Evaluating Large Language Models for Summarizing Bangla Texts

Mohammad Abu Tareq Rony
Department of Statistics,
Noakhali Science & Technology
University, Bangladesh
abutareqrony@gmail.com

Mohammad Shariful Islam
Computer Science & Telecommunication
Engineering, Noakhali Science &
Technology University, Bangladesh
shariful.ces43@gmail.com

Abstract

Large Language Models (LLMs) for Bangla text summarization condense texts while preserving key information by leveraging advanced Natural Language Processing (NLP) techniques. In this study, we used two popular Bangla news summarization datasets through the evaluation of 5 LLMs as well as human evaluation. we made two key observations. First, we found that GPT-4 with zero-shot model settings performs well in Bangla news summarization. Secondly, previous research has been constrained by low-quality references, resulting in an underestimation of human performance and diminished few-shot capabilities. To more accurately evaluate LLMs, we performed human assessments using high-quality summaries created by student writers. Despite notable stylistic differences, including the extent of paraphrasing, LLM-generated summaries were found to be comparable to those written by humans. Our model was assessed both qualitatively and quantitatively, and comparisons with other published results showed significant improvements in human evaluation scores due to the LLM techniques.

1 Introduction

Text summarization seeks to condense lengthy documents into concise, coherent, and easily readable formats while retaining the essential information from the original content(Fabbri et al., 2021a). This process, known as automatic text summarization, seeks to extract the most relevant information from a large text document. An effective summary should be coherent(Islam et al., 2024a), non-redundant(Islam et al., 2024b), grammatically correct, and retain the most important contents of the original document(Chowdhury et al., 2021).

To benchmark, we evaluated the BANS datasets (Bhattacharjee et al., 2021) and BNLPC (Haque et al., 2015) datasets, but found existing summaries had many issues. To address these quality concerns and better compare LLMs to human summary writers, we recruited Master's students from our lab to re-annotate 50 articles from the BANS and BNLPC test datasets. Comparing the top-performing LLM, GPT-4, with lab students' summaries, we observed that GPT-4 summaries are also effective. Again, manually annotating used in these summaries (Bhattacharjee et al., 2021) and (Haque et al., 2015), we found that GPT-4 paraphrases less frequently but can coherently combine copied segments.

We recruited annotators to compare the GPT-4 summaries with those of the lab's student writers. Overall, GPT-4 was rated as equal to the lab students, as shown in Figure 1. Examination of individual rater annotations revealed that each rater had a consistent preference for either GPT-4 or the lab students.

1.1 Main Contributions

We achieved a significant improvement in both LLM and human assessments compared to other existing Bengali news summarization techniques. Our main contributions are the following:

1. We conducted a systematic evaluation of five different LLMs on the Bangla news summarization datasets, comparing their outputs to human writers. The results show that LLM outputs are comparable to those produced by human writers.

2. Our evaluation reveals that GPT-4 is crucial for achieving zero-shot summarization capability in Bangla news summarization.

3. Finally, we evaluate our research both qualitatively and quantitatively, and the presented approach outperforms Bengali state-of-the-art approaches.

The organization of this paper is as follows: Section 2 reviews the relevant literature. Section 3 introduces the LLM methodology for Bangla text summarization. Section 4 evaluates the results of different LLM techniques. Finally, Section 5 summarizes our research findings.

2 Related Work

Prior studies on the research framework for Bangla text summarization can generally be divided into three main categories: evaluation metrics, datasets, and models. This study(Chowdhury et al., 2021) presents BenSumm, an unsupervised abstractive summarization system for Bengali texts. It uses a Part-Of-Speech tagger and a pre-trained language model to generate summaries without parallel data. They also created a new human-annotated dataset to evaluate the model, which outperforms existing unsupervised extractive methods. The model clusters sentences and constructs word graphs to achieve sentence fusion, providing an effective solution for Bengali text summarization.

This research(Bhattacharjee et al., 2021) presents Bengali Abstractive News Summarization (BANS), a neural attention-based model using a sequence-to-sequence Long Short-Term Memory (LSTM) network for summarizing Bengali news articles. It leverages a pre-trained Bengali language model to generate human-like summaries with core information. The researchers prepared a dataset of over 19,000 articles from different Bangla news sources and made the dataset publically available on Kaggle. The model, which employs attention mechanisms in both encoder and decoder, significantly outperforms existing methods in terms of human evaluation scores, ROUGE, and BLEU metrics.

In this work, (Hasan et al., 2023) implements a Bangla extractive update summarization task using the BNLPC dataset, which includes over 1,000 Bangla news articles. It evaluates a TF-IDF-based model and a pre-trained SentenceRank model, with the TF-IDF model performing better. The research aims to provide concise, non-redundant summaries to facilitate access to updated Bangla information and set a foundation for future advancements in Bangla summarization.

3 Methodology

In this study, we investigate the performance of LLMs for Bangla news summarization and identify key factors contributing to their success. We conduct a human evaluation of 5 LLMs with different pretraining methods, prompts, and model scales and find that instruction tuning, rather than model size, significantly enhances zero-shot summarization capabilities. Again, we also highlight issues with existing low-quality reference summaries, proposing the collection of higher-quality summaries from persons to improve evaluation accuracy and better understand LLM performance relative to our lab students' summaries.

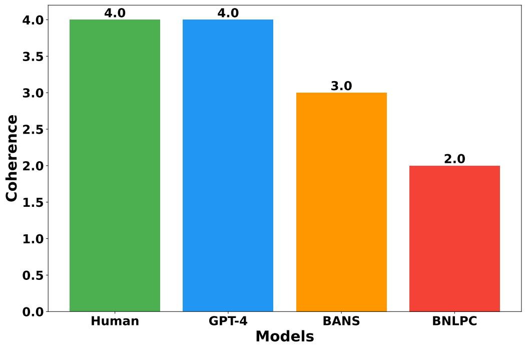

Figure 1: Annotators rated coherence Likert scale.

3.1 Dataset

We conducted our human evaluation on the BANS dataset (Bhattacharjee et al., 2021) and the BNLPC dataset (Haque et al., 2015), sampling 50 examples from each validation set. For limited context, we selected 5 articles between 50 and 100 tokens in length with GPT-3.5 tokenizer. For BANS, uniform sampling sometimes resulted in unreadable articles due to data preprocessing, so we manually selected articles from the training set. Table 1 compares the BNLPC dataset with the BANS dataset.

3.2 Model Details

We evaluated five LLMs, each with different pretraining strategies and model scales. We conduct zero-shot, five-shot, twenty-shot, and fifty-shot for all model settings.

GPT-3.5: GPT-3.5, by OpenAI, is an efficient LLM based on transformer model (Vaswani et al., 2017) with 175 billion parameters. We use the gpt3.5-turbo 0613 version of this model via OpenAI

GPT-4: GPT-4(Achiam et al., 2023), is another powerful language model in OpenAI's GPT series, is known for its enhanced reliability, creativity, and ability to process more nuanced instructions compared to GPT-3.5. However, it is about 25 times more expensive and considerably slower than GPT-3.5. The gpt4-0613 version of this model is used via OpenAI.

OPT: The OPT (Open Pre-trained Transformer) model by Meta is a transformer-based language model designed for efficient natural language processing. It offers robust language understanding and generation capabilities, balancing performance and efficiency(Zhang et al., 2022).

LLaMA2: LLaMA2, developed by Meta, is an advanced language model for natural language processing. It offers improved performance and efficiency, making it ideal for applications like chatbots (Islam et al., 2023) and text generation(Touvron et al., 2023).

PaLM-2: PaLM-2 is a transformer-based language model developed by Google, known for its advanced reasoning abilities(Sajol and Hasan, 2024) and improved computational efficiency(Anil et al., 2023). The text-bison@001 version of this model is utilized via Google's Vertex API.

3.3 Evaluation Toolkit and metrics

he evaluation toolkit contains 3 human evaluation metrics and 5 conventional evaluation methods described as follows;

- **Faithfulness:** Faithfulness in text summarization refers to the accuracy and precision with which the summary represents the original text, ensuring it is truthful and not misleading. (Delpisheh and Chali, 2024).
- **Consistency:** Consistency pertains to the factual accuracy between the summary and the source text(Abu Tareq Rony et al., 2024). A factually consistent summary contains only statements directly supported by the information in the source document (Fabbri et al., 2021a).
- **Relevance:** The selection of important content from the source document is crucial for effective summarization. Summaries should include only the essential information, and annotators were instructed to penalize those that contained redundancies or excess information(Fabbri et al., 2021a).
- **ROUGE:** ROUGE (Recall-Oriented Understudy for Gisting Evaluation) assesses the number of overlapping textual units between the generated summary and a set of reference summaries(Lin, 2004).
- **METEOR:** METEOR calculates alignment between the candidate and reference sentences by matching unigrams from the generated summary to those in the reference, taking into account stemming, synonyms, and paraphrases(Banerjee and Lavie, 2019).
- **BertScore:** BertScore calculates similarity scores by aligning the generated and reference summaries at the token level(Zhang et al., 2019).
- **BARTScore:** BARTScore is based on the BART model and evaluates the quality of generated text by assessing its fluency, coherence, and relevance(Yuan et al., 2021).
- **BLEURT:** BLEURT is an evaluation metric by Google Research that uses pre-trained transformers to assess machine-generated text quality(Sellam et al., 2020).

3.4 Human Evaluation

We recruited a total of 5 annotators from our summaries presented in random order and evaluated independently by annotators, who assess each summary based on three criteria: faithfulness, coherence, and relevance. These criteria are defined, and data is collected according to the guidelines in (Fabbri et al., 2021a). Coherence and relevance ratings are gathered using a 1 to 5 Likert scale, while faithfulness is rated as a binary value due to its binary nature. Our results show that the average pairwise agreement among annotators was 69% for faithfulness, 78% for coherence, and 88% for relevance, differing slightly from (Fabbri et al., 2021a).

3.5 Experimental Setup

All experiments were conducted on a machine equipped with high-performance hardware, ensuring efficient processing and accurate results with an Intel Xeon Gold 5218 CPU featuring 64 cores and 128 threads, coupled with four Nvidia Tesla V100 GPUs with VRAM memory of 30 GB per GPU card.

4 Experimental Results and Discussion

Table 2 presents the Performance comparison of various language models on BANS and BNLPC datasets based on Faithfulness(F), Coherence(C),

Table 1: Comparison of BANS with BNLPC dataset

Dataset	No of articles	Summary per article	No of summaries
BANS	19096	1	19096
BNLPC	200	3	600

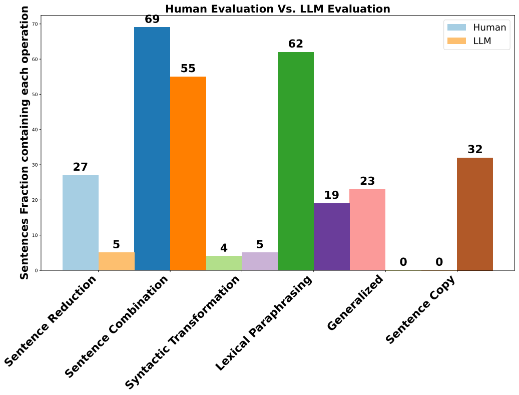

Figure 2: Human vs. LLM evaluation(GPT-4)

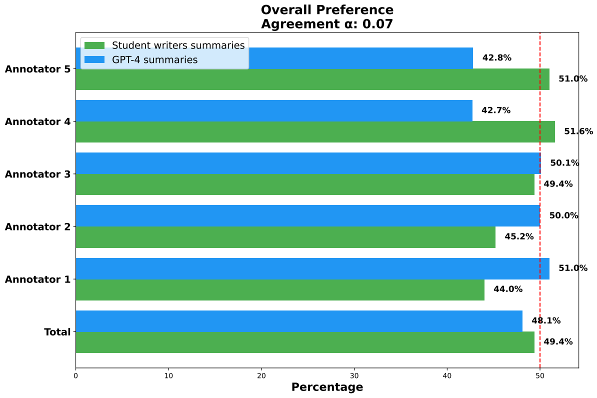

Figure 3: Human and LLM evaluation distribution

and Relevance(R). Among datasets, we found that zero-shot GPT-4 models outperformed the models with five, twenty, and fifty shots. Both datasets show higher faithfulness and relevance scores (0.95, 4.83, and 4.79 on BANS(Bhattacharjee et al., 2021) and 0.88, 4.81, and 4.39 in BNLPC (Haque et al., 2015).

Table 3 shows Kendall's tau rank correlations between human evaluation metrics and traditional evaluation methods. We observed significantly different trends in each dataset, warranting separate discussions. Within each model group, R-L and human evaluations showed higher correlations. Overall, reference-based metrics generally aligned better with human judgment scores across both datasets. While reference-free metrics are less affected by low-quality references, they primarily focus on measuring faithfulness.

Table 4 shows results for human-evaluated summaries on the BANS and BNLPC datasets. The zero-shot performance with GPT-4 and reference summaries are obtained from Table 2. Additionally, we observe that the difference between the student writer and GPT-4 in this evaluation is minimal.

Figure 2 displays the distribution of cut-and-paste operations, illustrating the fraction of sentences that contain each type of operation. We observed that student summaries performed well in sentence reduction, combination, lexical paraphrasing, and generalized or specification, and then the GPT-4 generated summaries. Again, we find that syntactic transformation is equally strong by both the writers evaluation and LLM. Finally, we find that the writers never directly copy an entire sentence from the article, whereas GPT-4 tends to do this more often.

Figure 3 presents the results of the paired comparison. A closer observation reveals significant variability in individual annotators' preferences, reflected in the low inter-annotator agreement (Krippendorff's alpha is 0.07). This indicates that the quality of generated summaries is approaching that of the student-written summaries, with comparisons largely influenced by each annotator's stylistic preferences.

4.1 Discussion

GPT-4 plays crucial role in enhancing the Bangla text summarization capabilities of LLM. These findings suggest that further research is necessary to deepen our understanding of how these factors influence the efficacy of GPT-4.

Evaluating high-performance LLMs is challenging as human assessments require more samples and precise measurements. (Fabbri et al., 2021b) noted difficulties in Bangla text summarization evaluation, suggesting fine-grained semantic units for better reference matching. However, our study finds existing reference summaries unreliable, and student writers' summaries often do not outperform LLMs. Thus, relying on reference summaries as ground truth is overly restrictive.

Despite potential quality issues, current benchmarks remain useful if applied correctly. As LLMs

Table 2: Comparison of LLMS on BANS and BNLPC based on Faithfulness(F) Coherence(C) Relevance(R).

Setting	**Models**	**BANS Data**			**BNLPC Data**		
		F	**C**	**R**	**F**	**C**	**R**
Zero-shot	GPT-3.5	0.78	4.33	4.39	0.19	4.55	3.81
	GPT-4	**0.95**	**4.83**	**4.79**	0.81	**4.81**	**4.39**
	OPT	0.66	2.54	3.41	0.48	2.67	3.62
	Llama2	0.73	4.02	4.26	0.71	3.69	3.79
	PaLM-2	0.78	4.54	3.90	0.86	4.65	3.91
Five-shot	GPT-3.5	0.91	4.15	4.66	0.81	4.68	3.99
	GPT-4	0.91	4.71	4.67	0.71	4.96	3.81
	OPT	0.87	3.45	4.13	0.81	4.65	3.12
	Llama2	0.86	3.64	4.33	0.77	4.80	4.01
	PaLM-2	0.86	3.86	3.15	0.85	0.85	0.85
Twenty-shot	GPT-3.5	0.88	4.52	4.34	0.76	4.03	2.70
	GPT-4	0.89	4.77	4.23	0.77	4.16	3.81
	OPT	0.76	2.65	3.50	0.55	2.61	3.42
	Llama2	0.80	4.02	4.26	0.81	3.90	3.87
	PaLM-2	0.80	4.24	4.90	0.76	4.27	3.34
Fifty-shot	GPT-3.5	0.84	3.88	4.33	0.70	4.88	3.88
	GPT-4	0.88	3.69	4.58	**0.88**	4.79	4.00
	OPT	0.94	3.69	4.24	0.74	4.72	3.88
	Llama2	0.86	3.69	4.33	0.88	4.80	3.01
	PaLM-2	0.89	3.69	4.34	0.69	4.69	3.03

Table 3: Kendall's tau correlation with human metrics Vs. automated summarization metrics

Metrics	**BANS**			**BNLPC**		
	F	**C**	**R**	**F**	**C**	**R**
Rouge-L	0.71	0.58	0.79	0.31	0.69	0.39
METEOR	0.48	0.55	0.66	0.32	0.69	0.31
BertScore	0.54	0.57	0.55	0.32	0.69	0.37
BARTScore	0.56	0.34	0.55	0.22	0.67	0.28
BLEURT	0.55	0.57	0.69	0.18	0.57	0.29

Table 4: Human Vs GPT-4(Zero-shot) Vs Existing summaries

Methods	**BANS**			**BNLPC**		
	F	**C**	**R**	**F**	**C**	**R**
Human	0.93	4.53	4.75	0.93	4.69	4.39
GPT-4	0.95	4.83	4.79	0.88	4.81	4.39
Existing	0.74	3.88	3.33	0.60	3.67	3.49

improve, grounding evaluations with clear user values in real-world applications will enhance accuracy and reduce assessment subjectivity.

5 Conclusions

In this study, we comprehensively evaluated five LLMs and human performance across two Bangla news summarization benchmarks. Our experiments demonstrated that state-of-the-art LLMs, particularly GPT-4, produce summaries comparable to student writers. These results highlight the critical importance of high-quality reference summaries for developing and evaluating summarization models. We discuss the issue of reference quality, comparing zero-shot, five-shot, twenty-shot, and fifty-shot performance. Finally, we showed the human evaluation, which is crucial, even when tackling the quality issue.

Limitations

Our work has several limitations, including a smaller sample size than other text summarization datasets. The availability of credible sources for Bangla news summaries was quite limited. Additionally, the news data was collected in only one language, and the dataset does not facilitate creative language interpretation for low-resource languages.

Acknowledgement

We extend our sincere gratitude to Noakhali Science and Technology University (NSTU) for funding this research. We also appreciate the dedication of the data annotators, whose support was crucial to this project's success.

References

Mohammad Abu Tareq Rony, Mohammad Shariful Islam, Tipu Sultan, Samah Alshathri, and Walid El-Shafai. 2024. Medigpt: Exploring potentials of conventional and large language models on medical data. *IEEE Access*, 12:103473–103487.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

S Banerjee and A Lavie. 2019. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of ACL-WMT*, pages 65–72.

Prithwiraj Bhattacharjee, Avi Mallick, and Md Saiful Islam. 2021. Bengali abstractive news summarization (bans): a neural attention approach. In *Proceedings of International Conference on Trends in Computational and Cognitive Engineering: Proceedings of TCCE 2020*, pages 41–51. Springer.

Radia Rayan Chowdhury, Mir Tafseer Nayeem, Tahsin Tasnim Mim, Md Saifur Rahman Chowdhury, and Taufiqul Jannat. 2021. Unsupervised abstractive summarization of bengali text documents. *arXiv preprint arXiv:2102.04490*.

Narjes Delpisheh and Yllias Chali. 2024. Improving faithfulness in abstractive text summarization with edus using bart (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23471–23472.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021a. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021b. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Md Majharul Haque, Suraiya Pervin, and Zerina Begum. 2015. Automatic bengali news documents summarization by introducing sentence frequency and clustering. In *2015 18th International Conference on Computer and Information Technology (ICCIT)*, pages 156–160. IEEE.

Md Nahid Hasan, Rafsan Bari Shafin, Marwa Khanom Nurtaj, Zeshan Ahmed, M Saddam Hossain Khan, Rashedul Amin Tuhin, and Md Mohsin Uddin. 2023. Implementation of bangla extractive update summarization task on busum-bnlp-dataset: A multi-document update summarization corpus. In *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pages 1–6. IEEE.

Md Shazid Islam, ASM Jahid Hasan, Md Saydur Rahman, Jubair Yusuf, Md Saiful Islam Sajol, and Farhana Akter Tumpa. 2023. Location agnostic source-free domain adaptive learning to predict solar power generation. In *2023 IEEE International Conference on Energy Technologies for Future Grids (ETFG)*, pages 1–6. IEEE.

Mohammad Shariful Islam, Mohammad Abu Tareq Rony, Mejdl Safran, Sultan Alfarhood, and Dunren Che. 2024a. Elevating driver behavior understanding with rknd: A novel probabilistic feature engineering approach. *IEEE Access*.

Mohammad Shariful Islam, Mohammad Abu Tareq Rony, and Tipu Sultan. 2024b. Gastrovrg: Enhancing early screening in gastrointestinal health via advanced transfer features. *Intelligent Systems with Applications*, page 200399.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Md Saiful Islam Sajol and ASM Jahid Hasan. 2024. Benchmarking cnn and cutting-edge transformer models for brain tumor classification through transfer learning. In *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, pages 1–6. IEEE.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.