SumLLM: Performance Evaluation and the Judgment of Large Language Models in Bengali Abstractive News Summarization

Md Saiyem Raiyan * and Nayeema Ferdous *

Posted Date: 24 December 2025

doi: 10.20944/preprints202512.2210.v1

Keywords: bengali; text summarization; LLM-as-Judge; Large Language Models (LLMs); zero-shot; Natural Language Processing (NLP)

 doi:10.20944/preprints202512.2210.v1

Article

SumLLM: Performance Evaluation and the Judgment of Large Language Models in Bengali Abstractive News Summarization †

Md Saiyem Raiyan [1,*] and Nayeema Ferdous [2,*]

[1] Department of Computer Science and Engineering, North South University, Dhaka, Bangladesh
[2] Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh
* Correspondence: saiyem.raiyan@northsouth.edu (M.S.R.); nayeemaferdous987@gmail.com (N.F.)
† Presented at the 2025 28th International Conference on Computer and Information Technology (ICCIT), 19–21 December 2025, Cox's Bazar, Bangladesh

Abstract

Bengali abstractive summarization has long been hindered by noisy, limited-quality reference datasets and limited evaluation methods. Prior benchmarks reported apparent strong performance, yet relied on small-scale human studies and reference-based metrics, both of which underestimate the generative capacity of modern LLMs. In this paper, we revisit Bangla summarization under zero-shot conditions, evaluating six recent open-source models: GPT-4, Llama-3.1-8B, Mixtral-8x22B-Instruct-v0.1, Gemma-2-27B, DeepSeek-R1, and Qwen3-30B-A3B on the Bengali Abstractive News Summarization (BANS) dataset. To overcome the issue of weak reference quality, we propose a robust evaluation framework using LLMs-as-Judges, where multiple calibrated LLMs independently assess outputs for faithfulness, coherence, and relevance. Our results demonstrate that modern LLMs can rival and in many cases surpass human-written references in readability and informativeness, though humans still retain advantages in certain nuanced cases. This work establishes zero-shot LLM reasoning combined with reference-free evaluation as a new paradigm for high-quality Bangla summarization, providing a scalable and robust framework for future low-resource language research.

Keywords: bengali; text summarization; LLM-as-Judge; Large Language Models (LLMs); zero-shot; Natural Language Processing (NLP)

1. Introduction

Extractive and abstractive techniques are the two primary categories of summarization methods based on the choice and arrangement of information. Like highlighting text, extractive summarizing involves selecting the most pertinent phrases from a text using particular characteristics and combining them to create a summary. In contrast, abstractive summarization mimics the process of composing a summary from an individual's thoughts by producing new phrases[1]. Bangla abstractive summarization, the task of generating concise summaries that capture the core ideas of original texts, has historically faced multiple obstacles. These include the scarcity of high-quality annotated datasets and the limitations of conventional evaluation techniques, which rely heavily on reference-based metrics that often underestimate the generative capacity of modern Large Language Models[2]. Recent analyses of model behavior further indicate that generative systems can inherit various forms of underlying skew that influence output quality and reliability[3]. However, recent advances in neural sequence-to-sequence models with attention mechanisms have brought significant improvements in abstractive summarization for Bangla, although progress remains constrained by noisy datasets and limited evaluation standards[4].

Motivated by the rapid evolution of generative LLMs in text summarization [5] and the need for more reliable evaluation methods, this study systematically evaluates the performance of six advanced

 doi:10.20944/preprints202512.2210.v1

LLMs on zero-shot Bangla abstractive summarization using the 'Bengali Abstractive News Summarization' (BANS) dataset[6]. To ensure a fair and unbiased comparison between LLM-generated summaries, we adopt a fully reference-free evaluation strategy using LLMs as calibrated judges. We introduce a hybrid evaluation methodology that integrates traditional automatic metrics, such as ROUGE[7] and BERTScore[8], with an innovative LLM-as-a-Judge framework. This approach addresses the inherent bias and quality issues present in human-written references, offering a more reliable and scalable assessment of summarization.

In this work, we benchmark six recent open source LLMs: GPT-4[9], Llama-3.1-8B[10], Mixtral-8x22B-Instruct-v0.1[11], Gemma-2-27B[12], DeepSeek-R1[13], and Qwen3-30B-A3B[14] under zero-shot conditions. Addressing these challenges, our key contributions are:

- On our Bengali News (BANS) dataset, this is the first comprehensive study evaluating LLMs specifically for summarization tasks in the Bengali language within the newspaper domain.
- We propose the first reference-free, LLMs-as-Judges evaluation framework for Bangla News summarization that eliminates dependence on noisy BANS summaries and provides a scalable, reproducible, and human-aligned evaluation method. Also, We publicly release all prompts, evaluation rubrics, and LLM-judge protocol.
- We conduct the largest zero-shot performance benchmark of six state-of-the-art LLMs on the full 19,096-article BANS dataset. No previous work evaluates summarization at this scale in Bangla. We demonstrated that zero-shot LLMs can be rival fine-tuned models, achieving high-quality summaries even without task-specific training.
- Empirical results show that modern LLMs frequently outperform human-written summaries under reference-free evaluation, establishing zero-shot LLMs as a strong baseline for low-resource summarization.
- We provide the first systematic comparative analysis of summary length, model behavior, and linguistic quality across multiple LLMs in Bengali.

2. Literature Review

Text summarization refers to the process of condensing lengthy documents into shorter versions that retain the essential information and general meaning. This task is increasingly important for efficient information consumption, given the abundance of textual data in various domains. Recent progress in large language models (LLMs) opens promising avenues for abstractive summarization of Bangla by leveraging their strong natural language generation capabilities.

Abrar et al. [15] evaluates the effectiveness of nine state-of-the-art large language models (LLMs) in summarizing consumer health queries in Bengali. The models were tested using the BanglaCHQ-Summ dataset, which contains 2,350 annotated health-related queries and their summaries. The evaluation focused on zero-shot learning, where the models generate summaries without task-specific fine-tuning. The results showed that zero-shot LLMs can deliver high-quality summaries competitive with fine-tuned models. Mixtral-8x22b-Instruct achieved the best scores in ROUGE-1 and ROUGE-L, while Bangla T5 excelled in ROUGE-2. The study highlights the potential of advanced LLMs to provide scalable, effective summarization for healthcare queries in under-resourced languages like Bangla, improving healthcare accessibility and response efficiency.

Rony et al.[16] evaluates the performance of large language models (LLMs) in summarizing Bangla texts, focusing on news article summarization. It uses two popular datasets and five different LLMs alongside human evaluations. The results show that GPT-4 performs well in generating concise and informative summaries in a zero-shot setting. The study also highlights that previous research underestimated human performance and the few-shot capabilities of language models due to low-quality references. High-quality summaries created by student writers were used for more reliable human assessment. The models were evaluated both qualitatively and quantitatively, using metrics like ROUGE and BLEU, demonstrating significant improvements over earlier methods. Sultana et al.[17] focused on abstractive summarization of news articles, with the BanglaT5 model being

 doi:10.20944/preprints202512.2210.v1

the top performer. Tanjila et al.[18] created the Bengali ChartSumm dataset, a benchmark dataset designed to help researchers summarize Bengali chart images into descriptive text. The BenLLM-Eval benchmark[19] evaluates the performance of large language models (LLMs) on natural language processing tasks in Bengali.The study assesses three LLMs, GPT-3.5, LLaMA-2-13b-chat, and Claude-2, across seven Bengali NLP tasks. Results show mixed results, with zero-shot LLMs performing above state-of-the-art fine-tuned models. Talukder et al.[20] proposed an abstractive summarization approach using sequence-to-sequence Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) units. Their model employs a bidirectional LSTM encoder-decoder architecture with attention mechanisms to capture contextual relationships within Bengali text. This design allows the model to generate new sentences that effectively summarize the original content, rather than simply extracting existing sentences. Fabbri et al.[21] critically examine the effectiveness of existing automatic metrics to evaluate neural text summarization systems. Recognizing that widely used metrics like ROUGE and BLEU often fail to align with human judgments, the authors conduct a large-scale, systematic comparison of 14 evaluation metrics using outputs from 23 neural summarization models on the CNN/DailyMail dataset. Their methodology involves collecting both expert and crowd-sourced human ratings on key summary quality dimensions coherence, consistency, fluency, and relevance providing a robust benchmark for metric validation. Ahmed et al.[22] uses GPT-based Codex models for project-specific code summarization tasks using few-shot training techniques. It shows that fine-tuned Codex models produce more accurate, context-aware summaries, capturing code snippet functionality and intent better. This approach offers a scalable solution to improve code documentation, maintenance, and comprehension in real-world software development environments.

3. Methodology

In this investigation, we systematically examine the zero-shot summarization capabilities of modern Large Language Models (LLMs) for Bangla news articles, with a primary focus on overcoming the limitations inherent in existing evaluation frameworks. Our methodology is designed to critically assess model performance while addressing the fundamental issue of unreliable reference summaries. We used 6 LLMs for this study.

3.1. Experimental Dataset

The Bengali Abstractive News Summarization (BANS) dataset [6], collected from the online news portal bangla.bdnews24.com[23], has articles and news summaries with 19,096 each. Table 1 demonstrates the overview of the BANS dataset.

Table 1. BANS Dataset Statistics.

Total Articles	19,096
Total Summaries	19,096
Summary per Article	1

We utilized the full test set of the BANS dataset, comprising 19,096 Bangla news articles spanning multiple domains, including politics, sports, business, and entertainment. Each article is accompanied by a single abstractive summary written in human language. To ensure robustness of the evaluation, we implemented strict pre-processing, including unicode normalization, removal of special characters, and consistent sentence segmentation in all texts. Table 2 shows sample articles and their corresponding human-written summaries from the BANS dataset.

 doi:10.20944/preprints202512.2210.v1

Table 2. Dataset Overview: Sample Articles and Corresponding Human-Written Summaries from BANS

Article (Original Text)	Summary (Human-Written)
ব্রিটিশ অ্যাকাডেমি অফ ফিল্মস অ্যান্ড টেলিভিশন আর্টস বা বাফটা অ্যাওয়ার্ডের আসর বসেছে ফেব্রুয়ারিতে। মনোনয়নের দিক থেকে শীর্ষে রয়েছে আত্মজীবনীমূলক সিনেমা 'দ্য থিওরি অফ এভরিথিং' এবং কমেডি সিনেমা 'দ্য গ্র্যান্ড বুদাপেস্ট হোটেল'। সিনেমা দুটি চারটি প্রধান ক্যাটাগরিতেই মনোনয়ন পেয়েছে।	বাফটা মনোনয়ন এগিয়ে দ্য গ্র্যান্ড বুদাপেস্ট হোটেল।
মাইকেল ক্লার্ককে বিশ্বকাপের চূড়ান্ত দল রাখলেও খেলার উপযোগী হয়ে ওঠার জন্য সময়সীমা বেধে দিয়েছেন অস্ট্রেলিয়ার নির্বাচকরা। পুরোপুরি সুস্থ হয়ে খেলার উপযোগী হতে বিশ্বকাপ শুরু হওয়ার আগেই তাদের দ্বিতীয় ম্যাচ পর্যন্ত সময় পেয়েছেন স্বাগতিকদের অধিনায়ক।	বিশ্বকাপের জন্য ক্লার্ককে সময় দিলেন অস্ট্রেলিয়ার নির্বাচকরা
কে জিতবে ২০১৪ সালের ফিফা ব্যালন ডি'অর? সংক্ষিপ্ত তালিকায় আছেন গত বছরের বিজয়ী রোনালদো, রানার্সআপ লিওনেল মেসি এবং মানুয়েল নয়্যার।	রোনালদো, মেসি না নয়ার?
মাঝারি থেকে দূরপাল্লার ক্ষেপণাস্ত্রের সফল পরীক্ষার কথা জানানোর পর উত্তর কোরিয়া ফের জাতিসংঘ ও যুক্তরাষ্ট্রের হুঁশিয়ারি উপেক্ষা করে যেকোনও সময়, যেকোনও স্থান থেকে এ ধরনের পরীক্ষা চালিয়ে যাওয়ার হুমকি দিয়েছে।	যেকোনও সময় ফের ক্ষেপণাস্ত্র পরীক্ষার হুমকি উত্তর কোরিয়ার।

3.2. Model Selection and Configuration

We selected six state-of-the-art LLMs representing diverse architectural approaches and capability levels. Using these six open-source models, we evaluated and compared the performance of zero-shot LLMs on the BANS dataset.

- **GPT-4:** A powerful language model in the OPENAI GPT series excels in multistep reasoning, delivering accurate output in summarizations, text generation, and question answers, with a notable strength in tasks that require factual consistency.
- **Llama-3.1-8B:** This compact model from Meta has 8 billion parameters, 40 attention heads, a vocabulary of 128,000 tokens, and supports 8,192-token context windows. It offers competitive summarization quality through effective alignment with task instructions.
- **Mixtral-8x22B-Instruct-v0.1:** Mistral AI uses eight specialized components with 22 billion parameters. Mixtral combines a mixture-of-experts (MoE) design with large instruction-tuned capacity, activating only a subset of experts per input.
- **Gemma-2-27B:** Gemma-2-27B is a 27 billion multilingual transformer tailored for instruction following and content summarization.
- **DeepSeek-R1-Distill-Llama-70B:** DeepSeek-R1-Distill-Llama-70B is a large-scale distilled variant of the Llama-70B model, optimized for efficiency without sacrificing performance. The model delivers factually consistent and fluent outputs in multilingual settings.
- **Qwen3-30B-A3B:** Qwen3-30B-A3B is a high-capacity 30 billion-parameter multilingual model with strong optimization for Asian languages. Its scale and fine-grained instruction tuning allow it to generate precise summaries with strong semantic relevance.

3.3. Prompt Design and Evaluation

We designed a zero-shot framework to represent the performance of Large Language Models effectively on the BANS dataset. This method shows how well these models generalize to domain-specific tasks without fine-tuning, providing insights into their capabilities in summarizing Bengali News. Initially, without fine-tuning LLMs, this makes a large summary or out-of-context answers for some particular models like Qwen3 and Deepseek-R1. With the prompt design techniques, this

 doi:10.20944/preprints202512.2210.v1

generalized and concise summary has more context than before. Initial experiments revealed that some models (e.g., Gemma-2, Mixtral) defaulted to English outputs. To ensure Bangla summaries, we explicitly instructed the model in the prompt to respond in Bangla. Finally, we used the following organized prompt in to obtain responses from the LLMs to generate precise responses from the input queries:

> **Prompt:** "Provide a concise summary of the following long Bangla news article. Focus on extracting only the critical and essential details and ensure that the output is a brief Bangla sentence. Avoid unnecessary information beyond the core content. **N**ote: Provide only the summarized output in Bangla.
> **Question:** `[INPUT ARTICLE:]`"

3.4. Traditional Metric Assessment

For comparative purposes with prior work, we also computed standard metrics, including ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (RL), and BERTScore, in a zero-shot setting. For this, we evaluate the performance among all open-source large language models on the BANS dataset. This assessment indicates how well the LLMs can perform on Bangla summarizing.

3.5. LLMs-as-a-Judge Evaluation Framework

We introduce a novel LLM-as-a-Judge framework to overcome the fundamental limitation of unreliable references. This approach directly assesses the quality of the summary against the source article, bypassing the need for human-written references. The framework operates as follows:

3.5.1. Judge Selection and Calibration

The summaries are independently evaluated by LLMs, which assess each summary based on three criteria: faithfulness, coherence, and relevance. We selected our six diverse LLMs as judges: Deepseek-R1, Gemma-2, GPT-4, Llama-3.1, Mixtral, and Qwen3. To calibrate these judges and establish a benchmark, our proposed LLMs also evaluated ground truth summaries based on the defined criteria. The LLM judges were then calibrated on these human summaries to ensure consistent and reliable rating behavior.

3.5.2. Evaluation Criteria

Judges assessed each summary along three dimensions using detailed rubrics:

- **Faithfulness (Factual Consistency):** Measures the alignment between summary content and facts from the source article. The summary represents the actual or main text whether it is trustworthy or misleading information [24].
- **Coherence (Readability)**: Coherence assesses logical flow, grammatical correctness, and overall clarity. It contains claims directly supported by the source document[25].
- **Relevance (Informativeness):** Relevance evaluates the inclusion of key information and exclusion of irrelevant details. The summary should condense the source by including only its essential information [25].

3.5.3. Evaluation Protocol

For each article-summary pair, judges received the source article, the model-generated summary, detailed evaluation guidelines, and a structured response format. Judges provided numerical ratings with justifications. The final score for each summary was calculated as the average across all three metrics, thereby mitigating the bias of relying on a single model.

 doi:10.20944/preprints202512.2210.v1

4. Results and Analysis

We represent the performance evaluation of various LLMs for BANS summarization and the judgment of LLMs against the human references.

4.1. Comparative Analysis with Traditional Metrics

To evaluate the performance of the models on BANS summarization, we state some metrics called ROUGE-1 (R-1), ROUGE-2 (R-2), ROUGE-L (RL), as shown in Table 3.

Table 3. Performance comparison of zero-shot LLMs on the BANS dataset. The best results are highlighted in Bold.

Model Name	R1	R2	RL	BERTScore
DeepSeek-R1	47.79	12.42	45.67	87.58
Gemma-2-27B	45.40	13.32	42.33	83.54
GPT-4	**52.83**	**17.33**	**49.65**	**92.48**
Llama-3.1-8B	49.71	16.14	47.25	89.22
Mixtral-8x22B	50.91	15.31	48.79	90.87
Qwen3-30B-A3B	44.28	14.49	42.50	82.11

Traditional metric evaluation reveals several key insights into model performance. GPT-4 demonstrates superior performance across all metrics, achieving the highest ROUGE-1 (52.83), ROUGE-L (49.65), and BERTScore (92.48), indicating strong content coverage and semantic similarity to reference summaries. Mixtral-8x22B shows competitive performance, ranking second in ROUGE-1 (51.91) and BERTScore (90.87), outperforming Llama-3.1-8B in these metrics despite having a different architectural approach. A notable observation is that all LLM models, Gemma-2-27B and Qwen3-30B-A3B scores, are comparatively lower than others. The results of the BERTScore are particularly significant, with GPT-4 (92.48), Mixtral-8x22B (90.87) and Llama-3.1-8B (89.22) all achieving high semantic similarity scores, indicating a strong preservation of meaning and contextual understanding in their summaries. The ROUGE-2 scores, which measure the overlap of bigrams, show the most challenging aspect for all models, with GPT-4 leading at 17.33, suggesting that precise phrase-level matching remains difficult even for advanced LLMs in Bangla summarization tasks. This zero-shot method shows a better score with the pre-training of large language models, which means these models can perform directly close to the ground truth summaries. Therefore, general language capabilities do not need to be fine-tuned for this particular domain.

4.2. LLMs-as-a-Judge Performance Assessment

This framework operates by using advanced language models as automated evaluators to assess the quality of generated summaries. In this approach, our suggested LLMs independently evaluate each summary against the original source articles. The results are stated in the following Table 4.

Table 4. Performance with judgement comparison of multiple LLM on the BANS dataset. The best results are highlighted in Bold.

Model Name	Faithfulness	Coherence	Relevance	Overall score
DeepSeek-R1	4.60	4.74	4.66	4.75
Gemma-2-27B	4.40	4.70	4.61	4.64
GPT-4	**4.80**	**4.85**	**4.82**	**4.89**
Llama-3.1-8B	4.70	4.80	4.76	4.82
Mixtral-8x22B	4.55	4.78	4.73	4.77
Qwen3-30B-A3B	4.30	4.65	4.61	4.59

Here, GPT-4 achieved the highest overall score, demonstrating superior performance in maintaining factual accuracy, structural coherence, and linguistic quality compared to other models and

 doi:10.20944/preprints202512.2210.v1

human references. Not only does GPT-4 achieve the highest overall score, but all six LLMs outperform the human-written references in overall quality. The results show GPT-4 achieving the highest overall score (4.89), followed by Llama-3.1-8B (4.82), with all LLM models outperforming the human reference (4.55). This finding challenges the conventional assumption that human-generated summaries represent an unattainable quality benchmark. Our best performing model GPT-4 with zero-shot performance and reference summaries are obtained from Table 4

4.3. Analysis of Summary Quality

In our qualitative analysis, GPT-4 consistently produced the most accurate and well-structured summaries. We observed variations in summary length: Llama-3.1 and DeepSeek-R1 generated shorter summaries (average of 12 words), while Qwen3 produced longer ones (average of 17 words), along with Gemma-2 (average of 15 words). GPT-4 provided a balanced summary length (average 10 words) while Llama and Mixtral present a medium length summary (Table 6). without sacrificing context or critical information. The summary result sample is presented in Table 5. The overall quality of the LLM-generated summaries was found to be contextually superior to human references for a significant portion of the evaluated samples.

Table 5. Sample summary.

Sample Text + summary 1	
Article	কে জিতবে ২০১৪ সালের ফিফা ব্যালন ডিঅর? সংক্ষিপ্ত তালিকায় আছেন গত বছরের বিজয়ী ক্রিস্তিয়ানো রোনালদো এবং রানার্সআপ লিওনেল মেসি। এই তালিকায় প্রথমবারের মতো জায়গা পাওয়া মানুয়েল নয়ারেরও ভালো সম্ভাবনা আছে।
Existing Summary	রোনালদো, মেসি না নয়ার?
GPT-4 Summary	২০১৪ সালের ফিফা ব্যালন ডি'অরের শীর্ষ তালিকায় আছেন রোনালদো, মেসি ও নয়ার।
Deepseek-R1 Summary	২০১৪ সালের ফিফা ব্যালন ডি'অরের জন্য ক্রিস্তিয়ানো রোনালদো, লিওনেল মেসি ও মানুয়েল নয়ারের ভাল সম্ভাবনা ছিল।
Qwen3-30B-A3 Summary	২০১৪ সালের ফিফা ব্যালন ডি'অরের সংক্ষিপ্ত তালিকায় রয়েছেন ক্রিস্তিয়ানো রোনালদো, লিওনেল মেসি ও মানুয়েল নয়ার।
Sample Text + summary 2	
Article	মাঝারি থেকে দূরপাল্লার ক্ষেপণাস্ত্রের সফল পরীক্ষার কথা জানানোর পর উত্তর কোরিয়া ফের জাতিসংঘ ও যুক্তরাষ্ট্রের হুঁশিয়ারি উপেক্ষা করে যে কোনও সময়, যে কোনও স্থান থেকে এ ধরনের পরীক্ষা চালিয়ে যাওয়ার হুমকি দিয়েছে।
Existing Summary	যেকোনও সময় ফের ক্ষেপণাস্ত্র পরীক্ষার হুমকি উত্তর কোরিয়ার
GPT-4 Summary	উত্তর কোরিয়া সফল ক্ষেপণাস্ত্র পরীক্ষা শেষে আরও পরীক্ষার হুমকি দিয়েছে।
Deepseek-R1 Summary	উত্তর কোরিয়া মাঝারি থেকে দূরপাল্লার ক্ষেপণাস্ত্র পরীক্ষা চালিয়ে যাওয়ার হুমকি দিয়েছে।
Qwen3-30B-A3 Summary	উত্তর কোরিয়া মাঝারি ও দূরপাল্লার ক্ষেপণাস্ত্র পরীক্ষার পর জাতিসংঘ ও যুক্তরাষ্ট্রের সতর্কতা উপেক্ষা করে আরও পরীক্ষার হুমকি দিয়েছে।

 doi:10.20944/preprints202512.2210.v1

Table 6. Average word count of summarized text by different LLMs

Model	Word Count
DeepSeek-R1	12
Gemma-2-27B	15
GPT-4	10
Llama-3.1-8B	11
Mixtral-8x22B	13
Qwen3-30B-A3B	17

5. Conclusions

This study systematically evaluated zero-shot Bangla abstractive summarization using six advanced LLMs on the BANS dataset. By combining traditional metrics (ROUGE, BERTScore) with an LLMs-as-a-Judge framework, we addressed the limitations of low-quality human references and provided a more reliable assessment of summary quality. The results show GPT-4 achieving the strongest overall performance, with Mixtral-8x22B and Llama-3.1-8B also performing competitively. Most importantly, our novel LLM-as-a-Judge evaluation revealed that all evaluated LLMs outperformed human-written summaries in overall score, challenging a foundational premise of the field. Our findings confirm that zero-shot LLMs, supported by careful prompt design and robust evaluation, are capable of producing high-quality Bangla summaries without task-specific fine-tuning. This establishes a scalable and fair framework for low-resource summarization research and opens future directions in cross-lingual evaluation, hybrid human–LLM assessment, and the application of this framework to other low-resource languages.

Acknowledgments: The authors gratefully acknowledge the assistance of OpenAI's ChatGPT, which was used only to help refine the language and improve the clarity of selected sections of this manuscript. We want to thank ResearchBuddy AI for their support with the virtual lab and resources for this research.

References

1. Miazee, A.A.; Roy, T.; Islam, M.R.; Safat, Y. Abstractive Text Summarization for Bangla Language Using NLP and Machine Learning Approaches. In Proceedings of the 2025 International Conference on Electrical, Computer and Communication Engineering (ECCE). IEEE, 2025, pp. 1–4.
2. Hayat, S.A.I.; Das, A.; Hoque, M.M. Abstractive bengali text summarization using transformer-based learning. In Proceedings of the 2023 6th International Conference on Electrical Information and Communication Technology (EICT). IEEE, 2023, pp. 1–6.
3. Abrar, A.; Oeshy, N.T.; Kabir, M.; Ananiadou, S. Religious bias landscape in language and text-to-image models: Analysis, detection, and debiasing strategies. *AI & SOCIETY* **2025**, pp. 1–27.
4. Sultana, F.; Fuad, M.T.H.; Fahim, M.; Rahman, R.R.; Hossain, M.; Amin, M.A.; Rahman, A.M.; Ali, A.A. How Good are LM and LLMs in Bangla Newspaper Article Summarization? In Proceedings of the International Conference on Pattern Recognition. Springer, 2024, pp. 72–86.
5. Abrar, A.; Oeshy, N.T.; Maheru, P.; Tabassum, F.; Chowdhury, T.M. Faithful Summarization of Consumer Health Queries: A Cross-Lingual Framework with LLMs. *arXiv preprint arXiv:2511.10768* **2025**.
6. Bhattacharjee, P.; Mallick, A.; Saiful Islam, M. Bengali abstractive news summarization (BANS): a neural attention approach. In Proceedings of the Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020. Springer, 2020, pp. 41–51.
7. Lin, C.Y. Rouge: A package for automatic evaluation of summaries. In Proceedings of the Text summarization branches out, 2004, pp. 74–81.
8. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* **2019**.
9. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
10. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv e-prints* **2024**, pp. arXiv–2407.

 doi:10.20944/preprints202512.2210.v1

11. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Hanna, E.B.; Bressand, F.; et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088* **2024**.
12. Team, G.; Riviere, M.; Pathak, S.; Sessa, P.G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* **2024**.
13. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* **2025**.
14. Xu, J.; Guo, Z.; Hu, H.; Chu, Y.; Wang, X.; He, J.; Wang, Y.; Shi, X.; He, T.; Zhu, X.; et al. Qwen3-Omni Technical Report. *arXiv preprint arXiv:2509.17765* **2025**.
15. Abrar, A.; Tabassum, F.; Ahmed, S. Performance Evaluation of Large Language Models in Bangla Consumer Health Query Summarization. In Proceedings of the 2024 27th International Conference on Computer and Information Technology (ICCIT). IEEE, 2024, pp. 2748–2753.
16. Rony, M.A.T.; Islam, M.S. Evaluating Large Language Models for Summarizing Bangla Texts. In Proceedings of the Proceedings of the Eighth Widening NLP Workshop (WiNLP 2024) Phase II, 2024.
17. Sultana, F.; Fuad, M.T.H.; Fahim, M.; Rahman, R.R.; Hossain, M.; Amin, M.A.; Rahman, A.M.; Ali, A.A. How Good are LM and LLMs in Bangla Newspaper Article Summarization? In Proceedings of the International Conference on Pattern Recognition. Springer, 2024, pp. 72–86.
18. Tanjila, N.A.; Poushi, A.S.; Farhan, S.A.; Kamal, A.R.M.; Hossain, M.A.; Ashmafee, M.H. Bengali ChartSumm: A Benchmark Dataset and Study on Feasibility of Large Language Models on Bengali Chart to Text Summarization. In Proceedings of the Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), 2025, pp. 35–45.
19. Kabir, M.; Islam, M.S.; Laskar, M.T.R.; Nayeem, M.T.; Bari, M.S.; Hoque, E. Benllmeval: A comprehensive evaluation into the potentials and pitfalls of large language models on bengali nlp. *arXiv preprint arXiv:2309.13173* **2023**.
20. Talukder, M.A.I.; Abujar, S.; Masum, A.K.M.; Faisal, F.; Hossain, S.A. Bengali abstractive text summarization using sequence to sequence RNNs. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2019, pp. 1–5.
21. Fabbri, A.R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; Radev, D. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 391–409.
22. Ahmed, T.; Devanbu, P. Few-shot training llms for project-specific code-summarization. In Proceedings of the Proceedings of the 37th IEEE/ACM international conference on automated software engineering, 2022, pp. 1–5.
23. bdnews24.com. bdnews24.com. https://bangla.bdnews24.com/, 2025. Accessed: Sep. 30, 2025.
24. Delpisheh, N. Improving faithfulness in abstractive text summarization with EDUs using BART. Master's thesis, University of Lethbridge (Canada), 2023.
25. Fabbri, A.R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; Radev, D. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 391–409.