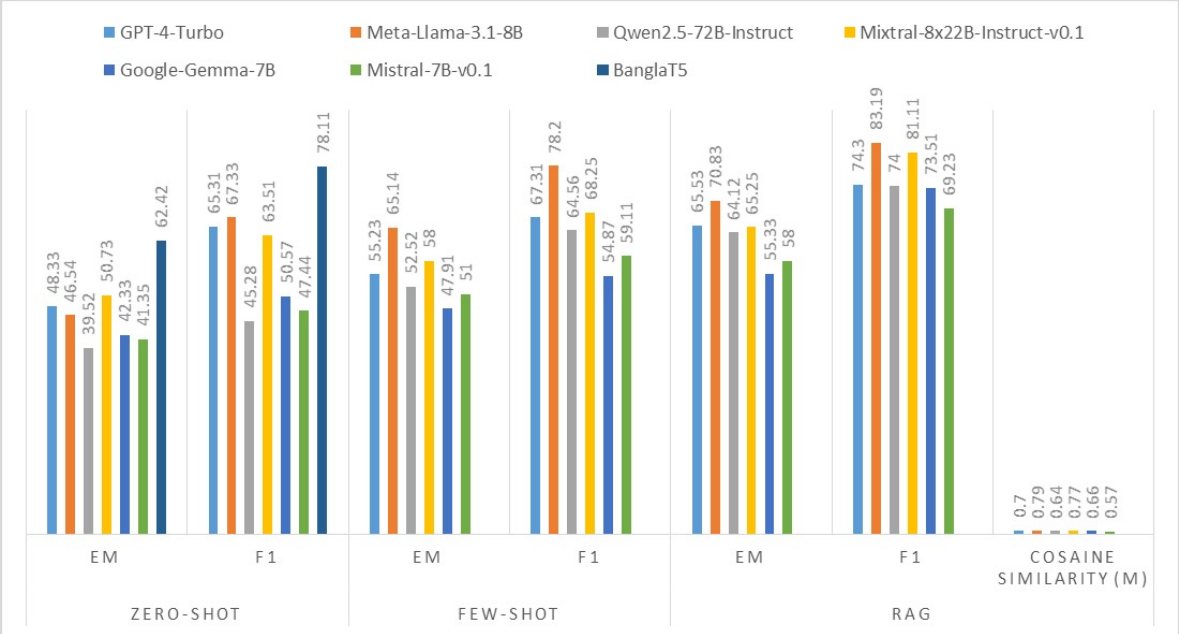

Figure 3. Performance graph of BHRE and RAG system EM and F1 score with average cosine similarity on LLMs.

4.2. *Performance on Different Types of Questions*

Among our approaches, RAG outperformed BHRE on the BanglaRQA test set, with Llama-3.1-8B showing the best results. We therefore analyzed its performance across different question types, as summarized in Table 3.

Table 3. Performance of Llama-3.1-8B (RAG) on different types of questions

	Factiod	**Causal**	**Confirmation**	**list**
EM_{RAG}	77.30	68.53	91.88	52.50
$F1_{RAG}$	85.67	80.31	91.88	76.66

The model performs significantly well on the confirmation types of questions such as BanglaT5, as it had to choose between two possible answers, ” (Yes)” or ”(No)”. The performance on factoid types of questions was better than the entire performance as the answers for this type of question are mostly single-span, whereas the casual and list types of questions perform less better than other two types, especially for list type. But the score is overall good after we use our RAG method.

 doi:10.20944/preprints202601.1821.v1

4.3. Performance on Different Types of Answers

As with the types of evaluation questions, we analyze how it performs on different types of answers. We found a very good similarity for all type questions that contain Table 4.

Table 4. Performance of Llama-3.1-8B (RAG) on different types of answers

	single span	**multiple spans**	**yes/no**
EM_{RAG}	76.66	51.31	92.21
$F1_{RAG}$	86.33	71.22	92.21

The model performs impressively well in conformational or yes/no types of answers. However, for multiple span-type answers, this showed least performance for both 51.31% EM and 71.22% F1 scores. But this enhances the performance more than the BanglaT5 model in the BanglaRQA data set.

4.4. Impact of RAG Components

Our experiments clearly show that the combination of BHRE exemplars with retrieval leads to performance beyond the previously reported state-of-the-art baseline (BanglaT5 with 62.42 EM and 78.11 F1). The BHRE few-shot block improves answer formatting and increases exact match by reducing hallucinated outputs. When the retrieval layer is added, additional gains are observed, especially in causal and list questions. We measured retrieval quality using the cosine similarity between the encoded question and passage chunks. Higher similarity values strongly correlate with better downstream EM and F1, confirming that accurate retrieval is a major key factor in model success. For example, questions with the top 1 retrieved passage similarity above 0.7 produced correct spans, while the lower similarity often led to partial or wrong answers. Therefore, BHRE-RAG demonstrates that carefully chosen zero-shot exemplars combined with similarity-based retrieval deliver stronger results than both zero-shot and few-shot prompting and the BanglaT5 baseline, providing clear evidence that even without fine-tuning, LLMs can surpass domain-specific SOTA systems on BanglaRQA.

5. Conclusions

This study explored the ability of large language models to tackle Bangla reading comprehension using the BanglaRQA dataset. By combining Bangla Holistic Reasoning Evaluation (BHRE) with the Retrieval Augmented Generation (RAG) system, we demonstrated that pre-trained LLMs can deliver strong performance without fine-tuning, in some cases surpassing the fine-tuned state-of-the-art models like BanglaT5. The retrieval layer, supported by cosine similarity, provided critical evidence alignment, while BHRE ensured strict answer formatting, narrowing the gap between generated outputs and gold references. The results highlight the practical strength of simple but effective strategies for under-resourced languages, showing that careful prompt design and retrieval can unlock competitive accuracy even in low-data settings. While further work is needed to improve reasoning for multi-span and causal questions, and should extend this framework with richer retrieval sources and cross-lingual transfer to strengthen Bangla NLP benchmarks.

References

1. P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.
2. Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *arXiv preprint arXiv:1809.09600*, 2018.
3. N. Raihan and M. Zampieri, "Tigerllm-a family of bangla large language models," *arXiv preprint arXiv:2503.10995*, 2025.
4. T. Mahfuz, S. K. Dey, R. Naswan, H. Adil, K. S. Sayeed, and H. S. Shahgir, "Too late to train, too early to use? a study on necessity and viability of low-resource bengali llms," *arXiv preprint arXiv:2407.00416*, 2024.
5. S. M. S. Ekram, A. A. Rahman, M. S. Altaf, M. S. Islam, M. M. Rahman, M. M. Rahman, M. A. Hossain, and A. R. M. Kamal, "Banglarqa: A benchmark dataset for under-resourced bangla language reading

 doi:10.20944/preprints202601.1821.v1

comprehension-based question answering with diverse question-answer types," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2518–2532, Association for Computational Linguistics, 2022.
6. A. Bhattacharjee, T. Hasan, W. U. Ahmad, and R. Shahriyar, "Banglanlg and banglat5: Benchmarks and resources for evaluating low-resource natural language generation in bangla," *arXiv preprint arXiv:2205.11081*, 2022.
7. L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.
8. OpenAI, "Gpt-4 turbo documentation." https://platform.openai.com/docs/models/gpt-4-turbo, 2025.
9. M. LLaMA, "Llama 3.1-8b." https://huggingface.co/meta-llama/Llama-3.1-8B, 2025. Accessed: 2025-09-11.
10. Q. Team, "Qwen2.5-72b-instruct." https://huggingface.co/Qwen/Qwen2.5-72B-Instruct, 2025. Accessed: 2025-09-11.
11. Mistral AI, "Mixtral-8x22b-instruct-v0.1." https://huggingface.co/mistralai/Mixtral-8x22B-Instruct-v0.1, 2025. Accessed: 2025-09-11.
12. Google, "Gemma-7b-it." https://huggingface.co/google/gemma-7b-it, 2025. Accessed: 2025-09-11.
13. Mistral AI, "Mistral-7b-v0.1." https://huggingface.co/mistralai/Mistral-7B-v0.1, 2025. Accessed: 2025-09-11.
14. S. Sultana, S. S. Muna, M. Z. Samarukh, A. Abrar, and T. M. Chowdhury, "Banglamedqa and banglammedbench: Evaluating retrieval-augmented generation strategies for bangla biomedical question answering," *arXiv preprint arXiv:2511.04560*, 2025.
15. C.-C. Chang, C.-F. Li, C.-H. Lee, and H.-S. Lee, "Enhancing low-resource minority language translation with llms and retrieval-augmented generation for cultural nuances," in *Intelligent Systems Conference*, pp. 190–204, Springer, 2025.
16. A. Abrar, N. T. Oeshy, P. Maheru, F. Tabassum, and T. M. Chowdhury, "Faithful summarization of consumer health queries: A cross-lingual framework with llms," *arXiv preprint arXiv:2511.10768*, 2025.
17. A. Abrar, N. T. Oeshy, M. Kabir, and S. Ananiadou, "Religious bias landscape in language and text-to-image models: Analysis, detection, and debiasing strategies," *arXiv preprint arXiv:2501.08441*, 2025.
18. M. R. A. H. Rony, S. K. Shaha, R. A. Hasan, S. K. Dey, A. H. Rafi, A. H. Sirajee, and J. Lehmann, "Banglaquad: A bengali open-domain question answering dataset," *arXiv preprint arXiv:2410.10229*, 2024.
19. S. Veturi, S. Vaichal, R. L. Jagadheesh, N. I. Tripto, and N. Yan, "Rag based question-answering for contextual response prediction system," *arXiv preprint arXiv:2409.03708*, 2024.
20. W. W. Faieaz, S. Jannat, P. K. Mondal, S. S. Khan, S. Karmaker, and M. S. Rahman, "Advancing bangla nlp: Transformer-based question generation using fine-tuned llm," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pp. 1–7, IEEE, 2025.
21. A. S. Ipa, M. A. T. Rony, and M. S. Islam, "Empowering low-resource languages: Trase architecture for enhanced retrieval-augmented generation in bangla," in *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pp. 8–15, 2025.
22. S. Shafayat, H. Hasan, M. Mahim, R. Putri, J. Thorne, and A. Oh, "Benqa: A question answering benchmark for bengali and english," in *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1158–1177, 2024.
23. A. R. Fahad, N. Al Nahian, M. A. Islam, and R. M. Rahman, "Answer agnostic question generation in bangla language," *International Journal of Networked and Distributed Computing*, vol. 12, no. 1, pp. 82–107, 2024.
24. S. C. Roy and M. M. H. Manik, "Question-answering system for bangla: Fine-tuning bert-bangla for a closed domain," *arXiv preprint arXiv:2410.03923*, 2024.
25. M. Rakibul Hasan, R. Majid, and A. Tahmid, "Bangla-bayanno: A 52k-pair bengali visual question answering dataset with llm-assisted translation refinement," *arXiv e-prints*, pp. arXiv–2508, 2025.
26. E. AI, "Generative ai api." https://www.edenai.co/technologies/generative-ai, 2025. Accessed: 2025-09-11.
27. Chroma, "Chroma: An open-source search and retrieval database for ai applications." https://www.trychroma.com/, 2025. Accessed: 2025-09-13.