An Evaluation of Bangla Text Summarization and Long Context Understanding Using LLMs

Submitted By

Abul Hasnat

ID: 0242220005343007

Department of Software Engineering

Daffodil International University

Submitted to

Afsana Begum

Assistant Professor & Coordinator M.Sc.

Department of Software Engineering

Daffodil International University

The enclosed project documentation constitutes the culminating requirement for the Master of Science degree in Software Engineering.

APPROVAL

This thesis/project/internship titled on **"An Evaluation of Bangla Text Summarization and Long Context Understanding Using LLMs"**, submitted by **Abul Hasnat**, ID: **0242220005343007** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Masters of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

---------- Chairman

Dr. Abdul Kadar Muhammad Masum
Professor
Department of Software Engineering
Daffodil International University

---------- Internal Examiner 1

Afsana Begum
Assistant Professor
Department of Software Engineering
Daffodil International University

---------- Internal Examiner 2

Dr. Marzia Ahmed
Assistant Professor
Department of Software Engineering
Daffodil International University

---------- External Examiner

Md. Tanvir Quader
Senior Software Engineer,
Technology Team,
a2i Program

DECLARATION

I hereby state that I have taken this thesis under the supervision of **Afsana Begum, Assistant Professor, Department of Software Engineering, Daffodil International University.** I also acknowledge that neither this thesis nor any part of this has been submitted elsewhere for the award of any degree previously by others.

Abul Hasnat

Abul Hasnat
ID: **0242220005343007**
Department of Software Engineering
Faculty of Science & Information Technology Daffodil International University

Certified by:

Afsana Begum
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

ACKNOWLEDGEMENT

In the name of Allah, this thesis has reached its culmination, a feat made possible by the invaluable contributions and insightful direction of several esteemed individuals. It is with profound sincerity that I express my deep appreciation to all who played a role in this academic undertaking. A particular debt of gratitude is owed to Assistant Professor Afsana Begum of the Department of Software Engineering at Daffodil International University, whose exemplary mentorship and unwavering guidance proved instrumental throughout the entire process. Her thoughtful supervision, characterized by both intellectual rigor and compassionate support, warrants special and emphatic acknowledgment. Her expertise and dedication have been a constant source of inspiration and direction.

Furthermore, I extend my heartfelt gratitude to my family, especially my parents, whose unwavering love and support have provided a bedrock for all my endeavors. My educators, who imparted knowledge and a love of learning, and my fellow students, whose camaraderie, insightful counsel, and generous assistance were vital in navigating the challenges inherent in this academic pursuit, also deserve profound thanks. Their collective encouragement formed an essential pillar of strength, without which the successful completion of this work would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

Abstract

In the realm of natural language processing, summarizing text in Bangla involves overcoming significant challenges, owing to the language's complex grammatical intricacies and rich linguistic variations. This study explores the effectiveness of state-of-the-art (SOTA) models in summarizing Bangla texts while preserving their essential meaning. The findings reveal that the Gemma series, particularly the Gemma 2 9B model, outperforms the latest Llama series model, Llama 3.1 8B, in capturing the essence of Bangla content, even as context length increases. The Mistral-based Microsoft FILM model, however, emerges as a formidable contender, closely rivaling both Gemma and Llama. Interestingly, despite its advanced architecture, the Llama 3 8B model struggles to match the performance of the Gemma 2B model.

Keywords: Bangla summarization, State-of-the-art models, Context length, Gemma, Llama, Mistral, Model performance.

CHAPTER 1

INTRODUCTION

1.1 Introduction

The art of condensing source material into concise yet comprehensive versions, known as summarization, plays a vital role across diverse fields - from consolidating news content to facilitating scholarly investigations and organizing digital information systems. To condense a text's essence, be it through extracting key sentences known as the extracting approach, or through an abstractive approach that rewrites the content for improved readability, particularly for complex texts [1].

Large text summarization, however, presents unique challenges, including the need to maintain coherence, context, and relevance over extended narratives [2]. As documents grow in length, the complexity of preserving the overall structure and meaning increases, making it difficult to generate concise yet comprehensive summaries [3].

The rapid proliferation of digital content in Bangla, one of the predominant linguistic systems, has not been matched by a corresponding growth in Natural Language Processing (NLP) research tailored to this rich linguistic tradition [3]. This disparity calls for intensified research efforts, particularly in developing robust Bangla text summarization systems. Research in this field has been limited, with only a handful of studies conducted. [4]

Modern Natural Language Processing has undergone a remarkable evolution through the advent of Transformer-based Large Language Models, which have set new standards of excellence across numerous applications such as language understanding, generation, and translation [5][6][7]. The field of text summarization has experienced remarkable progress due to Large Language Models' sophisticated ability to process and create text that emulates human composition.[7][3]. With faster and more capable hardware some of these models can handle huge context length [8], which is essential for long context summarization.

In this investigation, I have performed a comparative assessment of various SOTA pre-trained Large Language Models to examine their competence in summarizing Bengali text in a very minimal setting. The objective was to identify a model that could produce high-quality summaries while efficiently handling the nuances and complexities of the Bengali language.

1.2 Motivation

Processing and summarizing large context text are paramount in today's information-intensive era. Bangla's complex morphology, rich syntax, and the potential for long-range dependencies within texts pose significant challenges for traditional summarization techniques. To address these limitations, my research focuses on developing Bangla better and more efficient summarization models with better context-handling capability using very low resources. By emphasizing the ability to handle extensive textual information, I aim to unlock the valuable insights hidden within large-scale Bangla corpora, enabling more comprehensive and informative summarization.

1.3 Research Objective

- A better Bangla text summarization model using minimal resources.
- Assess the proficiency of leading-edge models in understanding extended Bengali passages and producing accurate summaries.
- Engage in a comparative assessment of the outcomes.

1.4 Research Gap

- A lack of in-depth studies on SOTA model adaptability after fine-tuning for Bangla text summarization tasks.
- The ability of Bangla models to process and understand extensive textual information is still underdeveloped.
- There is no evaluation of the performance of the newest SOTA model.

1.5 Dissertation Chapters and Structure

- Chapter 1: This initial chapter establishes the preliminary context of the research investigation.
- Chapter 2: This section encompasses a detailed review of contextual information, academic publications, and antecedent research contributions relevant to this investigation.
- Chapter 3: The methodologies and the architectural framework of the model are thoroughly presented in this chapter.
- Chapter 4: Contained within this section is a detailed account of the research methodology, observed results, and subsequent evaluation procedures.
- Chapter 5: Herein, examine the academic contributions to the field and explore potential avenues for future investigation.

CHAPTER 2
LITERATURE REVIEW

2.1 Literature Review

During the initial years of the twenty-first century, approaches to Bengali text summarization predominantly relied on statistical techniques and rule-based methodologies. Some of the pioneering work involved basic sentence extraction techniques that utilized features such as sentence position, word frequency, and sentence length. For instance, Islam et al. (2004) [9] developed Bhasa, a corpus-based Bangla information retrieval and summarization system. Their work provided early frameworks that relied on statistical measures for extracting important sentences from documents . Further contributions during this period included studies by Uddin et al. (2007) [8] who experimented with different extractive summarization techniques for Bangla, using sentence ranking and word scoring methods similar to the TF-IDF model These approaches represented the first steps toward automating the text summarization process in Bangla, although they were limited by small datasets and a lack of linguistic diversity in the corpora.

The 2010s marked the beginning of integrating more sophisticated machine-learning techniques into Bangla text summarization. One of the key developments was Sarkar et al. (2012) [10], who introduced sentence extraction methods that incorporated various statistical features, including sentence position, sentence length, and word co-occurrence frequency. Efat et al. (2013) [11] proposed a sentence scoring-based three-stage system for Bangla text summarization.

The year 2014 witnessed Uddin et al. [12] presenting a novel technique, the first of its kind, for summarizing multiple documents in the Bengali language. The authors found that by using a term frequency-based approach, they could extract the most significant content from Bengali documents. They used cosine similarity to determine the relationships between sentences.

Their research indicated that this approach mitigated several key difficulties inherent in multi-document summarization, specifically concerning comprehensive coverage, ease of understanding, accurate representation, and brevity. A heuristic-based method for sentence and word analysis was the foundation of the extractive Bengali text summarization technique proposed by Abujar et al. (2017) [13]. Key steps include stop word removal, stemming, word frequency calculation, and sentence position analysis. They tested their method on three Bengali documents and compared the system-generated summaries with human-generated ones, evaluated on a 1-5 scale. While specific accuracy metrics aren't provided, the authors demonstrate the effectiveness of their system through qualitative comparisons of news article summaries.

Ghosh et al. (2018) [14] introduced a rule-based summarization model using graph-based and corpus-level features, aggregate similarity assessment, bushy path algorithms, and in-sentence keyword analysis to improve the selection of relevant sentences for summarization. Their model was tested on 200 Bangla news documents. The authors recognized limitations such as the absence of synonym identification, which they aim to address in future research.

Sikder et al. (2019) [15] authors argued that current Bengali summarization tools, primarily extractive, struggle to capture the full theme of a document. Their proposed method combines mathematical rules (like term frequency and cosine similarity for sentence ranking and relevancy) with Bengali grammatical rules to refine the summarization process. This approach aims to address issues like redundancy and inconsistency in extracted summaries by simplifying sentences and joining related ones using grammatical rules. The authors believe their method, tested against human-generated summaries, provides a more accurate and comprehensive representation of Bengali text although no quantitative data is provided.

In 2019 Talukdar et al. (2019) [16] created an architecture utilizing a bi-directional recurrent neural network, incorporating LSTM layers for encoding and an attention-based mechanism for decoding. They trained their model on a dataset, which they collected themselves due to the lack of available online resources. The model was able to generate fluent and meaningful summaries from the given texts, achieving a training loss of 0.008. The authors acknowledge limitations in their model, such as the limited size of their dataset and the aptitude of the model for summarizing texts with a restricted quantity of words. Islam et al. (2021) [17] also reported the create a summarization model using bi-directional RNNs with LSTM. They claimed they had achieved a training loss of 0.007.

The advent of the Transformer model, as presented by Vaswani et al. (2017) [18], constituted a significant advancement in deep learning for natural language processing tasks. With the success of the Transformer architecture, a new methodology consisting of precursory training and subsequent optimization arose. Initially, pre-trained models undergo training on extensive textual datasets utilizing self-supervised learning objectives, for instance, language modeling or masked language modeling. Having undergone pre-training, these models are suitable for calibration on downstream applications using comparatively less task-specific data, yielding state-of-the-art performance in a wide variety of natural language processing domains.

A significant advancement in NLP was marked by the development of Transformer-based models such as the encoder-focused BERT (Devlin et al., 2018) [19], the decoder-centric GPT (Radford et al., 2018) [20], and the encoder-decoder hybrid T5 (Raffel et al., 2019) [21]. Leveraging these innovations, BanglaBERT and BanglaT5 (Bhattacharjee et al., 2022) [22][23] provide tailored solutions for a range of Bangla language processing tasks, effectively applying the core principles of BERT and T5. According to the analysis of Kabir et al. (2024) [23] the BanglaT5 model text summarization rouge score (Lin, 2004) [24] on a Bangla summarization dataset [25] is 0.28(R1), 0.11(R2), and 0.24(RL) after fine-tuning.

GPT-based models GPT 3.5 and Llama 2 13b [30]; summarization was also analyzed by the authors using zero-shot prompts. Mahfuz et al. (2024) [31] found that LLaMA-3-70B [28] showed strong performance in cross-lingual summarization surpassing the fine-tuned BanglaT5 model.

CHAPTER 3

METHODOLOGY

3.1 Methodological Framework

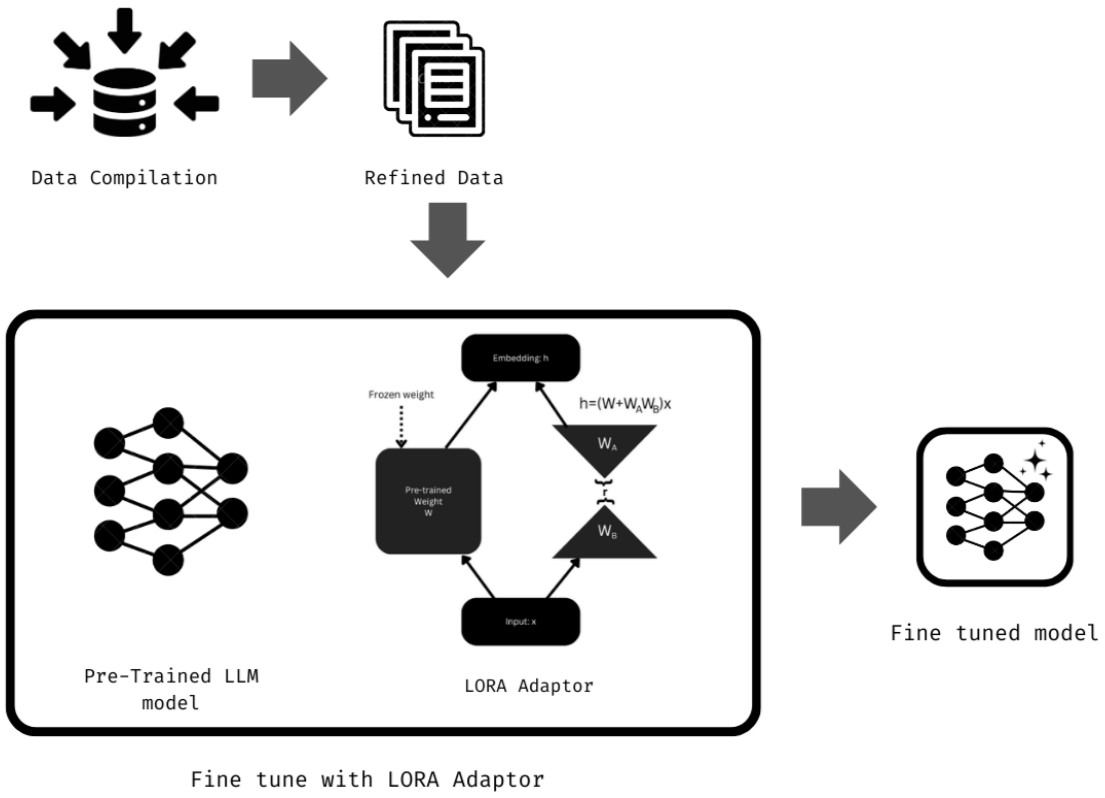

Figure 1: The Methodological Framework

This study employs a sophisticated fine-tuning methodology utilizing the Low-Rank Adaptation (LoRA) approach (Hu et al. 2021) [33] The framework consists of three primary components: a pre-trained model architecture, a LoRA adapter mechanism, and the resulting fine-tuned model.

The initial component comprises a pre-trained LLM coupled with processed input data. This pre-trained model is the foundation for subsequent adaptation and represents the accumulated knowledge from the initial training phase.

The core of this methodology revolves around the LoRA adapter implementation, which introduces a specialized weight modification mechanism. This adapter incorporates two essential matrices, WA and WB, configured in a low-rank decomposition arrangement. The forward pass through the adapter processes the input embedding h through the following mathematical transformation:

$$h = (W + W_A W_B)x$$

where:

- x represents the input features
- W denotes the frozen pre-trained weights
- W_A and W_B constitute the trainable LoRA parameters

Within this framework, the original model parameters remain unaltered, while adaptive learning occurs exclusively through LoRA configurations. This methodology safeguards the model's baseline knowledge while permitting task-oriented modifications. The final architecture represents a fusion of preserved foundational elements and optimized LoRA parameters, achieving targeted adaptability without compromising the underlying structure. This methodological approach offers several advantages, including computational efficiency, preservation of pre-trained knowledge, and effective task-specific adaptation while minimizing the risk of catastrophic forgetting.

3.2 Data Compilation

The empirical data employed for this investigation was sourced from the work of Hasan et al. (2021) [25], titled "XL sum." I collected it from Huggingface. It consists of over one million article-summary pairs, which were professionally annotated from BBC articles. The dataset covers 44

languages, making it one of the most diverse summarization datasets available. Given my focus on Bangla text summarization, I am exclusively working with the Bangla portion of this dataset. The dataset is structured with five columns and contains 10,100 rows in the Bangla segment. For training purposes, I am utilizing 8,100 rows. The remaining 2,020 rows are evenly divided between the test and validation sets, each containing 1,010 rows.

3.3 Systematic Data Refinement

The original dataset contained five distinct fields: 'id', 'url', 'title', 'summary', and 'text'. Given the specific requirements of this investigation, only the 'summary' and 'text' fields were retained, with the extraneous columns being eliminated. Subsequently, a prompt template was employed to restructure the dataset, aligning the paired 'summary' and 'text' entries from each record into a prescribed format. The resultant formatted content was consolidated into a new column designated as "final."

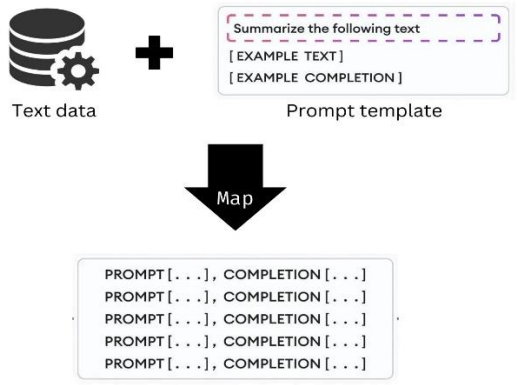

Figure 2: Mapping the dataset

```
alpaca_prompt = """You are an abstructive text summarizer in Bangla language.
Below is an instruction in Bangla,
paired with an input that provides the context for summarization.

### Instruction:
নিচের লেখাটির সারমর্ম লিখো

### Input:
{}

### Response:
{}"""
```

Figure 3: The prompt template that has been used

```
You are an abstructive text summarizer in Bangla language.
Below is an instruction in Bangla,
paired with an input that provides the context for summarization.

### Instruction:
নিচের লেখাটির সারমর্ম লিখো

### Input:
বাংলাদেশে গত কয়েকবছর ধরে ফুডপান্ডা, হাঙ্গরি নাকি বা পাঠাও ফুডের মত অনলাইন ডেলিভারি সার্ভিস ব্যবহার করে রেস্টুরেন্ট থেকে খাবার অর্ডার করার চল তৈরি হলেও শাকসবজি, মাছ-মাংস বা রান্নার জন্য প্রয়োজনীয় বাজার সদাই কেনার ক্ষেত্রে অনলাইন ডেলিভারি সিস্টেম ব্যবহার করার প্রবণতা মানুষের মধ্যে বেশ কম ছিল। কিন্তু সাধারণ ছুটির সময়কার দুই মাসে মানুষের এই প্রবণতা হঠাৎ করেই বহুগুণ বেড়েছে। একইসাথে বেড়েছে ইলেকট্রনিক পণ্য, মোবাইল ফোন বা দৈনন্দিন ব্যবহারের পণ্য অনলাইনে কেনার প্রবণতাও। তবে মানুষের মধ্যে অনলাইনে অর্ডার করার চাহিদা বাড়লেও মানুষের প্রত্যাশার সাথে সামঞ্জস্য রেখে অনলাইনে এই ধরণের সেবাদানকারী প্রতিষ্ঠানগুলো সেবা দিতে পারছে না বলে হতাশা প্রকাশ করতে দেখা গেছে সাধারণ মানুষকে। পণ্য ডেলিভারির সময় নিয়ে অভিযোগ তুলে ঢাকার বাসিন্দা উম্মে হানি সালমা বলেন, "শাক-সবজি, মাছ-মাংসের মত পণ্য অর্ডার দিলে আগে একদিন, খুব বেশি হলে তিনদিন সময় নিতো। কিন্তু সাধারণ ছুটির মধ্যে অর্ডার ডেলিভারি করতে ১০-১২দিন পর্যন্ত সময় নিতে দেখেছি।" আরেকজন সেবা গ্রহীতা মাকসুদা মোমিন জানান তিনি অনলাইনে একটি ব্র্যান্ডের পণ্য অর্ডার করলেও তাকে ডেলিভারি দেয়া হয় আরেকটি ব্র্যান্ডের পণ্য। "আর তা নিয়ে কাস্টমার কেয়ারে অভিযোগ করার পর সঠিক পণ্যটি পেতে দীর্ঘসময় অপেক্ষা করতে হয়, ভোগান্তিও কম হয়নি।" অথচ করোনাভাইরাস ম হামারির সময় মানুষের অনলাইনে কেনাকাটার চাহিদা বৃদ্ধির সুযোগ কাজে লাগানোর যথেষ্ট সুযোগ তৈরি হয়েছিলো এই খাতের প্রতিষ্ঠানগুলোর। 'দারাজ' বাংলাদেশের সবচেয়ে বড় ই-কমার্স সাইট চাহিদামত সেবা দিতে না পারার কারণ কী? বাংলাদেশে অনলাইনে নিত্য প্রয়োজ নীয় পণ্য কেনার অন্যতম জনপ্রিয় প্রতিষ্ঠান চালডালের চিফ অপারেটিং অফিসার জিয়া আশরাফ বলেন হঠাৎ বাড়তি চাহিদার সাথে খাপ খাইয়ে গ্রাহকদের সেবা দিতে না পারার অন্যতম প্রধান কারণ প্রতিষ্ঠানের যথেষ্ট পরিমাণ সক্ষমতা না থাকা। জিয়া আশরাফ বলেন, "সাধার ণ ছুটি শুরু হওয়ার আগ পর্যন্ত ঢাকায় দৈনিক গড়ে আড়াই হাজার অর্ডার আসতো আমাদের, আর আমাদের সক্ষমতা ছিল দৈনিক সাড়ে তিন হাজার মানুষকে অর্ডার দেয়ার।" "কিন্তু ছুটি শুরু হওয়ার পর প্রথম কয়েকদিন আমরা দেখলাম দিনে ১৬ থেকে ১৭ হাজারের মত অর্ডার আসছে, অর্থাৎ আমরা যেই পরিমাণ অর্ডার প্রতিদিনে ডেলিভারি দিতে পারি তারও চার-পাঁচগুণ বেশি।" এই কারণে অনেক অর্ডার ডেলিভারি দিতে সাধারণ সময়ের চেয়ে বেশি সময় লেগেছে বলে মন্তব্য করেন জিয়া আশরাফ। এছাড়া সাধারণ ছুটির সময় মানুষের মধ্যে নিত্য প্রয়োজনীয় পণ্য অতিরিক্ত পরিমাণে কিনে মজুদ করে রাখার প্রবণতার কারণে পণ্যের স্টক শেষ হয়ে যাওয়ায় এই সমস্যা আরো বেশি প্রকট হয়েছে বলে মনে করেন জিয়া আশরাফ। তবে মি. আশরাফ জানান চাহিদার সাথে সামঞ্জস্য রেখে নতুন ওয়্যারহাউজ খোলায় এ বং নতুন জনবল নিয়োগ করায় অর্ডার ডেলিভারি করার ক্ষেত্রে আগের মত দীর্ঘ সময় লাগছে না তার প্রতিষ্ঠানের। বাংলাদেশে দ্রুত বাড়ছে অনলাইনে নিত্য প্রয়োজনীয় পণ্য কেনার প্রবণতা 'ঋণ ও আর্থিক সুবিধা না পাওয়া অন্যতম প্রধান সমস্যা' চালডালের চিফ অপারেটিং অফিসার জিয়া আশরাফ ই-কমার্স অ্যাসোসিয়েশন বাংলাদেশের একজন পরিচালকও। বাংলাদেশে এই ধরণের ই-কমার্স প্রতিষ্ঠানের যথাযথভাবে ব্যবসা করতে না পারার অন্যতম প্রধান কারণ হিসেবে ব্যাংক এবং আর্থিক প্রতিষ্ঠানগুলোর নীতিমালাকে দোষারোপ করেন তিনি। "এক টা ই-কমার্স স্টার্ট আপকে বাংলাদেশে কোনো ব্যাংক বা আর্থিক প্রতিষ্ঠান ঋণ দিতে চায় না বা সহায়তা করতে চায় না। ফলে এই প্রতিষ্ঠানগুলোও তাদের প্রয়োজনীয় ফান্ডিং পায় না উন্নতি করার জন্যে। অর্থায়নের সহজলভ্য ব্যবস্থা থাকলে ও সময়মতো ফান্ডিং পেলে চাহিদা তৈরি হওয়ার সাথে সাথেই প্রয়োজনীয় বিনিয়োগ করে পরিস্থিতি অনুযায়ী পদক্ষেপ নেয়া সম্ভব হতো, সেক্ষেত্রে গ্রাহকদের এত ভোগান্তি পোহাতে হতো না।" অর্থায়ন সহজলভ্য হলে ই-কমার্স প্রতিষ্ঠানগুলো সহজে প্রয়োজনীয় বিনিয়োগ পেতে সক্ষম হবে এবং সাধারণ মানুষও এ সব প্রতিষ্ঠান থেকে পাওয়া সেবা নিয়ে সন্তুষ্ট হবেন বলে আশা প্রকাশ করেন জিয়া আশরাফ। বর্তমানে চাহিদার তুলনায় সেবা দেয়ার সক্ষমতা অনেক কম বলে মন্তব্য করেন তিনি। "ঢাকায় যদি এখন নিউ ইয়র্কের মত পরিস্থিতি তৈরি হয়, অর্থাৎ সবাইকে ঘরে থাকতে হয় এবং দোকানপাট বন্ধ করে দেয়া হয় - তাহলে প্রতিদিন প্রায় এক লাখের মত মানুষকে হোম ডেলিভারির মাধ্যমে শাক-সবজি, চাল ডাল, মাছ-মাংসের মত পণ্য সরবরাহ করতে হবে। কিন্তু আমাদের সবগুলো ই-কমার্স প্রতিষ্ঠান মিলে সর্বোচ্চ ৪০ হাজার মানুষের কাছে দিনে পণ্য পৌঁছে দিতে পারবো।"

### Response:
মার্চ মাসের শেষদিক থেকে জুনের শুরু পর্যন্ত দুই মাসের বেশি সময় বাংলাদেশে সাধারণ ছুটি থাকায় অধিকাংশ দোকানপাট বন্ধ থাকায় এবং মানুষের ঘরের ভেতরে থাকার কারণে নিত্য প্রয়োজনীয় পণ্য অনলাইনে অর্ডার করার প্রবণতা বেড়েছে। তবে অনলাইনে অর্ডারের প রিমাণ অনেক বাড়লেও সেই অনুপাতে সেবা দিতে হিমশিম খেতে হয়েছে অনলাইনে নিত্য প্রয়োজনীয় পণ্য কেনা ও ডেলিভারি দেয়ার প্রতিষ্ঠানগুলোকে।
```

Figure 4: A sample of formatted text prompt after mapping.

3.4 Model Description

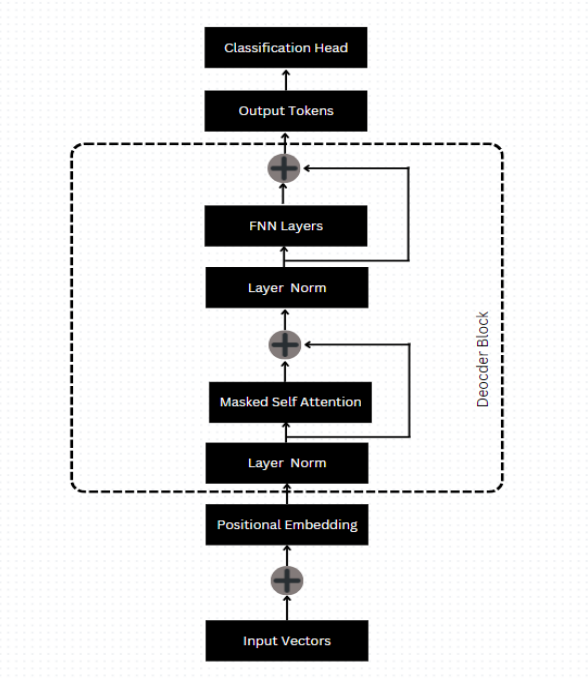

Figure 5: Simple Decoder only transformer architecture

In this experiment, I've used five SOTA decoder-only models.

Gemma 2

Gemma 2 (G Team et al., 2024) [26] developed by Google, is an advanced open large language model (LLM) designed to enhance text generation tasks. This model exists in 9 billion and 27 billion sizes, with both a foundational and a guideline-tuned version available for each. The model is built on the foundation of Google Deepmind's Gemini and features a context length of 8,192 tokens. Gemma 2 includes sliding window attention, logit soft-capping, knowledge distillation, and model merging, collectively improving its performance and efficiency. In this experiment, I used the Gemma 2 9 billion model.

Gemma

The original Gemma (G Team et al., 2024) [27] model family represents a suite of efficient, cutting-edge open models developed by Google, engineered to address a range of text generation applications, including question-answering, summarization, and reasoning. The architectural framework of these models utilizes a decoder-only approach for text processing, with their weights being openly accessible in both pre-trained and instruction-enhanced iterations. The relatively small size of Gemma models makes them suitable for deployment in environments with limited resources, democratizing access to advanced AI capabilities2. In this experiment, I used the Gemma 2 billion model.

LLama 3

Within the progression of Meta's LLama large language model series, LLama 3, as chronicled by Dubey et al. (2024) [28], emerges as the most recent development. The model architecture is deployed in two distinct configurations, featuring parameter counts of 8 billion and 70 billion respectively, with both base and instruction-tuned variants. LLama 3 introduces a new tokenizer that expands the vocabulary size to 128,256 tokens, enhancing its multilingual capabilities and efficiency in text encoding. The model also incorporates Grouped-Query Attention (GQA) to handle longer contexts more effectively. I used the Llama-3 8 billion model for this experiment.

LLama 3.1

With several enhancements, Meta introduces LLama 3.1 (Dubey et al., 2024) [28] based on the architecture of LLama 3. It is available in three sizes. LLama 3.1 supports a context length of 128,000 tokens and is optimized for multilingual dialogue use cases. Additionally, LLama 3.1 includes tool usage capabilities, allowing it to generate calls for specific tools like search and code execution. I used the Llama-3.1 8 billion model for this experiment.

Microsoft FILM

Microsoft FILM [29] is a 32K-context large language model designed to overcome the “lost-in-the-middle” problem. FILM-7B, a notable variant, is trained from Mistral [32] specifically Mistral-7B-Instruct-v0.2 using Information-Intensive (In2) Training. The architecture demonstrates superior capabilities in processing extended contextual information while excelling in tasks requiring limited context analysis. FILM-7B is particularly effective in tasks requiring extensive context utilization, thus rendering it a valuable asset for diverse natural language processing applications.

3.5 Training Setup

The research methodology employed supervised fine-tuning techniques, executed on a Kaggle-hosted Nvidia T4 graphics processing unit with 15GB of VRAM. The training protocol implemented 120 computational iterations within a single training epoch.

CHAPTER 4

Result and Discussion

4.1 Evaluation Matrices

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) was developed by Lin (2004) [24], ROUGE has become one of the most widely used evaluation metrics in the context of natural language processing, notably for summarization applications.

There are a few different variants of ROUGE but the two that have been used for the evaluation of this study are:

- ROUGE-1: This metric quantifies the unigram overlap between the generated and the ideal summaries. ROUGE-1 performance is quantifiable via precision, recall, and the harmonic mean of these, the F1 score.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Recall} = \frac{\text{Number of overlapping unigrams}}{\text{Total unigrams in reference summary}}$$

$$\text{Precision} = \frac{\text{Number of overlapping unigrams}}{\text{Total unigrams in candidate summary}}$$

- ROUGE-L: Through the computation of shared sequential patterns between proposed and reference summaries, this analytical framework evaluates both the structural composition and narrative coherence of the generated content.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{LCS Recall} = \frac{\text{LCS length}}{\text{Length of reference summary}}$$

$$\text{LCS Precision} = \frac{\text{LCS length}}{\text{Length of candidate summary}}$$

4.2 Analytical Evaluation

I've taken a total sample of 100 with various ranges of string tokens from the 'test' split of the dataset for the testing. Here is the result of all five models across multiple context lengths.

Rouge 1

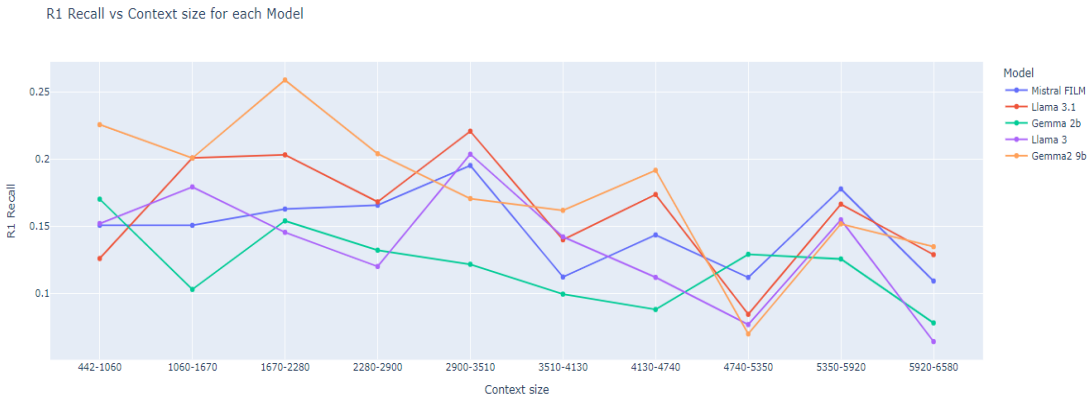

Figure 6: R1 Recall vs context length for various model

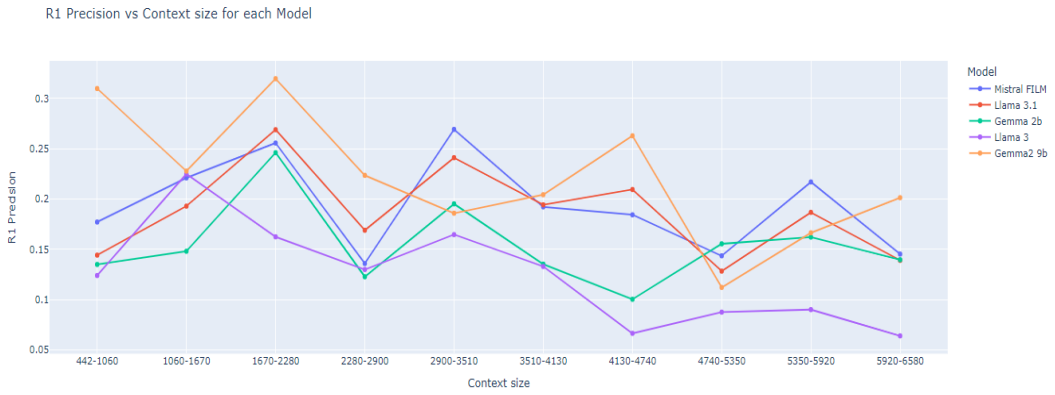

Figure 7: R1 precision vs context length for various model

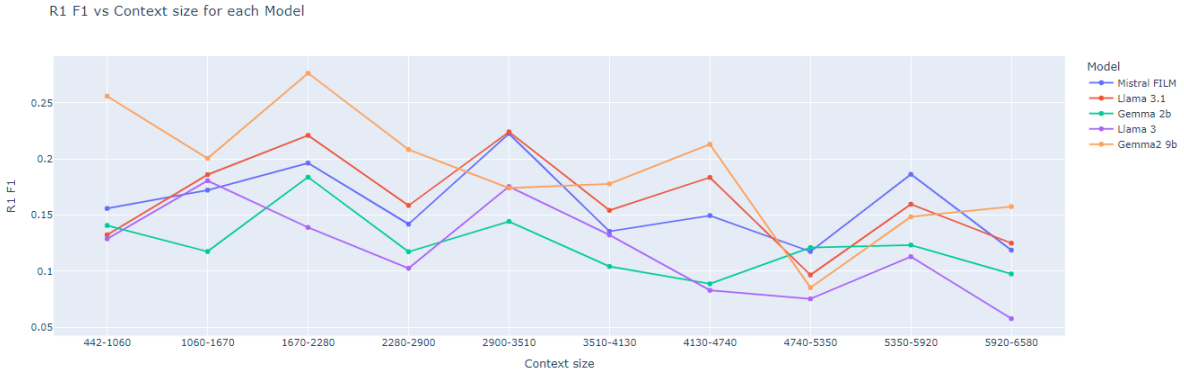

Figure 8: R1 F1 vs context length for various model

Gemma 2 9b, which performs the best for shorter contexts, shows a noticeable decline in precisions, F1, and recall as the context size grows larger. As the magnitude of contextual information expands, Gemma 2b exhibits a diminishing performance trajectory. Mistral FILM and Llama 3 .1 8b show some improvements at certain longer context sizes, but generally, their performance fluctuates and tends to decrease or stabilize at a lower level. Llama 3 8b struggles most in a larger context it even performs badly than the 2 billion model of Gemma.

Overall, models tend to struggle with larger context sizes, decreasing precision, recall, and F1.

	Model	R1 Precision	R1 Recall	R1 F1
0	Gemma 2b	0.153978	0.120111	0.123750
1	Gemma2 9b	0.221329	0.176950	0.189677
2	Llama 3	0.124687	0.135009	0.118648
3	Llama 3.1	0.187373	0.161199	0.164028
4	Mistral FILM	0.194095	0.147957	0.159568

Figure 9: Average R-1 score comparison table

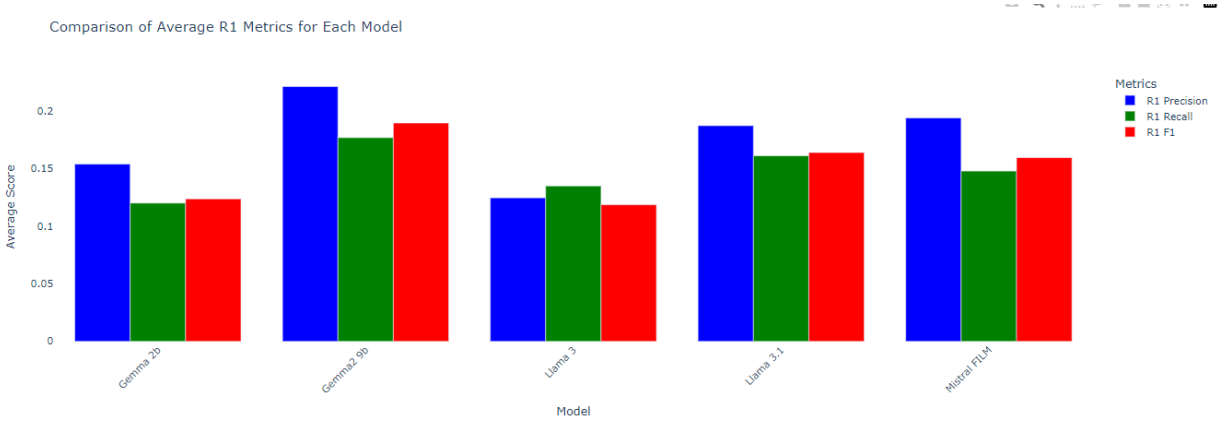

Figure 10: Comparison of Average R1 metrics

Gemma 2 9b stands out with the highest precision and recall, achieving robust F1 scores as well. In contrast, Llama 3 8b shows overall lower performance, with balanced but suboptimal precision, recall, and F1 scores. Llama 3.1 8b demonstrates improvement over Llama 3 8b, particularly in precision, with slightly enhanced recall and F1 scores. Mistral FLM performs well in precision but has lower recall and F1 scores, echoing the performance trends of Llama 3.1.

Rouge L

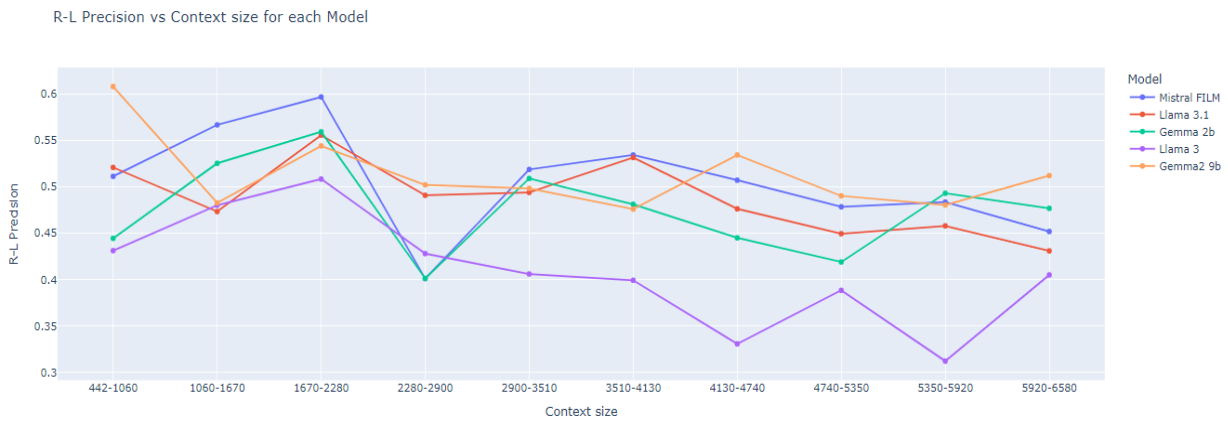

Figure 11: R-L precision vs context length for various

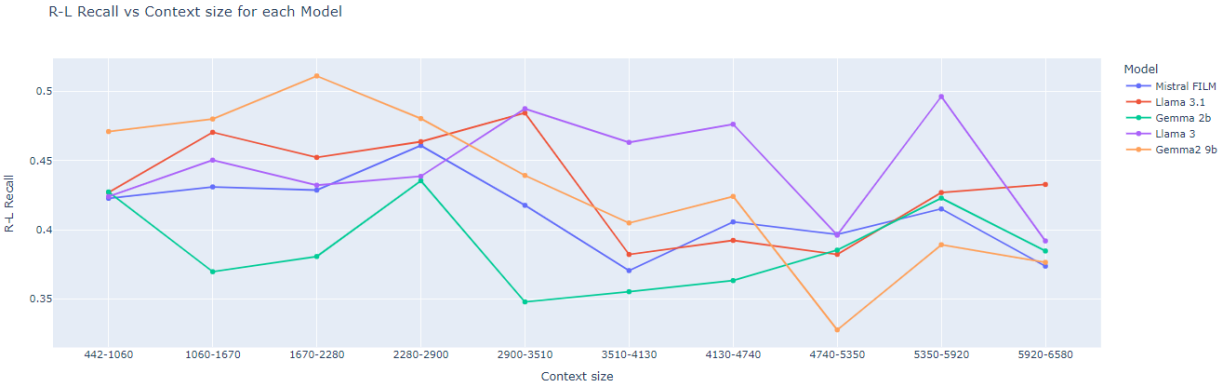

Figure 12: R-L Recall vs context length for various

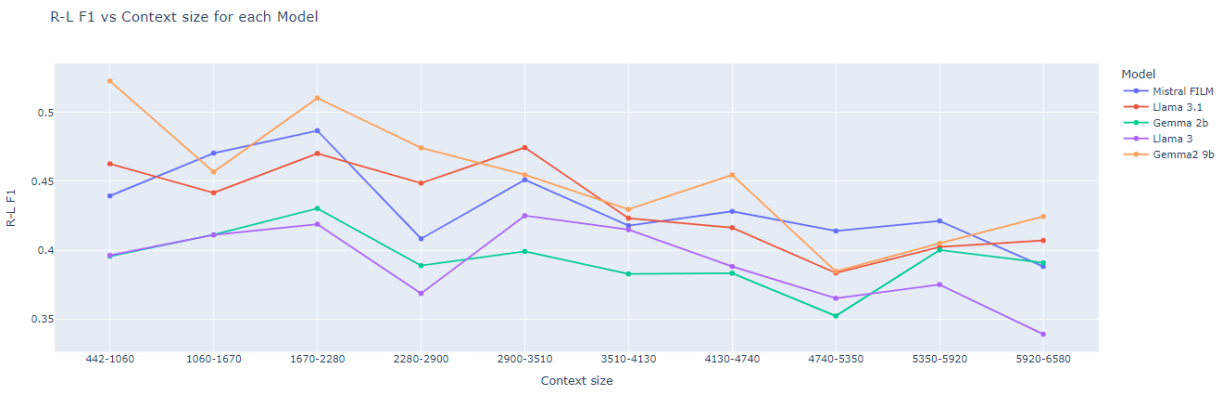

Figure 13: R-L F1 vs context length for various

	Model	R-L Precision	R-L Recall	R-L F1
0	Gemma 2b	0.475575	0.387345	0.393385
1	Gemma2 9b	0.512990	0.430421	0.451765
2	Llama 3	0.409070	0.445607	0.390156
3	Llama 3.1	0.488226	0.431421	0.433025
4	Mistral FILM	0.505167	0.412302	0.432554

Figure 14: Average R-L score comparison table

The R-L precision performance of models without Llama 3 8b, is quite similar in a large context. The Gemma 2 8b starts well for all metrics but the result degrades as the context size increases. The Llama 3 8b has better results in recall for large contexts but the lowest result in both F1 and precision. That means the summary might include extra, non-essential information or noise.

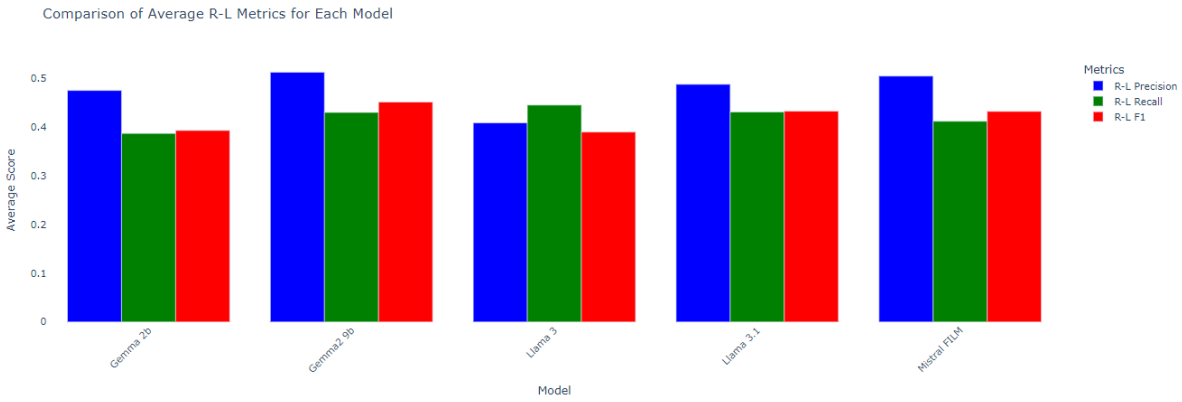

Figure 15: Comparison of various model R-L score

On average R-L scores the Gemma 2 9b also stand out from others. Performance assessment revealed scores of 0.512990 for average precision, while recall and F1 measurements reached 0.430421 and 0.451765 respectively. FILM and Llama 3.1 8b also show close results with the Gemma 2 9b model. The Llama 3 8b model is the worst among all five models.

CHAPTER 5
CONCLUSION

5.1 Outcomes and Scholarly Contributions

This study undertakes a systematic comparison of premier models in Bangla text summarization, analyzing their performance characteristics across different textual expanses. The findings indicate that as the context length increases, the performance of SOTA models tends to decline. Additionally, this work highlights that the Gemma series, particularly the Gemma 2 9B model, outperforms the latest Llama series model, Llama 3.1 8B, in Bangla summarization tasks. Interestingly, the Mistral-based Microsoft FILM model also demonstrates competitive performance, closely rivaling both the Gemma and Llama models. Notably, the Llama 3 8B model showed weaker performance compared to the Gemma 2B model.

5.2 Future Work

This research utilized a relatively small-scale model, such as an 8-billion parameter model. To fully comprehend the potential of Bangla text summarization and long context understanding, future research should explore a wider range of model sizes, including significantly larger models also with more diverse datasets. Additionally, investigating the impact of different and new architectures, such as mamba-based models, could provide valuable insights. Finally, thorough evaluations, incorporating both quantitative measurements and qualitative insights, are required for enhanced comprehension.

REFERENCES

[1] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

[2] Koh, Huan & Ju, Jiaxin & Liu, Ming & Pan, Shirui. (2022). An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics. ACM Computing Surveys. 55. 10.1145/3545176.

[3] Zhang, H., Yu, P. S., & Zhang, J. (2024). A Systematic Survey of Text Summarization: From Statistical Methods to Large Language Models. arXiv preprint arXiv:2406.11289.

[4] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

[3] M. S. Mahmud, M. T. Islam, A. J. Bonny, R. K. Shorna, J. H. Omi and M. S. Rahman, "Deep Learning Based Sentiment Analysis from Bangla Text Using Glove Word Embedding along with Convolutional Neural Network," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2022, pp. 1-6, doi: 10.1109/ICCCNT54827.2022.9984392.

[4] S. M. Tasnimul Hasan, M. A. Rahman, M. M. Hasan, M. R. Hasan and M. M. Hoque, "Advancing Abstractive Bangla Text Summarization: A Deep Learning Approach Using Seq2seq Encoder- Decoder Model and T5 Transformer," 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI), Dhaka, Bangladesh, 2023, pp. 1-6, doi: 10.1109/STI59863.2023.10464712.

[5] Qin, L., Chen, Q., Feng, X., Wu, Y., Zhang, Y., Li, Y., ... & Yu, P. S. (2024). Large language models meet nlp: A survey. arXiv preprint arXiv:2405.12819.

[6] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

[7] Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On Context Utilization in Summarization with Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2764–2781, Bangkok, Thailand. Association for Computational Linguistics.

[8] Uddin, M. N., & Khan, S. A. (2007, December). A study on text summarization techniques and implement few of them for Bangla language. In 2007 10th international conference on computer and information technology (pp. 1-4). IEEE.

[9] Islam, M. T., & Al Masum, S. M. (2004, December). Bhasa: A corpus-based information retrieval and summariser for bengali text. In Proceedings of the 7th International Conference on Computer and Information Technology.

[10] Sarkar, K. (2012). Bengali text summarization by sentence extraction. arXiv preprint arXiv:1201.2240.

[11] Efat, M. I. A., Ibrahim, M., & Kayesh, H. (2013, May). Automated Bangla text summarization by sentence scoring and ranking. In 2013 International Conference on Informatics, Electronics and Vision (ICIEV) (pp. 1-5). IEEE.

[12] Uddin, M. A. A Multi-Document Text Summarization for Bengali.

[13] Abujar, S., Hasan, M., Shahin, M. S. I., & Hossain, S. A. (2017, July). A heuristic approach of text summarization for Bengali documentation. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-8). IEEE.

[14] Ghosh, P. P., Shahariar, R., & Khan, M. A. H. (2018). A rule based extractive text summarization technique for Bangla news documents. International Journal of Modern Education and Computer Science, 10(12), 44.

[15] Sikder, R., Hossain, M. M., & Robi, F. M. R. H. (2019). Automatic text summarization for Bengali language including grammatical analysis. International Journal of Scientific & Technology Research, 8(6).

[16] Talukder, M. A. I., Abujar, S., Masum, A. K. M., Faisal, F., & Hossain, S. A. (2019, July). Bengali abstractive text summarization using sequence to sequence RNNs. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[17] Islam, M. M., Islam, M., Masum, A. K. M., Abujar, S., & Hossain, S. A. (2021). Abstraction based bengali text summarization using bi-directional attentive recurrent neural networks. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 2 (pp. 317-327). Springer Singapore.

[18] Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[20] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[21] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

[22] Bhattacharjee, A., Hasan, T., Ahmad, W. U., Samin, K., Islam, M. S., Iqbal, A., ... & Shahriyar, R. (2021). BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. arXiv preprint arXiv:2101.00204.

[23] Bhattacharjee, A., Hasan, T., Ahmad, W. U., & Shahriyar, R. (2022). BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. arXiv preprint arXiv:2205.11081.

[24] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[25] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4693–4703, Online. Association for Computational Linguistics.

[26] Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., ... & Ronstrom, S. (2024). Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118.

[27] Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., ... & Kenealy, K. (2024). Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

[28] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

[29] An, S., Ma, Z., Lin, Z., Zheng, N., & Lou, J. G. (2024). Make Your LLM Fully Utilize the Context. arXiv preprint arXiv:2404.16811.

[30] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[31] Mahfuz, T., Dey, S. K., Naswan, R., Adil, H., Sayeed, K. S., & Shahgir, H. S. (2024). Too Late to Train, Too Early to Use? A Study on Necessity and Viability of Low-Resource Bengali LLMs. arXiv preprint arXiv:2407.00416.

[32] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825.

[33] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

2nd time

ORIGINALITY REPORT

4%	3%	1%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
2	www.coursehero.com Internet Source	1%
3	Yang Li. "Ontological Knowledge Learning for Relation Extraction", University of Technology Sydney (Australia), 2024 Publication	1%
4	www.slideshare.net Internet Source	<1%
5	nikosfl.github.io Internet Source	<1%
6	Ana M Maitin, Alberto Nogales, Sergio Fernández-Rincón, Enrique Aranguren et al. "Application of large language models in clinical record correction: a comprehensive study on various retraining methods", Journal of the American Medical Informatics Association, 2024 Publication	<1%
7	docslib.org Internet Source	<1%
8	deepai.org Internet Source	<1%