Automatic MCQ with Answer Generation for Bangla Medium SSC-Level Students

by

Md Abu Tarabin Surzo
22101349
A. K. M Nihalul Kabir
23341032
Sm Azmain Faysal
22101576
Lawrence Amlan Gomes
23341031
Ariana Haque Ami
22101080

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2025

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md Abu Tarabin Surzo

22101349

A. K. M Nihalul Kabir

23341032

Sm Azmain Faysal

22101576

Lawrence Amlan Gomes

23341031

Ariana Haque Ami

22101080

Approval

The thesis titled "Automatic MCQ with Answers Generation for Bangla Medium SSC level students" submitted by

1. Md Abu Tarabin Surzo (22101349)
2. A. K. M Nihalul Kabir (23341032)
3. Sm Azmain Faysal (22101576)
4. Lawrence Amlan Gomes (23341031)
5. Ariana Haque Ami (22101080)

Of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 18, 2025.

Examining Committee:

Supervisor:
(Member)

FYS

Dr. Farig Yousuf Sadeque

Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi

Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

It has almost become a tradition for Bangladeshi students to join coaching centers before exams only to take part in Multiple Choice Question model tests. Since these coaching centers charge a substantial amount of fees for these MCQ exams, students not only waste their money but also their valuable time attending them. This widespread practice of pre-exam testing is specially noticeable in Bengali medium students. However, there has not been much research done in this regard to solve this issue. Our paper aims at solving this problem by developing a system that will not only automatically generate MCQs but also be able to predict the answers of inputted MCQs. The proposed framework in this paper incorporates natural language processing (NLP) functions for extracting and cleaning academic Bengali textual data from NCTB verified text books and utilizes a hybrid form of Graph-based Retrieval Augmented Generation (GraphRAG) to produce appropriate multiple choice questions along with the capability to predict answer of a given MCQ. This would support Bengali medium SSC candidates in designing their own MCQ model tests. Our research result demonstrates the demand for high-quality Bengali embedding models as well as provides implementation strategies for any future RAG-based automated educational tool designed for Bengali language. The MCQ generation framework developed by our team, would be able to provide teachers and students with numerous practice MCQs whereas the answer prediction pipeline, could help students study more efficiently. Integration of our system can lead towards effective learning environments in Bengali medium institutions.

Keywords: Secondary School Certificate; Automatic MCQ Question Generation; Natural Language Processing; Bangla Text Processing; NCTB; Retrieval Augmented Generation (RAG); GraphRAG; Answer Extraction.

Acknowledgement

Firstly, all praise to the Great Almighty for whom our thesis have been completed without any major interruption.
Secondly, to our supervisor Farig Yousuf Sadeque sir for his kind support and advice in our work. He helped us whenever we needed help.
And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

α Alpha (Greek Symbol)

β Beta (Greek Symbol)

ϕ Phi (Greek Symbol)

AI Artificial Intelligence

AIF Artificial Intelligence Feedback

$BERT$ Bidirectional Encoder Representations from Transformers

BGS Bangladesh and Global Studies

DB Database

$Distractor$ Incorrect Options of an MCQ

GPT Generative Pretrained Tranformer

LLM Large Language Model

MCQ Multiple Choice Questions

ML Machine Learning

$NCTB$ National Curriculum and Textbook Board

NER Named Entity Recognition

NLI Natural Language Inference

NLP Natural Language Processing

NN Neural Network

OCR Optical Character Recognition

PDF Portable Document Format

RAG Retrieval Augmented Generation

SSC Secondary School Certificate

TPU Tensor Processing Unit

UTF Unicode Transformation Format

Chapter 1

Introduction

In Bangladesh, students are bound to sit for the Secondary School Certificate (SSC) examination at the end of their 10th grade. It is a high stake, public examination where multiple choice questions (MCQs) usually carry around 40% of the total marks. However, in Bengali medium education system, students often have limited cheap or free high quality resources for unlimited practice of MCQs. As a result, this forces the students to join expensive coaching institutes, which provide pre-exam mock tests. In addition to the financial cost, this process consumes a huge amount of time and energy of the students during a very important stage in their academic lives.

The current state of research in terms of automatic MCQs generation, heavily skew towards English based curriculum, which creates a considerable research gap for the Bengali medium students. Moreover, the available structure of Bengali textbooks, which are frequently only accessible as scanned PDFs, further complicates the task - since it creates the necessity for the use of strong text extracting pipelines. To overcome these, we have built a system which will extract and clean semantically sound text content from the textbooks assigned by National Curriculum and Textbook Board (NCTB), construct knowledge graphs and offer a hybrid Retrieval Augmented Generation (RAG) framework to produce high quality multiple choice questions. An useful feature of our system is that it can predict answers from MCQs inputted by the students as well.

Our system provides a variety of advantages to the SSC candidates. They will be able to generate an unlimited amount of practice MCQs directly from any chapter, create custom test sets by topic and have immediate validation of answers. Moreover, teachers can also take advantage of this system to evaluate MCQ script more efficiently. In addition to its direct use as an SSC preparation tool, the system demonstrates a blueprint of a scalable, intelligent MCQ generation system with answer extraction feature, in low resourced languages like Bengali, opening the way to more effective educational technology.

1.1 Research Statement

Our research establishes automated pipelines for generating multiple choice questions for Bengali medium SSC level students and predicting the answers of their input MCQs. By automating the entire process, this paper fills the gap of affordable, high quality, curriculum-matched MCQ practice materials that the SSC candidates are so desperately in need of. With all content based upon official NCTB declared textbook materials, the system remains factual and relevant, providing students and teachers with a flexible on-demand evaluation system.

1.2 Research Objectives

The objectives of this research are given below :

1. To deploy an advanced framework which produces good-quality MCQs containing semantically appropriate distractors in Bengali language.
2. To provide a pipeline capable of predicting the answers of the MCQs that are given as inputs.

3. To create a pipeline capable of creating and traversing knowledge graphs from a given text book.

4.To provide a complete analysis of current NLP approaches applied to automatic MCQ production field.

5.To explore and implement a hybrid graph-based approach which enhances the retrieval quality of RAG based systems.

6. To analyze the transformer architecture based embedding models specially designed for Bengali language.

7.To study OCR technologies through a comparative evaluation of its performance for academic text extraction from scanned PDFs.

8. To develop metrics capable of evaluating the generated MCQs of our framework and show their effectiveness.

9. To perform a comparative analysis of three LLMs' performances in generating quality MCQs.

1.3 Report Structure

This document presents an overview of why the topic matters while addressing our research statement and objectives in the ”Introduction” chapter. The next chapter, ”Literature Review”, provides an in-depth summary of the research papers we found to be relevant with our topic, by explaining their goals, datasets, technologies and evaluation approaches. The ”Data” chapter contains complete data description and explanatory data statistics regarding the entire research project. In the ”Models” section of our ”Methodology”, we have discussed each of the models involved in our pipeline development process in great detail. The in depth details about the pipelines for MCQ generation, answer prediction and knowledge graph creation can be found in the ”Framework” section of the same chapter. In the next chapter, ”Results”, we have presented the description of our developed metrics capable of evaluating MCQs generated by our system and showed our evaluation data for both MCQ generation and answer prediction pipelines. Later in ”Result Analysis”chapter, we have analyzed our evaluation outcomes for individual stages of our pipelines. Our paper details a complete step by step workflow for our research in chapter ”Work Plan”. This work finishes with a summary of essential research principles in the last chapter followed by an evaluation of how an automatic MCQ generator system may impact the users.

Chapter 2

Literature Review

For our background research, we used search terms like “NLP based automated MCQ generator”, ”Bengali transformer based model”, “Contextual distractor generation for MCQ” and “Multilingual automatic MCQ generation” across websites of ”Institute of Electrical and Electronics Engineers”, ’Google Scholar’ and ’Semantic Scholar’ etc. Our research primarily focused on studies which were published following 2020 since we wanted to investigate contemporary technological frameworks. We had initially collected about 50 papers and then thoroughly examined abstracts and conclusion sections to perform our screening process. Twenty two papers were shortlisted through this process as they directly addressed our research issue. Our methodical selection process ensured that we selected only the newly published resources with positive research findings to guide our research path.

2.1 Related Works

Text Extraction

Ray Smith introduced the Tesseract OCR engine framework along with its development history in his presentation [31]. From 1984 to 1994, the HP engineering team designed Tesseract as an answer to the printing limitations that restricted commercial OCR solutions at that time. HP Labs Bristol researchers built the PhD engine before integrating it into future HP scanning hardware. HP postponed opening up Tesseract’s codebase although its performance exceeded that of similar systems until 2005. The testing platform consisted of binary images with marked textual regions to examine document scans at varying quality levels. Data from the UNLV Annual Test of OCR Accuracy served to evaluate the system performance through standardized tests in their benchmark role for OCR systems. The Tesseract engine performed text detection through connected component analysis before executing its two-pass recognition system with adaptive learning functionality. The engine built a line detection system that kept tabs on both straight and bent or slanted lines. The adaptive classifier implemented baseline/x height normalization features and polygonal approximations as distinctive features compared to competitor systems. The implemented technologies delivered better measurement results while boosting text format detection abilities and maintaining operational stability. Initial tests demonstrated that Tesseract outperformed commercial competitors through its accuracy metrics. The assessment used character error rates and word error rates to determine success outcomes. Tesseract achieved superior performance than current OCR solutions in the 1995 UNLV benchmark tests through substantial word error rate and character error rate reductions. Research results establish Tesseract as an OCR platform which meets current industrial standards of accuracy yet preserves its distinctive capabilities. The combination of the adaptive classifier with creative architectural principles provided Tesseract with exceptional capability for processing complex text recognition needs. The research shows Tesseract continues serving as a central development foundation for OCR technology while HMM-based character n-gram methods improve its accuracy.

Pipeline Insights

In their paper [2], Razia Marzia and Asheque Siddique presented a model that autonomously generates multiple-choice options and answers from user-provided question sources. The system addresses two phases in its goal by first creating three relevant distractors for each input question followed by generating an entire experiment set with multiple choice questions. Using their dataset, they implemented a corpus system based on question setters to generate answer options for various questions. Since their system lacks a standardized domain, it cannot process different forms of Bengali text using a uniform logic. Acquiring the required corpus proved challenging because Bengali remains a language with inadequate textual resources. Model development and system architecture followed a rule-based approach, although they omitted some elements from the first Bengali automated question-answering system, BFQA. The system accepts Bengali input queries for processing which includes question analytics followed by word stemming and ranking and topic-based answer extraction and distractors development. The second goal relies heavily on relevant methods and processes to generate both answers and multiple-choice options from the model's output. For their system the functional development team opted to use Oracle version 11g as a data storage component alongside Spring framework version 5.0.3 for middle tier development and ExtJs framework version 6.2.0.981 for building the user interface. The system performed more efficiently with domain-specific text collections than with generic text databases. The system displayed poor performance in crafting response choices for specific input query assessment although it produced effective multiple choice question sets along with swift test execution. Precision along with recall and Fscore were the evaluation metrics the team employed for model assessment. Answer extraction reached an 80% accuracy rate however distractor generation succeeded at only 53% accuracy. Study findings will serve as a base to unite DL, NN and ML systems toward performance optimization and future applications advancement. The successful extraction of answers marks progress yet developing effective distractors stands as the primary challenge to make high-quality multiple-choice questions.

Another method of working with a Bengali corpus was investigated through the scholarly work of Samina Tasnim Islam [3]. Three main objectives drove her investigation into developing an intelligent question-answering solution that recognizes precise answers from user-generated questions in Bangla language. The paper enhanced question quality with natural language processing methods while measuring performance with different metrics to conduct model comparisons with alternate systems. The decision tree model used 444 stop words and 29 suffixes during preprocessing to prepare the dataset. Decision tree classifier(C4.5) served to update the training list for the measurement unit. Two attributes appeared in her dataset including quantitative information and quantified statements. The decision tree classifier determined what position the tree root should occupy. The system used traditional NLP methods to extract answers from questions. The data preprocessing started with cleaning data followed by stemming and keyword extraction. The system used N-gram from keywords for its closest matching mechanism. Additional semantic and lexical attributes helped determine targeted answer types within specific questions. The decision tree model provides an easy solution for working with datasets directly in a system. The evaluative metric consisted of precision together with recall as well as F-score while examining both performance quality and accuracy. The Case-Based Reasoning system implemented a rule-dependent approach yielding precision rates at 36% while recall reached 66% and the F-score achieved 44%. The decision tree classifier produced higher performance than the rule-based approach with precision at 42% , recall at 84% and F-Score at 58%. Machine learning techniques improve answer accuracy to the point where they show potential advantages against standard rule-based decision systems.

The authors Musale, Shafali, Meghana, Ishwari, and Bhujbal of the research [4] studied a different approach to automate multiple-choice question creation from text input. Manual question generation required lengthy effort so the authors sought to assist teaching personnel. Their system worked by finding keywords to produce suitable questions alongside proper distractors. The article examined multiple MCQ generation strategies to show existing missing points and built a simplified process for making advanced MCQs. The researchers conducted their experiments on educational papers from textbooks alongside educational content and research publications. The system transforms text documents into MCQs along with occasional additional elements using UML and software engineering terminology. The system operates by discarding superfluous text when it creates questions. The system employed parts of speech tagging combined with syntactic parsing and ontology-based approaches. The implementation methods adapted based on the types of operations required. Parts of speech tagging enabled the researchers to discover keywords. Standard syntactic parsers employed a methodology to convert sentences into question structures. A rule-based system established the domain-fitted nature of all questions. Each method was picked because it demonstrated both competency and effectiveness. The research presented no specific success metrics yet the integrated system can be tested through question validity and user input combined with content difficulty evaluations. The assessment relied on accurate distractors for measuring success through established distractor accuracy standards. The system received evaluation according to the quality of its syntactic question production. The system achieved its target by generating high-quality MCQs yet more adjustments would improve outcome quality.

Chidinma and Ikechukwu's paper [6] introduced an NLP system that produces multiple-choice questions for the Computer-Based Testing Examination (CBTE). The researchers developed multiple choice questions with appropriate answers and correct well-structured distractors that used textbook materials for teacher support. Their research used multiple lecture materials of various sentence length. Teachers manually identified keywords in order to establish evaluation criteria for the model implementation. Information about sentence word length limits alongside sentence count averages guided the preprocessing steps for all datasets. Using Term Frequency and Inverse Document Frequency analysis, they determined term occurrence frequency within documents to identify important terms that extracted keywords from the text. With N-gram they analyzed sentence structures to produce questions that maintained proper contextual validity. The study adopted Natural Language Processing (NLP) techniques demonstrating successful implementation in contemporary scientific applications. The interface used Django as the main framework while employing a high-level Python web framework during design. Through their model they managed multiple NLP features including sentence tokenization and text summarization alongside normalization and stemming to achieve high competency in classification tasks. Noise removal needed the combination of term frequency and Inverse Document Frequency measures while N-gram performed word normalization. The model uses the weighted keyword as its focal point to replace it with a dash before populating the distractor field with randomly selected elements from extracted keywords. Research investigators performed model assessment by comparing automatically extracted keywords against teacher-supplied keywords. The research team used precision and recall as their key evaluation methods to compare extracted information with teacher-given key words and to measure the percentage of relevant lecture content retrieval. The research team unified precision results and recall values through the F-measure for a balanced evaluation approach. All subjects demonstrated high recall values showing the model retrieved most relevant terms. During the analysis precision scores showed that the model selected a few teacher-marked terms which were not relevant to the lecture content. The extraction algorithm recognized key words but its limited practical value became apparent because many identified keywords failed to match educational standards resulting in numerous incorrect matches.

In 2021 researchers Mehta et al created an MCQ generation system employing transformers in their paper [5]. They emphasized that minimizing The manual creation process for Multiple Choice Questions demands a significant reduction of time and work investments since online testing controls an increasing segment of the industry. Dominating role of online testing. The researchers believed their system would prove viable for use. A system creates both Multiple-Choice Questions and relevant distractors directly from specific text materials. The system would generate MCQs alongside suitable distractors from any text while requiring minimal expenses and labor time. This paper intended on developing system has been designed to transform basic input text of any academic field into MCQs. The research study omitted dedicated dataset specifications yet declared flexibility towards any input textual domain and discipline. One availability of testing resources known as the CNN/DailyMail dataset helped evaluate system performance. It used BERT which serves as a deep learning approach which 'google' created. The model focuses on Natural Language Processing operations as its primary objective. It was used because The transformer model demonstrates superior performance across large text archives and detailed summary generation. For summarization, the BERTSUM,a modified version of BERT, provided the capability to recognize appropriate sentence selections while conducting MCQ generation. The application used WordNet. A lexical database together with other resources. The system uses word association with hyponymic relationships to create distractors. ROUGE F1 scores evaluated the system performance to measure the summary quality. The assessed quality of text summaries from "BERTSUM" fell below the output of this approach compared to other methods. The system provided strong performance toward exam generation work. The system evaluates hypernym-hyponym relationships through oriented MCQs containing accurate distractors, with an average score accuracy of around 70.81% for the option generation task. This process demonstrates its ability to produce MCQs that have comparable quality to professional human-made questions. The goal of this work's implementation succeeded in reducing human labor requirements. The system's demonstrated capability to produce quality distractors establishes measurement standards for this type of technique. The system's ability to fulfill its objectives reflected in its delivery quality.

In the paper [9], Vatsal Raina and Mark Gales aimed to generate an automated system that produces MCQs-from English comprehension passages that includes both questions and answer options. Theoretically, they desired to replace the unreliable n-gram baseline evaluation metrics such as BLEU and ROUGE and argued for the proposal of a new framework that respects grammatical fluidity, answerability, diversity, and complexity of generated questions. The authors used the RACE++ dataset, a large English reading comprehension dataset split into RACE-M (middle school), RACE-H (high school), and RACE-C (college), hence covering both easy and hard difficulty levels. Each of these subsets was used to train and test a model, thereby presenting a wide scope of complexity for sturdier testing. They used two main models. The first one was T5 (Text-to-Text Transfer Transformer), which had automatic encoding-decoding capabilities and offered the best performances in text generation tasks. With this model, based on a given context, the question and the answers were generated. The second one was the GPT-3 used in a zero-shot setting for baseline purposes. For evaluation, an ELECTRA-based model ensemble was trained for multiple-choice (MCMRC) and question complexity classification. There were four metrics used for the evaluation. Grammatical errors (G) were almost none, noting fluency. Answerability (A) was measured by uncertainty estimates from an MCMRC ensemble; unanswerability was thus dropped from 0.8413 in unfiltered outputs to 0.6350 in filtered outputs. Complexity (C), 0 being easy and 1 being hard, was set to 0.3839 for filtered T5 outputs against 0.4402 for human-written questions. Diversity (D), being the entropy of binary question types (stand-alone and option-dependent), was 0.6629 in the filtered T5 set and 0.7750 in the human set. A total of 77.24% of generated samples had four distinct answer options. From these, a filtered subset was sampled in which all three MCMRC ensemble models agreed on option one as the correct answer, yielding 100% accuracy in agreement on that subset. The paper greatly contributes to MCQ generation. MCQs generated by the system are found to be slightly less complex than their human-designed counterparts. However, with the T5 architecture, its generation consists of valid-grammatical question grammar, answerable questions, and heterogeneous questions. The framework is found far more useful than setting up traditional metrics and, therefore, promotes a strong foundation for the evaluation of educational NLP systems.

A detailed investigation of Bangla Natural Language Processing (BNLP) emerged from the pen of Ovishake Sen and his team in 2017 [7]. Traditional Natural Language Processing (BNLP) techniques are described by the authors as they survey traditional and modern BNLP practices. ML and DL techniques. Researchers made value addition to modern scientific literature their main goal during the method evaluation process. The analysis of research publications included the examination and analysis of techniques among divergent domains within BNLP. A multi-article functionality includes simultaneous execution of sentiment analysis text summarizing and additional features including speech recognition capabilities. Electronic publication researchers analyzed seventy-five articles spanning from 1999 to 2021 with particular emphasis on articles released after 2015. The research entered its initial phase in 2015 then expanded by adding both written text content and spoken audio data to its collection methods. This research implements standard algorithms together with modern ML approaches and advanced DL models. Standard approaches used DTW and HMM and rule systems to execute processing tasks on received input information through speech or text. A research framework contained the ML and DL models which included SVM, CNN, RNN, LSTM, and GRU. The selected algorithms prove highly effective for dealing with difficult textual and spoken contextual patterns. Spoken word data shows its functionality by enabling applications for sentiment analysis as well as named entity recognition (NER). We conducted selection model evaluations primarily based on their distinct computational features. Data analysis tasks along with skilled linguistic detection accelerate these advances through large-scale data procedures The success of extracting essential information represents an essential requirement for building BNLP applications. The research exhibited continued accuracy enhancements in various BNLP operations. Our NER system achieved 0.72 F1 score maximum through CRF model adaptation with CRF. The sentiment analytical LSTM model displayed an accuracy level reaching 73.6%. Experimental results showed the Convulated Neural Network-based Bengali speech processing system reached 99.02% accuracy during numeric digit recognition. The majority of research prototypes displayed significant advancement since their first publication. Experimental performance metrics exceeded previous methodological outcomes indicating the authors fulfilled their research objectives. Both advanced classical and machine learning systems worked together to enable smooth operation in this method. This method enhanced the reliability of BNLP task assessments. The academic work sought to perform an extensive investigation of Biological Natural Language Processing tasks. The existing methods we have today exist in separate segments because of this. The reported framework enabled efficiency analysis by establishing different classification segments. A primary emphasis of BNLP development centered on the generation of new datasets and models along with new modeling approaches. The research delivers extensive theoretical foundations to support future Biological Natural Language Processing advancement.

Advancing the NLP field of the Bangla language, in 2022 [8], the authors at Bhattacharjee et al developed the BERT variant BanglaBERT through training experience on large Bengali language datasets. They developed datasets of the Bengali language which stands as a document-shortage resource. They studied Bengali language monolingualism while attempting to determine its performance relative to multilingual models. A pure Bengali language dataset-trained model exhibited what level of performance would result when compared to alternative models. Through this research, a Bangla Language Understanding Benchmark(BLUB) emerges as one major outcome. ”BLUB” represents an evaluation framework that judges model performance through different linguistic assessment tools in various NLP tasks. The datasets they extracted through their visits to more than a hundred popular Bengali sites and public repositories served as their research foundation. This large 27.5 GB of text within their dataset produced over 2 billion tokens giving them the name Bangla2B+. The ethical requirements forced them to eliminate inappropriate words and select exculsive websites. The research team excluded websites that included swearing language from their analysis. The training of any linguistic model can greatly benefit from using this collection. The dataset serves as a foundation for developing Bengali word meaning and contextual competence within any model structure and tokens. The development follows an alternative approach compared to standard Masked Language Modeling (MLM) methods. The research team applied ELECTRA instead of BERT’s typical predictive model because it supports pretraining through Replaced Token Detection. Besides, its computational efficiency, The discriminator component of ELECTRA functions automatically to detect replaced tokens while the generator predicts masked tokens. The model benefits from all sentence tokens as it receives predictions from the generator and identification from the discriminator through the(predicted masked tokens allowance feature) of the model architecture. The model addresses every token in a sentence while MLM can only process 15% of masked tokens. Besides, techniques like deduplica- The preprocessing involved deduplication followed by non-Bengali page filtration along with html and JS tags removal and batch tokenization. The dataset received preprocessing that covered several modifications. The team applied a Wordpiece vocabulary consisting of 32000 subword tokens after processing their corpus and selected a 400-character vocabulary for their model. To generate text in Romanized Bangla the developers settled on using a character vocabulary of 400 vocabularies and proceeded with vocabulary definition through the 32000-token Wordpiece corpus. The SentNob combined with BNLI together with MultiCoNER and BQA and TyDiQA supplied their experimental data. Their system executed multiple NLP processes including text classification, NER, NLI and question production roles. Research tasks were applied to evaluate their model using the BLUB for model evaluation and standing Benchmark. After applying different metrics F1 Score along with Accuracy and sample efficiency the researchers evaluated their model efficiency. BanglaBERT scored 77.78 points on BLUB to surpass all other models. BanglaBERT showed promise in many The NLP tasks show that BanglaBERT surpasses all the multilingual models by achieving remarkably high computational efficiency in these NLP tasks. The research demonstrates how models for distinct languages operate efficiently. This research demonstrates success in analyzing Bengali data while establishing opportunities for future Bengali NLP work.

In the next year, Samruddhi Deode, Janhavi Gadre, Aditi Kajale, Ananya Joshi and Raviraj Joshi [12] worked to confirm that vanilla multilingual BERT can be made into strong cross-lingual Sentence-BERT for under-resourced Indian languages without using parallel corpora. First, they used machine translations to build two benchmarks for ten East Asian languages - IndicXNLI (392 k entailment pairs) and STSb (8 k similarity pairs). Next, they included the real-world IndicNLP news-classification sets to make sure the results apply to real data. They relied on MuRIL for multilingual purposes and selected S-BERT subnetworks for each language, trained Siamese networks with many negatives, and achieved desired results in IndicXNLI; they chose MuRIL due to its pre-training on seventeen languages, and S-BERT because it is designed for sentence-level similarity, retrieval, and zero-shot tasks. Spearman embedding correlations for IndicSBERT and the ten SBERT-STS models went as high as 0.85 (Hindi 0.83, Gujarati 0.82), with each beating LaBSE by 0.07 to 0.13 points and beating its own original models by about 0.25. KNN accuracy for all eight languages in IndicNLP text classification stayed at least 0.95 and shared a high value of 0.99 for Gujarati and Marathi. The zero-shot similarity between English and Indic improved a lot, as IndicSBERT-STS scored 0.82 for en-hi, which is 0.10 better than LaBSE's 0.72. According to the authors, the goal was accomplished because using their method led to around a 45 percent increase in embedding results when measured by similarity, and about the same accuracy as before in downstream tasks; moreover, no parallel texts were employed and parameters stayed the same. Besides, the runtime worked efficiently since fine-tuning did not cause the model to grow, and training only needed 8 GB of VRAM. All this demonstrates that the method benefits not only researchers but also those working in industry with modest resources. HuggingFace was used to release the code and models developed, so the community could immediately access them again, for free.

In their paper [20], Salim and Das aim to develop a monolingual Generative Pre-trained Transformer(GPT) model, specifically tailored to the Bangla language. They addressed the limitation of the multilingual model for low-resource languages like Bangla. They mentioned that, based on recent studies language language-specific monolingual GPT models perform way better than multilingual GPT models. They used two types of datasets. One dataset was utilized for training the model, and the other dataset was used for testing the model. For testing purposes, they used a novel dataset named BanglaCLM. BanglaCLM comprises 26.24GB of Bangla text data. Approximately 50% of the data originates from the OSCAR corpus, while 24% of the data comes from Wikipedia. Additionally, 15% of the data was sourced from a popular newspaper called Prothom Alo, and 14% was derived from another newspaper named Kaler Kantho. For testing purposes, they made a separate dataset that contains 4960 sentences, which were collected from recent news articles, prothomalo.com, and bdnews24.com. Their BanglaGPT model is built on the GPT-2.0 architecture, which applies an encoder-only model from the GPT-2 model. It has multiple decoder block that contains masked self-attention block, feed-forward neural blocks, and normalization blocks. Their model has total trainable parameters of 123,239,808. They used Causal Language Modeling(CLM). In Causal Language Modeling, it predicts the next word in a sentence based on the previous tokens. For tokenization, BanglaGPT used Byte-Pair Encoding(BPE). Unicode normalization and rule-based replacements were used to standardize Bangla text. They used GPT-2 architecture because of its effectiveness in text generation tasks. To evaluate the BanglaGPT model, they used Perplexity and Loss metrics. Perplexity measures the model's ability to predict the next word in a sequence. A lower perplexity score reflects better performance of the models. BanglaGPT achieved a Perplexity score of 2.86, but Multilingual GPT achieved 6.27, and the LSTM-based Sequence to Sequence Model scored 10.512. The Loss metrics represent the model's error in predicting the next token. A lower loss score portrays a better performance of a model. BanglaGPT had a loss score of 0.45. BanglaGPT outperformed compared to these models based on Loss metrics. Their goal was successful, as BanglaGPT demonstrates better performance for Bangla text generation, evidenced by better perplexity and loss scores compared to other baseline models.

Pochiraju and colleagues with Chakilam and Betham explored a method in their paper [18]. which developed a natural-language processing system to build automated multiple-choice questions. Researchers studied natural language processing techniques to develop trainer tools that help educators create high-quality modular assessments. The researchers applied XLNet approaches to develop multiple-choice questions . Their research group evaluated the BERT methods against their implementations. The research model operated without demanding training data as well as pre-labeled examples. Using Python programming they built this implementation. From the given data the first stage involved the model processing the provided textual input by the user. The keywords were extracted from the summarized sentences. Their system operated with the main base of XLNet as text summarizing technology. The system employed XLnet to provide condensed text summaries as well as XLnet techniques for text compression tasks. They used YAKE as a keyword selector. The response system would produce the answer for the multi-step question. questions. The implementation used YAKE because training this tool does not depend upon any specific datasets. The system provides superior performance across datasets over the alternative RAKE selection framework. The methodology enabled more applicable distractors to the questions. The study team conducted performance comparisons between the XLNet model and BERT model and made findings. XLNet maintains dual functionality by uniting autoregressive features with bidirectional contextual analysis. One difference between the XLNet English model and the XLNet Large English model exists in their layer counts. The models develop two structures which contain 12 layers and a different structure with 24 layers. BERT has a token The token-based restriction of BERT serves to stop inferior output however XLNet operates without such limitations. The researchers tested the implemented model afterward while conducting a comparative assessment against BERT. The research utilized identical raw textual content in both model implementations. The model-generated questions proved superior to BERT's questions in terms of quality according to evaluation. The question structure underwent substantial modifications when using this BERT model. The model's distractor quality received enhancements according to their findings. This study presented quantitative insights into specific keyword extracters in context. Of all the extractors YAKE had better performance. The paper provides evidence which shows the data shows XLNet surpasses BERT while YAKE delivers superior key extraction results. These procedures demonstrate high utility which indicates their potential to enhance different modeling approaches.

In the same year, in the paper [14] by S Mahesh Kumar et al, the researchers implemented Word Sense Disambiguation WSD alongside batch processing and tokenization methods according to Hedge. The researchers deployed tokenization and batch processing to maximize the performance of natural language processing models. The authors utilized pre-trained language models BERT and T5 to specifically generate MCQs. Their focus was Resource consumption alongside computational complexity reduction served as the main objectives. Researchers used advanced methods involving batching together with tokenization. This paper wanted to create a framework which utilizes complex NLP optimization technologies to create a unified system for generating MCQs. Their research focused on model optimization of pre-trained systems T5 and BERT without requiring specialized datasets. dataset. The research project operated using WordNet as its main foundation but focused on a lexical English dataset. The selected models extract WordNet-based synsets, hypernyms and hyponyms to establish contextual word relationships. The discovery of contextual relationships between different words leads to better accurate distractors and questions. event MCQ and distractors. The SQuAD datasets functioned as an essential testing ground alongside other networks for model optimization. process of a T5 model. BERT functioned as their word sense disambiguation tool while the T5 served as their primary model. model for generating questions. Through advanced batching combined with tokenization-based methods, the process achieved better results. The time and computational performance together with memory consumption received improvement through these techniques. usage. The integration of WordNet with BERT used both NLTK and Transformers. The workaround to establish Google Colab with Google Drive began as Google identified a functional problem. The "T5" model received customized training through the SQuaD dataset for its applications. Additionally, the T5 model required further modification with SQuAD to improve its structural accuracy during contextual MCQ creation. Square root de- The development team applied composition techniques to boost performance levels in ways that optimized both running times and storage requirements. Process batch operations enabled the system to tokenize faster through its optimization. as well. The team found WSD increased their model's accuracy through experimental analysis. The application of square root decomposition technique enabled them to transform the batching process time complexity from exponential to linear. The team reported that their novel batching approach brought dual benefits for memory consumption and processing time efficiency. the memory usage of the tokenization process. Unfortunately, they avoided showing proof of these improvements and ample proof of scores for their statements. The authors failed to give proper priority to both of these elements in their work. The model evaluation included an analysis of both correct outcomes and contextually relevant results. They provided no information about multiple choice questions or their answers or the distractors during their evaluation procedure. Since they did Aside from providing no domain dataset the research team failed to establish benchmarks for comparison. their generated MCQ with.

Study-Buddy, an AI-assisted learning system designed to help students is described in a paper by Fernanda Martinez et al. [15]. Study-Buddy aimed to address typical educational tool drawbacks including inaccurate content and lack of personalization methods among various Al solutions that support students and teachers. Educational proficiency transforming was the key priority which aimed to serve each learner individually. Students can find articles, textbooks and instructional materials representing each academic subject and level in Study-Buddy's datasets. The system requires student feedback for continuous improvement while they enhance its functionality. The collection of multiple information resources led to the development of a knowledge graph system which linked students to connected information topics and learning materials. The educational platform Study-Buddy combined various essential technological components to enhance learning practices. Knowledge Graphs created student-to-topic-to-teacher-to-materials linkages so teachers could deliver individualized academic support to students. A natural dialogue between students and AI assistants became possible because of Large Language Models (LLMs) in the chatbot system. Students achieved better study discipline through Computational Persuasion methods that introduced fun incentives. The authors selected these technologies to build an effective teaching environment. The performance measurement of Study-Buddy tracked user engagement levels alongside student progress. The user engagement rate measured student-robot dialogue sessions while the process tracking feature enabled teachers to gauge student scholarship development alongside examining inquiries' academic quality. The paper lacks any declaration about its intended success rate. The development of Study-Buddy involves expanding its subject selection while enhancing the user interface and changing the feedback mechanism to strengthen its operational effectiveness.

Reddy, Dheeraj, and Vishal described their research in[19] dedicated to building a model that would create accurate relevant questions through Natural Language processing of user input data. The objective was to develop better examination formats suited for educational institutions from schools to colleges to coaching centers. The developers aimed to employ their model in the near future to produce exam questions through this system. The team concentrated exclusively on question generation with the T5 model. They lacked performance metrics to determine their system progress and failed to evaluate model importance through comparison against other models. Their combination of the T5 text transformer model used pre-trained data from the c4 dataset. Their initial processing phase concentrated on preparing the input user data. They utilized WordNet along with Sense2vec to create distractors from identified key sentences and keywords. The designers completed the process with a T5 text transformer model to generate their results. Text input data underwent processing with this model which operated exclusively as a summarization tool. The research team selected this model because it received previous training across multiple tasks for supervised and unsupervised applications keeping its format oriented toward text translation. Since the model received prior training the team does not require separate training of their own dataset. The system resolved complicated processing methods and administrative difficulty. Through T5 users can process various tasks by adding specific prefixes to individual activity inputs without the need for a separate T5 Transformer model. Since the system did not require a particular dataset for training the team chose to disregard all related metrics. Their work contained no indication regarding performance evaluations making it impossible to assess their results. In different language processing tasks the T5 transformer demonstrated a high overall success rate yet this achievement varied based on particular operational objectives and assessment parameters.

In 2023 Cheng Zhang proposed a published paper[23] that built an automated MCQ testing system for comprehension measurements through automatically generated questions. Producing QAPs with suitable alternative responses and adequate numbers of distractors backed up the effectiveness of MCQs during testing activities. Research focused on developing a system able to generate valid questions with proper grammar while maintaining logical sense to specific arrayed articles. The study based its model training on multiple standard datasets including SQuAD which included 100K QAPs extracted from Wikipedia articles. The analysis included the RACE exam dataset from Gaokao English testing along with the CoQA dataset for conversational pairs. For generating QAPs researchers applied two distinct strategies to their work. The Deep-learning-based approach (TP3) deployed the T5 Transformer model as its core technology component. T5 was chosen for its ability to perform multiple NLP tasks because it accepts text-to-text inputs which makes question and answer generation highly effective. MetaQA followed a Sequence-learning-based approach by producing QAPs through meta-sequence descriptions of actual sentences which incorporated concept/synonym tags. The researchers selected the MetaQA method due to its ability to build questions with proper grammar that TP3 was unable to create. Alongside each other both approaches functioned to harness deep learning capabilities and sequence learning ability for developing suitable question formats. The paper implements Positive Security tagging as well as named entity recognition and semantic role assignment and word vector representation technologies as distraction methods. All evaluated models achieved performance 17 based on automatic and manual evaluation scoring. In TP3 the training of T5-Base and T5-Large model cases proceeded using different learning rates and yielded superior results for T5-Large. Automatic assessment used BLEU alongside ROUGE and METEOR and BERT Score as evaluation metrics. The T5-Large model achieved a BLEU score of 23.83 which placed it ahead of alternative models in both ROUGE and METEOR evaluations. The SAT practice testing yielded a 97% accuracy for the MetaQA system which verifies its ability to create questions that are correct grammatically while maintaining contextual relevance. Trajectory for their generation of questions and distractors and the T5-Large approach yielded a smaller but superior performance compared to previous models. The system evidence combined BLEU and ROUGE scores and human assessment to establish that the produced inquiries were adequate and acceptable to a minimum rate of 90% per QAP. The research met its target of building an automated inquiry-generation program that created MCQs successfully. MetaQA reached success points through leveraging extensive datasets in combination with T5 model selection and MetaQA technology introduction.

A new artificial intelligence-based MCQGen system appeared in the paper by Hang et al.[21] which focused on developing Multiple Choice Questions (MCQs). The main goal of this work was to integrate LLMs, specifically GPT-4, systems for creating MCQs with a focus on individual learner adaptation needs in personalized learning practices. The framework adapts to modern learning techniques including blended learning and flipped classrooms because it offers students valuable efficient question generation features. MCQGen retrieved questioning material simultaneously from educator-originated materials and student-generated questions as part of its knowledge foundation. Student- and teacher-generated questions were processed separately in MCQGen, categorized by difficulty level (easy or hard) and guided by educational principles of variety and creativity. MCQGen's retrieval-augmented generation functionality received improvements which created more efficient operations. GPT-4 provided the text generation capabilities though question generation functions depended on retrieval-augmented generation with model sampling from external materials. Advanced prompt engineering techniques with CoT and self-refine strategies turned the LLM able to generate context-sensitive complex questions which revealed student misconceptions.GPT-4 served this application because it offered both educational content creation and textual coherence with comprehensive contextual outputs. MCQGen was evaluated using human and machine assessments across five criteria: flexibility of grammar, responsiveness, differentiation, density and significance. Both experimental evaluation results indicated strong performance from basic questions and the new framework demonstrated potential to produce more complex questions. Low-effort shorter MCQ development cycles are achievable through implementation of suggested structural operations compared to conventional methods. The achievement of strategic development objectives occurred while maximizing platform features to optimize learning outcomes. The development of candidate metrics such as diversity and complexity reveals opportunities to enhance automated question generation.

In 2024, Maity, Deroy and Sarkar introduced a paper[24] that developed an new method to generate multiple choice questions using multi-stage Prompting. They selected 4 different languages for their work which included Bengali as well. Various datasets served as their data source: 'SQuAD' for English, 'GermanQuAD' for German, 'HiQuAD' for Hindi and "BanglaRQA" for Bangla questions. They leveraged the comprehensive capabilities GPT model through paid APIs to generate plausible distractors for their generated MCQs. To determine accuracy, the team employed Bilingual Evaluation Understudy metrics, while the Longest Common Subsequence assessment measured recall performance and the team measured answer quality using cosine similarity comparisons between right and incorrect solutions. The authors conducted a comparative evaluation between Multistage (MSP) and single stage prompting (SSP) methods. Their experimental data indicated that the multi-stage promotional method with GPT showed better reliability than the SSP.

Research from Roy and Manik[25] demonstrated the creation of a domain-focused QA system which operates in Bengali through BERT-Bangla model optimization for KUET queries. The research developed a solution to address the lack of Bengali-speaking domain-specific QA systems through building a model that handles domain-relevant question processing efficiently. Researchers extracted their data from KUET's official website and Wikipedia articles focusing on KUET. The evaluation was based on a collection of 100 context segments totaling 250,000 words each. Analysis found that a majority of text blocks contained two to five matched question and answer pairs. The authors selected particular sources that featured information about KUET academic programs combined with admissions data and research facilities alongside university infrastructure. The research utilized BERT-Bangla which represents a transformer model specifically designed to work with the Bangla language. BERT functions exceptionally well for NLP applications by understanding word meaning within textual contexts. The team selected BERT-Bangla as it underwent extensive Bengali text pre-training using diverse textual resources while demonstrating success at handling Bengali language challenges. Additional modifications were made to BERT-Bangla for domain questions because they demanded expertise in precise technical language. The system tested for quality analysis using EM metrics alongside F1 metrics along with PPL metrics to deliver consistent assessment results. Thus, scientists achieved an fl score of 55.26% by measuring precise alignment. Research based findings demonstrated the classifier achieved a precision level of 74.21% at measuring correct predictions against total instances. Longitudinal testing demonstrated that the system performs well with basic scientific questions but faces difficulties when encountering ambiguous questions. The BERT-Bangla model demonstrated its capacity to address domain-specific questions yet additional research must be done to make it more effective. The authors proposed enhancements including increased datasets and simplified fine-tuning stages and external information implementation to boost query identification precision. The paper establishes a starting point for improving Bengali closed-domain QA systems yet additional work must be completed to handle complex textual queries.

Generative Model Research

The Gemini Team at Google expanded their Gemini 1.5 family of multimodal models with Gemini 1.5 Flash and Gemini 1.5 Pro during this year[22]. The project developed efficient algorithms to solve multimodal problems with massive text and audio and video content processing capabilities. The state-of-the-art models set new benchmarks in document and video question-answering systems, as well as speech recognition tools to support mobile applications and additional language services. The collection included Web documents alongside images and audio and video data and human subject preference data for additional models fine-tuning capabilities. Operational performance evaluation utilized real operational multi-linguistic documents alongside extensive text records for examination. The Gemini 1.5 Pro system used Dependency-Level (SWalcon planning) to process 10 million token contexts with sparse mixture-of-experts (MoE) transformers. The power-law context scaling framework presented its new iteration as part of computational components that combined multimodal input sequences. The system developers implemented memory capacity extensions and reasoning performance upgrades across diverse database networks. The performance peak of Gemini 1.5 Flash emerged from tensor processing units that optimized framework delivery which provided swift precise results at no cost to accuracy.Research performed on Gemini 1.5 Pro system during ”needle-in-a-haystack” synthetic operations demonstrated recall achievement exceeding 99% with text and video and audio tokens that exceeded ten million. Measures of precision outperformed basic human capabilities when analyzing text translations from English to Kalamang along with various authentic real-world scanning operations. The translation quality measurement by BLEURT used precision along with recall metrics to support outcome validation. The new multimodal QA and ASR tasks achieved 20-75% better performance when compared to existing solutions and competing models. Redefinition of Gemini 1.5 shows substantial advancements in its ability to process diverse data formats alongside industry-leading results across multiple application domains. Through Gemini 1.5, the team managed scalable issues, enabling their position as an essential breakthrough in large language model development. This research yielded important results that propel progress for practical applications and multimodal AI systems which focus on language preservation while boosting professional productivity.

Google DeepMind Gemma Team began their report [27] by indicating they had trained a family of lightweight open language models, Gemma 3, with 1B to 27B parameters, and their objective was to build on the previous Gemma series by adding multimodal vision understanding, 128K-token long context, and wider multilingual coverage, keeping within the capability of being deployed on consumer-grade hardware. They pre-trained these models on a mixed corpus of about 2T tokens (1 B model), 4T (4 B), 12T (12 B), and 14T (27 B). The data in this token budget consisted of text and image data, with monolingual and parallel multilingual sources and low-quality or unsafe data removed; image data was passed through a SigLIP encoder to generate 256 vision tokens per example. Architecturally, they kept a decoder-only Transformer backbone but inserted a single global self-attention layer every five local sliding-window layers (span=1024 tokens), rescaled RoPE positional frequencies, and embraced Grouped-Query Attention and QK-norm to restrain KV-cache memory expansion . They trained on larger teacher models in pre-training and used a refined post-training recipe, which consists of knowledge distillation, Supervised Fine-Tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF) to hone math, coding, chat, instruction following, and multilingual abilities. As a form of evaluation, they passed the instruction tuned (IT) models through human rated benchmarks as well as automated benchmarks. Gemma 3 27B IT also participated in LMSys Chatbot Arena and got an Elo rating of 1338, which is in the top 10 open models and better than its predecessor, Gemma 2 27B IT (1220) . The 27B model achieved 67.5% on MMLU-Pro, 91.8% on MATH, 74.9% on FACTS Grounding, and 89.0% on Global MMLU-Lite on zero-shot static benchmarks , and the same uplift was observed on LiveCodeBench, Bird-SQL, and GPQA. Benchmarks on vision also improved: DocVQA increased to 85.6 (val), InfoVQA to 59.4 and TextVQA to 68.6 . All in all, they achieved their goals: Gemma 3 outperformed Gemma 2, had competitive features to much larger proprietary models (e.g., Gemini 1.5 Pro), and allowed efficient vision-and-long-context processing on commodity hardware. That said, it is evident that the work is a significant step towards efficient, open-source, multimodal language modeling with the balance between scalability, performance, and responsible deployment that will serve as a milestone to guide future research and applications.

In their paper [26], the Qwen team aims to introduce Qwen3, a large language model (LLM). Qwen3 was their latest model of the Qwen series, which was developed to accelerate the performance of multiple tasks and domains. The Qwen team extended multilingual support for 119 languages in this Qwen3 series. The new dataset included 36 trillion tokens. Qwen2.5 had support for 29 languages, but Qwen3 had support for 119 languages, which is three times higher than Qwen2.5. In their pre-training dataset, they included multiple topics such as coding for code-related data and code support. To enhance the logical capabilities, reasoning tasks were added to the dataset. Many books were added to update general knowledge capabilities. They used previous series Qwen2.5-VL, Qwen2.5-Math, and Qwen2.5-Coder models during the pre-training phase for data processing and synthetic data generation. Qwen3 includes six dense models: Qwen3-0.6, 1.7B, 4B, 8B, 14B, 32 B. Qwen's tokenizer got a huge upgrade with their byte-level byte-pair encoding(BBPE) with a vocabulary size of 151,669. Those support multilingual text processing across 119 languages. Qwen3 used the ABF technique, YARN, and Dual chunk Attention for long context handling. Long chain-of-Thought(CoT) was used for finetuning during the post-training phase. GRPO was used for updating model parameters. For Qwen3's performance evaluation, thinking and non-thinking modes were used. Thinking mode is optimized for complex reasoning. On the other hand, the non-thinking mode is optimized for context-driven response. In the paper, a comparative evaluation was shown among Qwen3, DeepSeek-R1, Grok-3-Beta, and Gemini2.5-Pro. Those models were evaluated across 23 benchmarks covering different tasks and capabilities. In Thinking mode, Qwen3-235-A22B (MoE Model) outperformed DeepSeek-R1. Qwen3-235-A22B had an average success rate of 76.1%, and Qwen3-32B achieved a success rate of 80.5%. Multilingual capabilities had a success rate of 83.3%, which indicates the overall capabilities of 119 supported languages. The integrated system, which has the capability of dynamic mode switching, was deployed with an accuracy of 98.9 percent. The overall performance and versatility of Qwen3 have met their ultimate project goal.

Chapter 3

Data

3.1 Data Sources and Statistics

For our academic text data, we mainly chose three SSC-level NCTB books - “Bangla Shohopath”, “Bangla Shahitto” and “Bangladesh and Global Studies” (BGS). Our research focused on these literature heavy SSC books because of mainly two reasons: one is that generative LLMs are usually better at literature heavy contexts than science and the other is that text contents of literary books are generally easier to extract than the science related books because scientific subjects are usually full of symbols and equations. Moreover, the BGS book added value to our literary texts by bringing in humanities elements which improved the assessment criteria.

All academic books available for SSC level can be found completely free on the official NCTB website as scanned PDFs. These PDFs were the primary source of our text data. All the tables and the chapter-ending exercises or questions were removed as our primary focus was on the main chapter contents of these books.

For evaluating the answer extraction pipeline’s performance, our team has selected a total of 100 MCQs that were previously given in board exams, from publicly available question banks like [28]. We have ensured an equal distribution of board exam MCQs for all three selected subjects while creating this MCQ dataset. This dataset was also useful for developing one of our metrics, the details on which can be found in the ”Results” chapter. This dataset included-

- Basic recall required questions
- The questions that included multiple answer options with formats such as ”i, ii” or ”i, ii, iii.”
- Questions that include traditional ’fill in the blanks’ formats.

Table 3.1: Data Statistics

Subjects	Token Count	Page Count	Paragraph Count	Avg. word count in Paragraphs
”B.G.S. book”	74900	223	322	237
”Bangla Shahitto path”	118319	311	468	224
”Bangla Shoho path”	26517	85	98	272

Text Data Extraction

Since the official PDFs were scanned copies, we could not directly extract the text data. Therefore, a combination of two essential extraction approaches were utilized for acquiring data from these PDF digital scans. In the first approach, using a python library named 'Pdf2image' we have successfully transformed PDFs into image formats. After that, we have used Google's 'Tesseract 4' OCR tool on these and extracted the Bengali texts from the images. More about Tesseract-OCR in the "Models" section. Another approach was the manual extraction utilizing built in OCR functionality of "Google lens". However, details about google lens' built in OCR technology are not publicly available anywhere. Each chapter within 'Bangla Shahitto' and 'Bangla Shohopath' yielded text lengths between 2600-2900 words while BGS chapter texts measured between 2100- 2400 words.

3.2 Data Preprocessing and Formatting

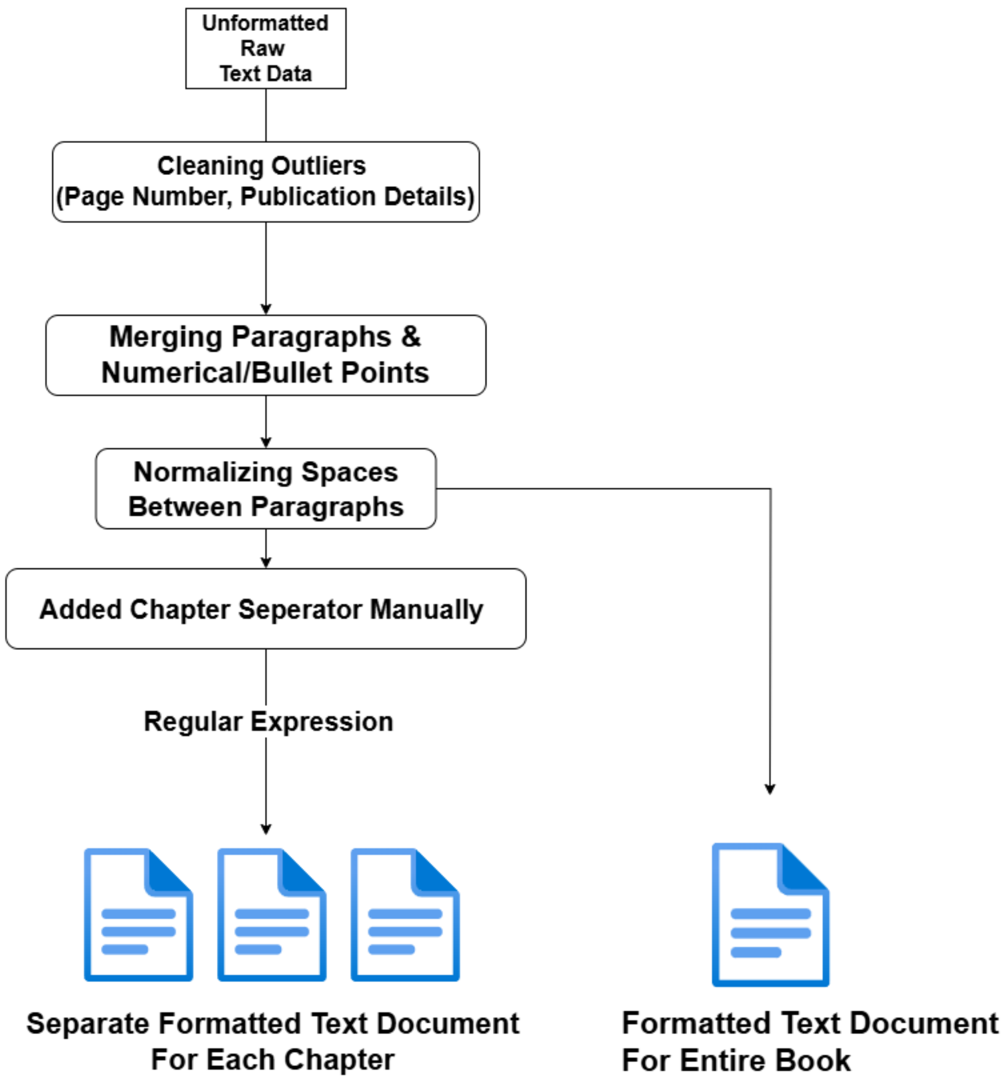


Figure 3.1: Preprocessing & Formatting Steps

The main preprocessing we had to do to our books' text data was adding the chapter separator so that we can separate each chapter's text data for the second approach of our pipeline. More about these approaches in the "Methodology" chapter. We have first separated the chapter text of a book using regular expressions through "UTF-8" encoding and created one text file for each chapter of the book. Then noises like page numbers and publication related details inside pages were cleaned by us. However, we did not remove any punctuation from our text since they could carry semantic significance which our generative LLM might need later. We had to manually format the entire book, by merging tiny paragraphs as well as the bullet/numerical points into bigger paragraphs. This was an important step in preprocessing because it prevented too much variation in the chunk size while chunking with "RecursiveCharacterTextSplitter" tool. Furthermore, we have used regular expressions to normalize the spaces between paragraphs as well.

3.3 Database

ChromaDB functioned as our primary vector database where we have stored all of our text chunk embeddings, as it provides efficient vector database management. It lets us effortlessly use our custom embedding model to vectorize the text chunks and persist in a local directory. Chroma's "Langchain Wrapper" was specifically used in our pipeline to perform the chunk vectorization as well as the retrieval of top k closest chunks. It also lets us retrieve the similarity scores along with the chunks' contents, through the use of the "similarity_search_with_relevance_scores()" function. Therefore, in our proposed methodology all the retrievals for the traditional RAG part was done with ChromaDB.

To search for contextual similarity in the vector space, ChromaDB uses HNSW (Hierarchical Navigable Small World) algorithm and provides us with options of three vector search types - Cosine Similarity, Euclidean Distance and Dot Product. The details on how it was incorporated in the framework can be found in the "Methodology" chapter of this paper. Moreover, chroma contains the functionality to perform a keyword based search in the database, which was applied for our answer extraction pipeline.

Chapter 4

Methodology

4.1 Models

4.1.1 Extraction Model

We decided to use Tesseract 4 by Google as our primary Optical Character Recognition (OCR) tool for extracting the bengali text data from PDFs. With a pre-trained model experienced with Bengali language, we do not need to perform any additional OCR model training. Our decision to use Tesseract for OCR stemmed from its open-source design together with its ability to work across multiple languages.

Tesseract 4 is a bidirectional LSTM based open-source OCR engine which is trained on more than 100 languages which includes Bengali as well [31]. It is a RNN based tool which leverages Long Short Term Memory type neurons to properly label a sequence. It was mainly trained on real world labeled text images as well as synthetic datasets of various languages. This newer version is mainly focused on line based recognition however it also supports the character pattern recognition of older versions of Tesseract, using which it generates ranked output character possibilities. Recognizing characters in the legacy engine is done by adaptive thresholding, connected component analysis, and using classifiers like polynomial classifiers. This combination of both engines improves the accuracy of extraction. We used the python wrapper available for Tesseract called "Pytesseract" for our research. For preprocessing, "Tesseract 4" uses techniques such as binarization, noise reduction, and normalization to make images in the file look clearer. The adaptive thresholding technique is used to process images in several variations of lighting. The engine splits the image into text blocks, lines, and words using connected component analysis and different heuristic-based methods. It is important to do this step for handling documents that have several columns or pictures with both text and graphics. Tesseract isolates text lines with the help of projection profiles or geometric analysis before providing the lines to the LSTM.

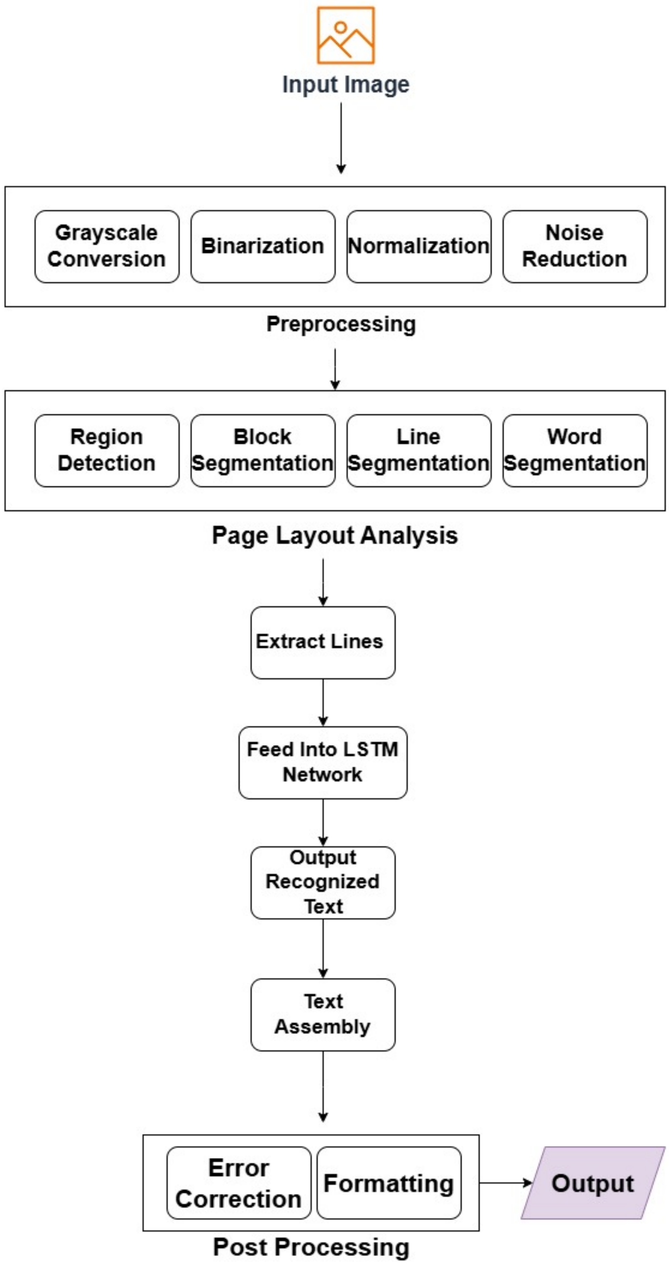


Figure 4.1: Tesseract 4 inner workings

4.1.2 Embedding Model

We have chosen "l3cube-pune/bengali-sentence-similarity-sbert" to be our primary embedding model. Our inspiration for choosing this embedding model came mainly from the outstanding effectiveness it showed when tested against "sagorsarker/bangla-bert-base" and "csebuetnlp/banglabert" [29]. This model is part of the "MahaNLP" project which mainly focused on Indian languages. It produces 768 dimensional sentence embeddings through mean pooling and taking attention mask into account **l3cube2023**.

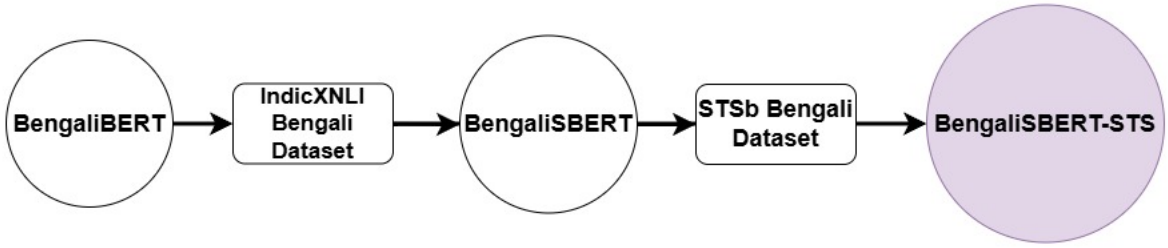


Figure 4.2: Training process of chosen embedding model

BengaliBERT, a monolingual BERT model was first trained on the IndicXNLI (Bengali) dataset created by l3cube which consists of 392702 sentence pairs. This turned the monolingual BERT model into a SBERT model capable of natural language inference such as detecting entailment, contradiction or neutral relationship among sentences. Then the trained model was further fine-tuned on the STSb dataset, which comprises sentence pairs with similarity scores annotated manually by humans using cosine similarity loss function [12].

4.1.3 Generative Models

We decided to select a total of 3 generative models to get a comparative analysis of the generation quality of our framework. The architecture of these models are as follows:

Gemini 2.0 Flash

We have chosen Gemini 2.0 Flash as one of our primary generative models mainly because Gemini's api is free of cost for 1000 requests per day and 15 requests per minute. Gemini 2.0 Flash is an innovative AI model by Google, made for smoothly handling different forms of data quickly and effectively. In Gemini 2.0 Flash, a Mixture-of-Experts (MoE) structure is applied, so each expert network takes care of particular types of data. A gating network selects the needed experts, so the system is efficient. It may handle each type of input (such as texts, images, audios, videos) with its own encoder and only combine the results in the final output. The Vertex AI documentation [30] states that Gemini 2.0 Flash launched in February 2025 as a part of the Gemini 2.0 family and will be discontinued a year later, in February 2026. Although Gemini 2.0 Flash's architecture is not explicitly outlined in public documentation, it seems to be based on the Mixture-of-Experts (MoE) architecture used in Gemini 1.5, as discussions on the evolution of the Gemini family show [10].

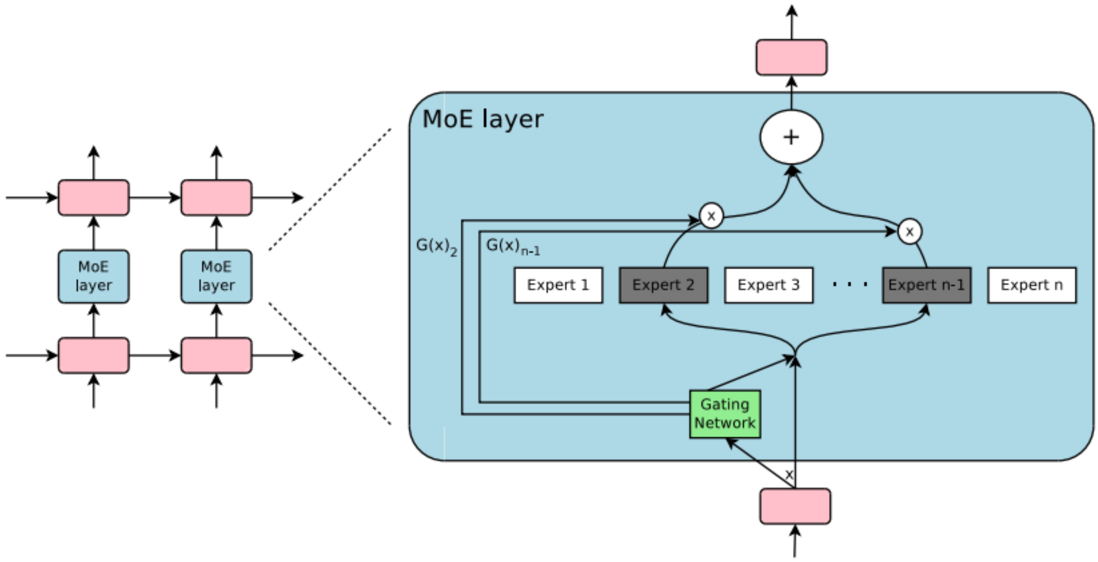


Figure 4.3: Inner Architecture of MoE

Gemini models were trained with TPUv5e and TPUv4 accelerators deployed in SuperPods containing 4096 chips, all the chips trained at the same time, along with in-memory fault tolerance. The model was trained on multilingual as well as multi-modal dataset comprising web docs, books, codes, images, audios, videos etc. The tokenizer they used was SentencePiece tokenizer [1].

Qwen 3: 14B

We have decided to use the "Qwen 3: 14B" model as another of our generative models to test our framework's generation quality across multiple LLMs. The reason behind choosing the 14 billion parameter model was mainly due to our hardware limitations. Since we used the locally installed version of this model with the help of "Ollama", we did not need the api. The architecture of the 14B model according to [26] is given below:

Model	Layers	Heads (Q/ KV)	Tie Embedding	Context/Generation Length	License
14B	48	40/8	No	128K/8K	Apache 2.0

Table 4.1: Model Architecture of Qwen 3: 14B

"Qwen 3: 14B" is a model of dense architecture which uses a "Transformer decoder", and is enhanced by "Grouped-Query Attention" (GQA) for efficient "Key-Value" (KV) cache use, SwiGLU (A variant of the Gated Linear Unit) for improved nonlinear activation, "Rotary Position Embedding" (RoPE) for position encoding, "Query-Key-Value" (QKV) Bias in attention to make it more stable, and also with "Root Mean Square Normalization (RMSNorm) and Pre-Normalization for training [26]. In the supervised fine tuning step, they introduced a diverse set of 70,000 new queries spanning several domains to enhance the logical reasoning capability of the model. These queries also included multiple-choice questions, which is one of the reasons behind our choosing this model. Another reason that leverages our research is that they used a translation model to generate low-resource language's corresponding responses converted from high resource language.

Gemma 3: 12B

"Gemma 3: 12B" model was chosen as another of our generative models for its strong multilingual capabilities, enhanced reasoning, and instruction-following abilities. As shown in the technical report [27], the Gemma 3 12B is a member of the Gemma 3 family of light models released by Google. Gemma 3 12B utilizes a decoder-only transformer, GQA instead of standard multi-head attention, an interleaving of attention layers where local attention is done 5 times more often than global attention, and RMSNorm for layer normalization. The architecture of this model according to [27] is given below:

Model	Vision Encoder	Embedding Parameters	Non-embedding Parameters
12B	417M	1,012M	10,759M

Table 4.2: Model architecture of Gemma 3: 12B

The tokenizer they used is the same as "Gemini 2.0 flash" which is a SentencePiece tokenizer. The 12B model was trained on a huge set of data that features 12 trillion tokens of texts and images. Since it includes way more examples, the model can perform well on various tasks. Gemma 3 was enhanced by using datasets that contain languages other than English inspired by the strategy of [11]. The advanced multilingual enhancement is one of the reasons why we chose this as one of our generative LLMs.

			Shards		
Model	**Type**	**#Chips**	**Data**	**Seq.**	**Replica**
12B	TPUv4	6144	16	16	24

Table 4.3: Training hardware and sharding configuration

4.1.4 Evaluation Models

BanglaGPT

"BanglaGPT" was our primary model for evaluating the perplexity of our generated MCQ questions. This is a monolingual Generative Pretrained Transformer (GPT) model built on GPT-2 architecture specifically tailored for the Bangla language [20]. This is the primary reason we chose this model for perplexity calculation of our generated MCQs. BanglaGPT has a total of 124,249,808 trainable parameters. These trainable parameters also reflect the capability of the model to capture complex patterns in training data.

Parameters	Learning Rate	Weight Decay	Batch Size	Training Steps	Epochs
124M	$5e^{-5}$	0.01	32	40772228	40

Table 4.4: Model Details of BanglaGPT

For testing and validation purposes, a novel dataset named BanglaCLM was used which comprises 26.24GB of Bangla text data from which 90% was allocated for the training set, while the remaining 10% was designated for the validation set [20]. The training of the model was implemented using a selective set of hyper-parameters to enhance its core performance. The model applied a 0.01 weight decay rate, which plays a significant part in regularizing the model and prevents it from becoming overfit to the training dataset. The Byte Pair Encoding (BPE) algorithm was used to perform tokenization, which breaks down sentences into individual characters, diacritics, and sub-words. The tokens were made of vocabulary containing 50,256 individual words or sub-words. Their perplexity score was 2.86, which is significantly lower than the multilingual Generative Pretrained Transformer model(1). Details about the integration of this model into our framework are available in the “Perplexity” section.

GPT-4 for AI Feedback

We have chosen the ChatGPT 4.0 Model because is it one of the most efficient, advanced and popular Large Language Models (LLM) available in our current world’s market. It’s world wide use and top notch performance made us use it to evaluate the effectiveness of our designed metric [17]. GPT-4 is a transformer-based model which has about 1.8 trillion parameters across 120 layers. It employs a Mixture of Experts (MoE) system, with 16 expert neural networks, where each of the neural networks comprise approximately 111 billion parameters. We have used this model to judge our distractor scoring metric’s eligibility by comparing our method’s score with GPT’s sharp reasoning and language skill [16].

4.2 Framework

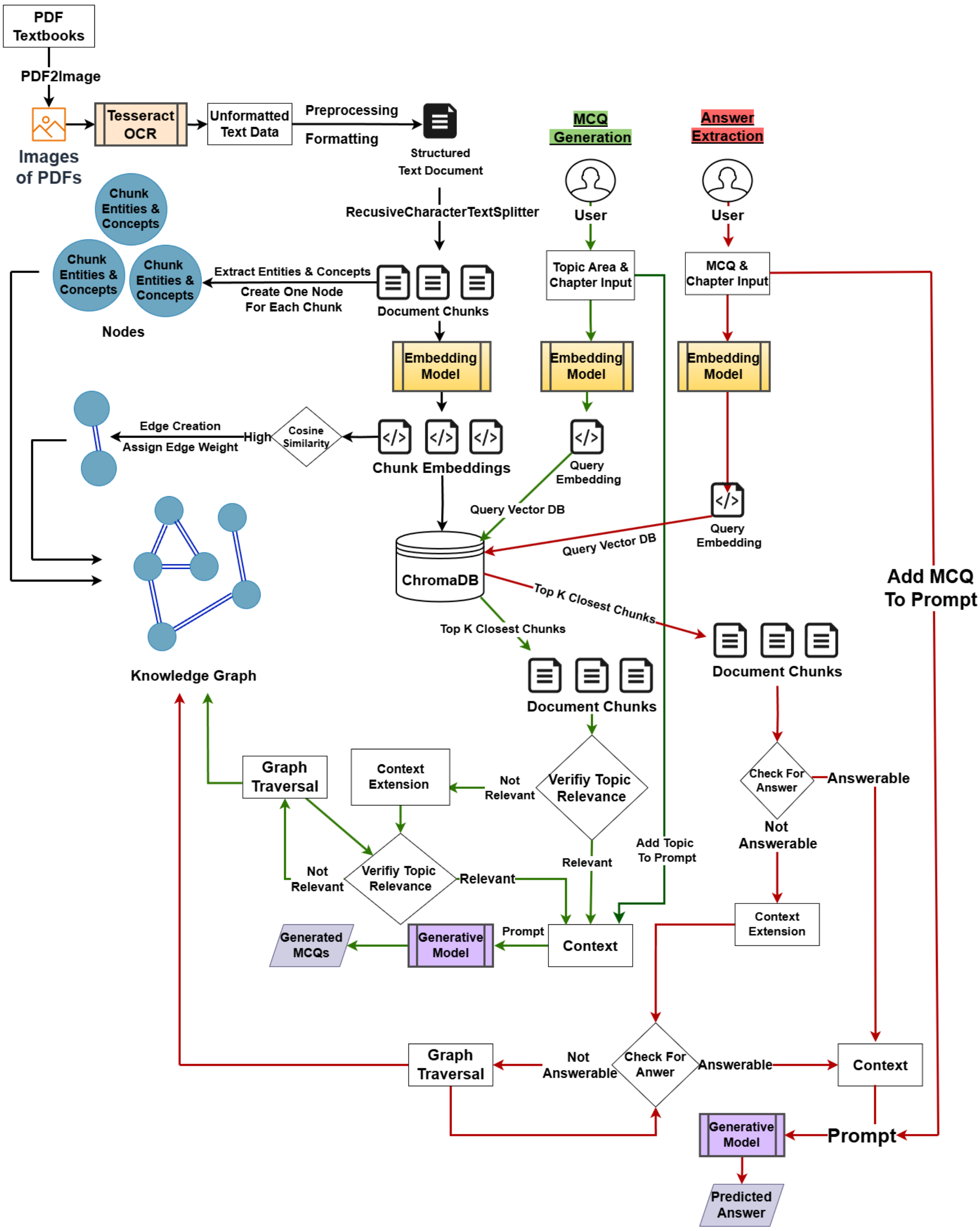


Figure 4.4: Entire framework

We have divided our entire methodology into two separate parts - MCQ generation pipeline (Green) and answer extraction pipeline (Red). For both of these, we have leveraged a modified version of hybrid GraphRAG technique inspired by [13]. We have implemented two approaches based on the retrieval domain - one where we perform retrieval from the entire book and another from each chapter of the book separately. The reason we went for two approaches is explained in detail in the "Result Analysis" chapter of this paper. The entire methodology for both pipelines (after data preprocessing) is sequentially described below in detail-

4.2.1 Document Processing

Document Splitting

We have used Langchain's "RecursiveCharacterTextSplitter" utility to split the entire book's texts into smaller chunks. Through its unique splitting technique, RecursiveCharacterTextSplitter keeps natural breaks between sentences or paragraphs intact to generate self-contained text sections. We had to manually adjust the chunk size through a process of experimentation. For example, a sentence,'দুঃখের মধ্যে দিয়েই জীবনের অর্থ খুঁজে নিতে হয়। ত্যাগের মধ্যেই জীবনের প্রকৃত স্বার্থকতা নিহিত। তাই আত্ম-উৎসর্গই মূলমন্ত্র।', the splitter treats the sequence of Unicode encoding as a continuous 100-character segment. In our test case we have decided on a chunk size of 1500 characters and a chunk overlap size of 150 characters, however, these parameters can be tuned depending on the document.

Document Vectorization

After splitting the document, we have vectorized these chunks using our chosen sentence embedding model, "l3cube-pune/ bengali-sentence-similarity-sbert" through "ChromaDB's Langchain Wrapper" and persisted them in a local directory to reuse later for creating a knowledge graph of the entire book. This entire procedure is mostly common for both of our pipelines.
However, in the second approach, we have split each chapter's texts separately and stored their embeddings into their own chapter-separated persistent directories. Thus we have kept one persistent directory per chapter to reuse later and create separate knowledge graphs for each chapter.

4.2.2 Knowledge Graph Creation

Nodes Creation

We have extracted all the named entities and important concepts using our primary generative LLM, "Gemini 2.0 Flash", from each of the chunks that we got from splitting in the first step of "Document Vectorization". If gemini's output was not properly formatted or contained any error, we have utilized a locally installed Qwen3: 14B as our fallback model. Through prompt tuning, we have extracted the LLMs' outputs in our desired JSON format and created a combined set of named entities and concepts for each chunk. After that we have used "networkx" python library to create a node representing each chunk where the node will contain mainly two attributes - the combined set of entities-concepts and the text content of that chunk.

Edge Creation

For creating an edge between two nodes representing two chunks, the cosine similarity between their chunk embeddings need to be higher than a threshold. In this case we decided on the threshold being 0.6 but this is a tunable parameter for our knowledge graph creation framework. We assigned the edge weight by using the following mathematically intuitive formula:

$$EDGE\ WEIGHT = \alpha * SIMILARITY + \beta * \frac{NUMBER\ OF\ COMMON\ CONCEPTS}{MAX\ POSSIBLE\ NUMBER\ OF\ COMMON\ CONCEPTS}$$

Here alpha and beta are two more tunable parameters, where alpha denotes the percentage effect of cosine similarity and beta denotes the normalized percentage effect of common concepts. The motivation behind this mathematical intuition came from the paper [13]. This entire process of checking for cosine similarity threshold and calculating edge weight was done for all possible combinations of unique chunk pairs of the entire book. For creating a graph for only one chapter in the second approach, we have performed the same procedure for all possible combinations of unique chunk pairs only from that chapter. We have used the previously created chunk-separated persistent directories for finding the text chunks of only that chapter.

As an example for proper visualization, the Knowledge Graph created in the first approach for the entire “Bangla Shohopath” book, containing only two chapters, is given below:

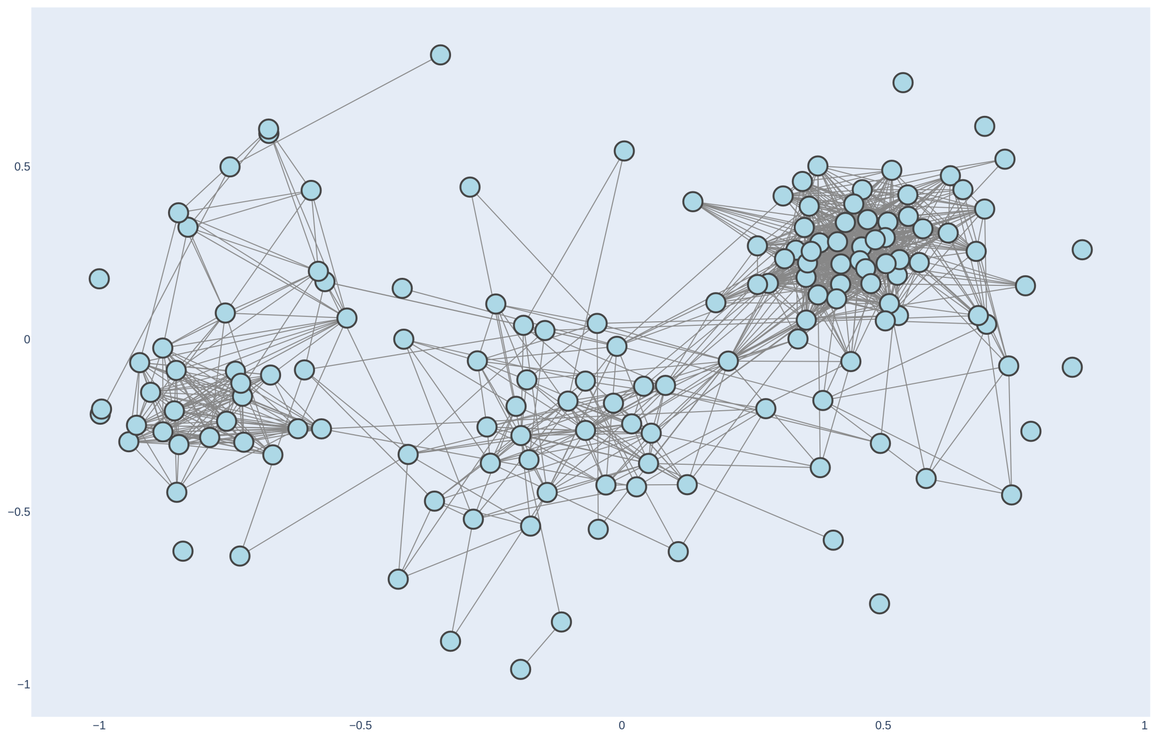


Figure 4.5: Knowledge Graph of ”Bangla Shohopath”

Below are the knowledge graphs for the two chapters of the same book created in the second approach:
First chapter:

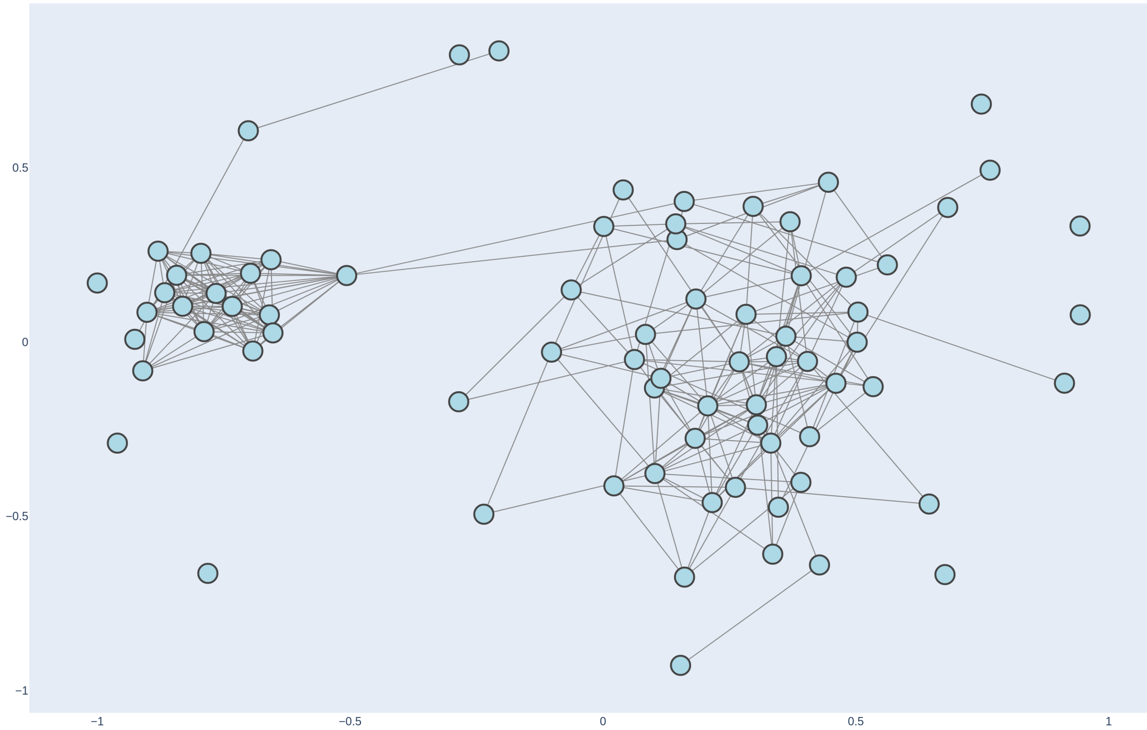


Figure 4.6: Knowledge Graph of the first chapter of ”Bangla Shohopath”

Second chapter:

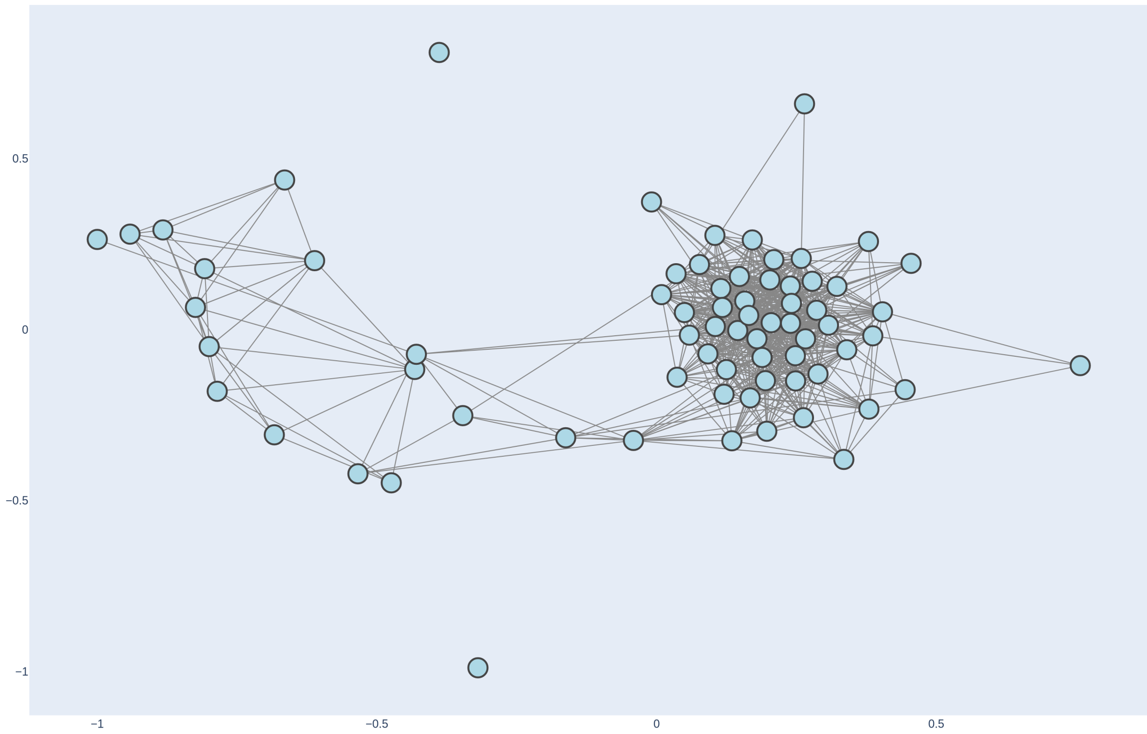


Figure 4.7: Knowledge Graph of the second chapter of ”Bangla Shohopath”

4.2.3 MCQ Generation

We have integrated 3 significant parts in our hybrid GraphRAG framework: a traditional RAG, an extension of context, and finally a graph traversal. At the beginning of our MCQ generation pipeline, the user inputs a topic area text from where they want the MCQs to be made.

After that, first the traditional RAG method will vectorize the topic area text by our embedding model. Then it will query the vector database by using ChormaDB as a retriever object and retrieve top K chunks from the vector space. In our test case we considered this K to be 5. Now for each of these retrieved chunks, we have utilized our primary generative model's semantic understanding to verify if that chunk is relevant to the input topic area. If it is relevant to the topic, it is appended to the context list for LLM but if not, we apply context extension.

For the context extension method, we traverse the database to extract the sequentially previous and next text chunk of our normal RAG-retrieved chunk, join in that sequential order and check again for the topic verification. There are mainly two reasons behind extracting serially previous and next text chunks to extend the context: one is that "RecursiveCharacterTextSplitter", sometimes splits a big paragraph discussing about one concept/topic into multiple smaller paragraphs and another is that oftentimes, textbooks tend to discuss a topic/concept in multiple paragraphs. If we still can not verify the topic relevance in the extended context of that chunk, we start our traversal of the knowledge graph considering our retrieved chunk's representative node as the starting node.

The traversal method is explained elaborately in the "Graph Traversal" section. During the traversal if we got to a node, the content of which verifies topic relevance, we add it to the context list. Finally, after conducting the same process for all the normal RAG-retrieved chunks, we join the elements of the context list and send it to our generative LLM. We have applied zero shot prompting in our case. We have also kept a parameter for our debugging and understanding that lets us perform only one of the 3 retrieval parts (traditional RAG, context extension, graph traversal) for any particular topic.

The only difference in the second approach is that we have to take an additional input parameter from the user which specifies chapter name. According to the name of the chapter, we find the mapped knowledge graph and use the persistent directory for that chapter to implement the traditional RAG, extended context and graph traversal in the same way described above for the first approach.

4.2.4 Answer Extraction

For the answer extraction we combined a total of 4 types of retrievals to ensure that our generative LLM finds the relevant context to predict the correct answer. These are keyword retrieval, traditional RAG, context extension, and graph traversal. We have also utilized our primary generative model's contextual understanding and designed a "Check Answer" function that checks if an MCQ can be answered from a given context chunk.

After the user inputs the MCQ question and its options in a list, first we perform a keyword based search in the database of chunks and retrieve the first keyword matched chunk for each word of the question as well as for each option. Then we join all these retrieved contexts and check for answer with the joined context and the MCQ. If "Check Answer" returns false, we apply the traditional RAG exactly like our MCQ generation pipeline and then for each retrieved chunk check for the answer. If the MCQ can be answered, we simply return the correct answer but if not, we go for the context extension and check for the answer again. If it still fails to extract the answer of the MCQ, we finally start the graph traversal and check for the answer after each node's expansion of context in the traversal path. The graph traversal technique for answer extraction is described broadly in the next section. In the chapter separated approach, we perform the same operation but only with that particular chapter's persistent directory and knowledge graph mapped to the inputted chapter name.

4.2.5 Graph Traversal

Since our knowledge graphs are mostly dense, we have implemented a BFS like knowledge graph traversal in both of our pipelines. Though the traversal is mostly similar in both pipelines, there exist differences in how we are exploring a node in the traversal path. The graph traversal is performed for each chunk retrieved from the normal RAG. It will start by first finding the associated node for that chunk and considering it as our starting node. Then starting from that node we will keep visiting top N unvisited neighbors for M levels deep. Here N and M are two tunable parameters where N denotes how many unvisited neighbors we want to explore at each level and M denotes how many levels we want the traversal to continue. We have considered N and M both to be 3 in our test case for less time complexity. The top N neighbors are selected based on the edge weights of my current node and the neighboring nodes.

While visiting a node in the MCQ generation pipeline, we will verify topic relevance of that node with a topic verifier function. If a node's content gets topic verified, the traversal for that retrieved chunk will stop and the next chunk's traversal will start in the same way.

However, while exploring a node in the answer extraction pipeline, we keep an expanded context list and add current node's content to that list. After visiting the node, we check if the input MCQ is answerable with the expanded context till now with our designed "Check Answer" function. If at any point we find the MCQ to be answerable with the expanded context, we return the correct answer option and stop the entire traversal immediately. However, if the MCQ is not answerable yet, we keep adding content of the node currently being explored to the expanded context list until we reach level M and thus finish traversal for one chunk. If traversal for the first chunk does not produce the answer of the MCQ, we keep doing the same traversal operation on all the retrieved chunks in the normal RAG.

Chapter 5

Results

This chapter of our research focuses entirely on evaluating each phase of our framework.

5.1 Performance of Extraction

Google's Tesseract 4 Engine generally achieves excellent accuracy in reading PDF images but produces errors during specific cases. It generates incorrect results containing noise effects especially when processing pages with images in between sentences or paragraphs. We needed to use manual extraction with Google Lens for all instances where Tesseract-OCR failed. By uniting Tesseract-OCR with manual google lens based data extraction, we achieved higher accuracy in identifying correct text from PDF images and eliminated substantial dataset errors. The performance of google lens' OCR was almost 100% so we decided not show the performance here.

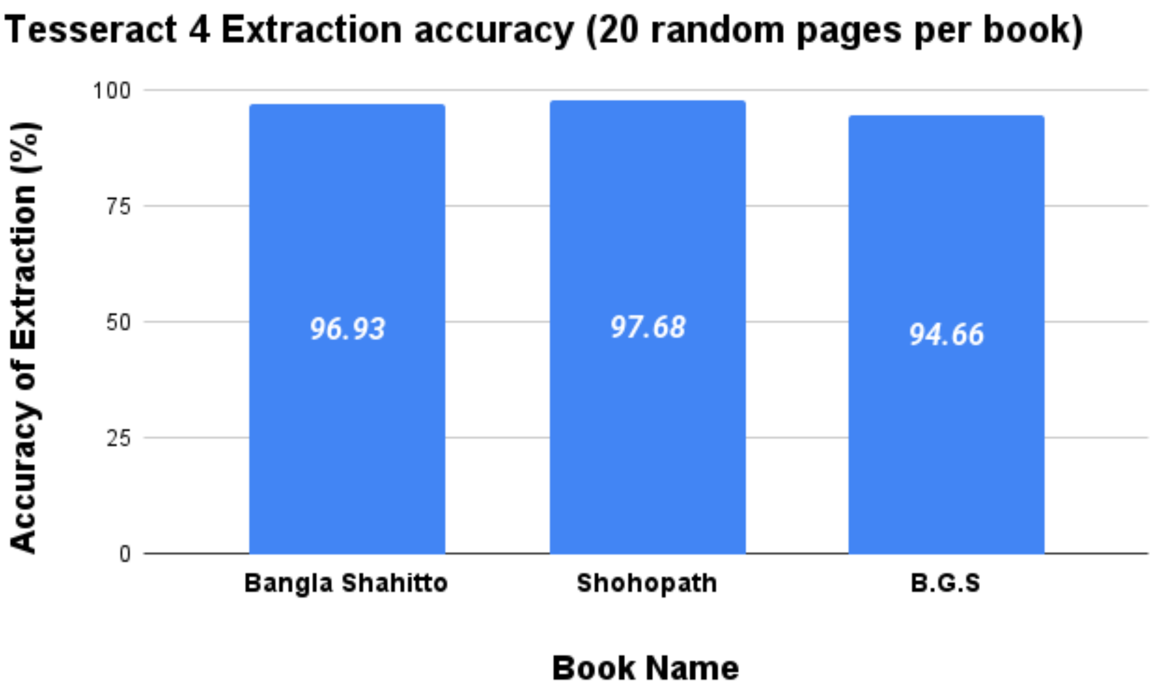


Figure 5.1: Tesseract's Performance

5.2 Performance of Question Generation

Our inability to measure the quality of multiple-choice questions through basic metrics, led us to develop our own assessment system for evaluating our framework's generation quality. We have chosen in total of 4 metrics to evaluate our MCQ generation quality:- "Distractor quality", "Contextual relevance", "Perplexity" and "Diversity".

5.2.1 Distractor Quality

The most important metric to evaluate an MCQ is how good its distractors are. Therefore, we have designed the below framework to score our generated MCQs based on their option qualities:

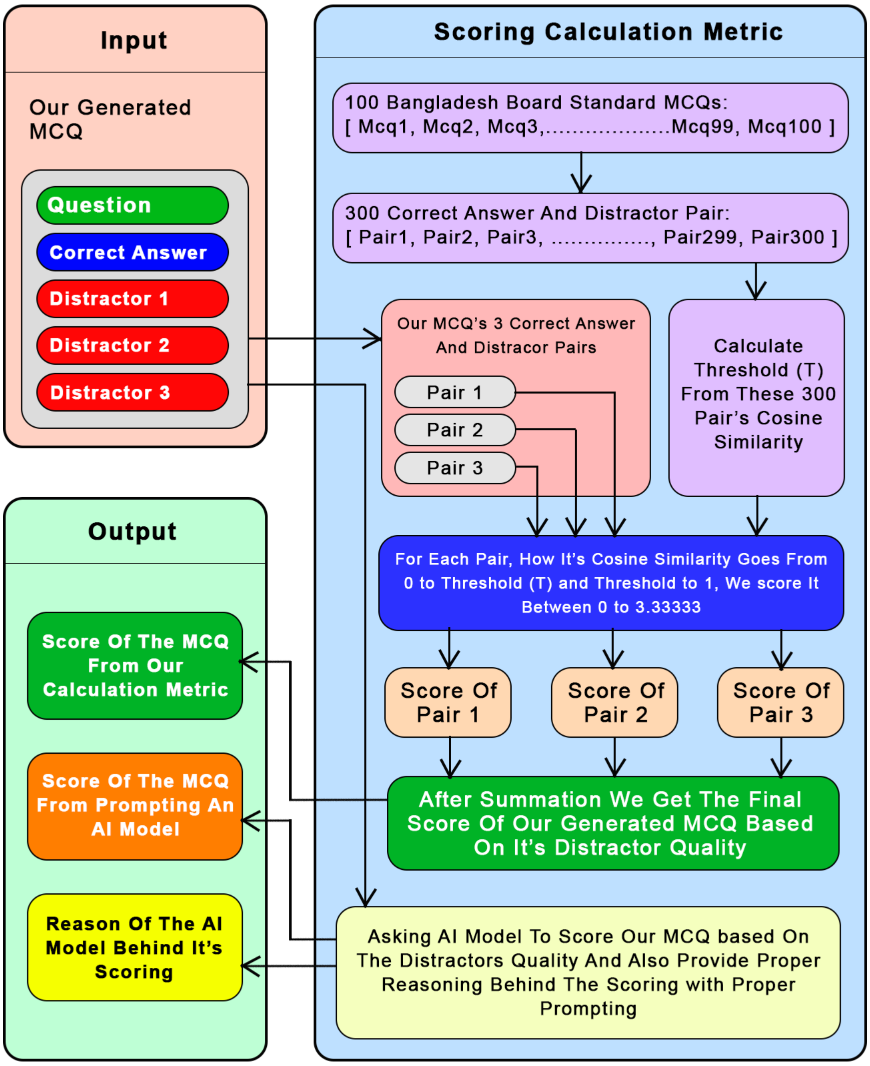


Figure 5.2: Framework for Distractor Quality

Equations

Each MCQ has one correct answer and 3 distractors, which result in 3 pairs of options : correctAnswer-distractor1, correctAnswer-distractor2 and correctAnswer-distractor3. Therefore, we wanted to score each pair separately (out of 3.33) and then consider the sum of 3 pairs as our metric's final score (out of 10). For scoring each pair, we have depended on their cosine similarity score. If the score is less than the gold standard threshold, we have scored the pair linearly with the equation of a line that passes through (0, 0) and (Threshold, 3.33). If the pair similarity is higher than the threshold, we scored it with a linear equation passing through (Threshold, 3.33) and (1, 0).

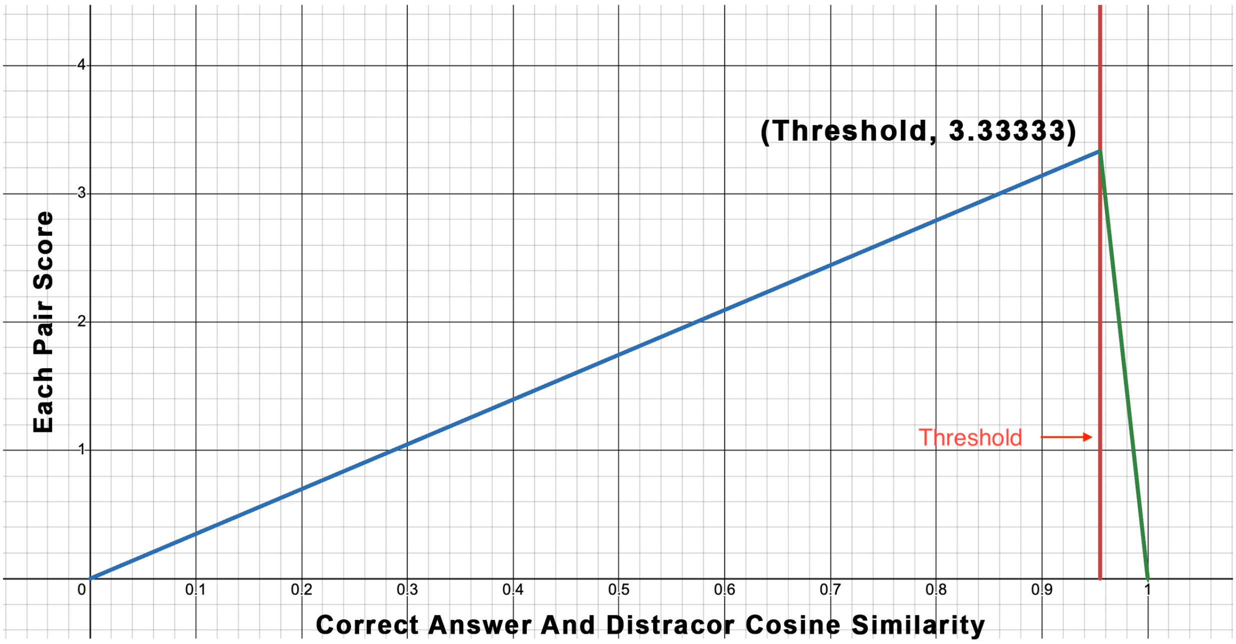


Figure 5.3: Graph of Pair Scoring Equation

The main intuitions here were - if the pair similarity decreases from the gold standard, the distractor's relevance to the correct answer also decreases, which should eventually decrease the pair score. This behaviour is characterized by the first linear equation (Blue line). However, if the pair similarity tends to 1, the distractor is semantically too close to the correct answer, which should also penalize the score significantly. This phenomenon can also be observed for our second equation (Green line) where the pair score becomes 0 as the similarity reaches 1.
Therefore the main mathematical equations that we have applied for evaluating each pair of our correctAns-distractor is:

$$\text{Pair Score} = \begin{cases} \text{Max Pair Score} \cdot \dfrac{\text{Pair Similarity}}{\text{Threshold}} & ; \text{if Pair Similarity} \leq \text{Threshold} \\ -\dfrac{\text{Max Pair Score}}{1 - \text{Threshold}} \cdot \text{Pair Similarity} + \dfrac{\text{Max Pair Score}}{1 - \text{Threshold}} & ; \text{if Pair Similarity} > \text{Threshold} \end{cases}$$

Threshold Calculation

Our goal was to take the 100 board exams' MCQ questions, from our created answer prediction's evaluation dataset and extract a gold standard similarity threshold from their (correctAnswer-distractor) pairs. We have used "shihab17/bangla-sentence-transformer" to vectorize all 4 options of the 100 MCQs into embeddings and calculated the cosine similarities between the 300 pairs generated by them.

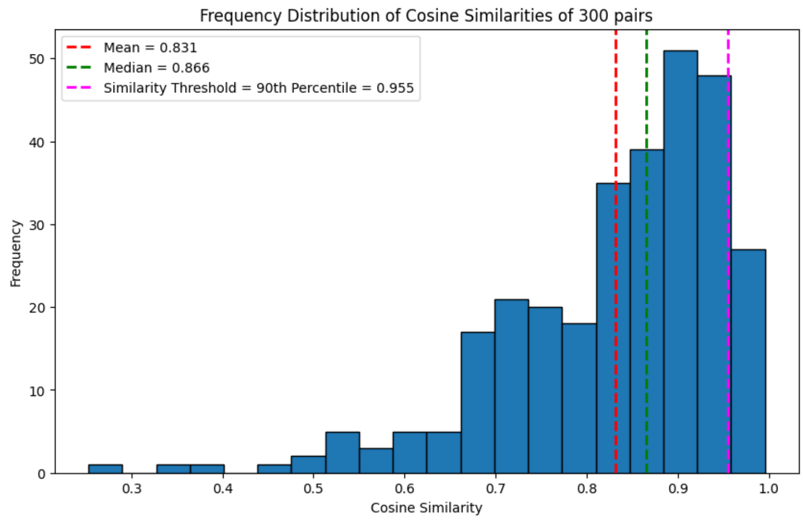


Figure 5.4: Frequency Distribution of Cosine Similarities of 300 pairs

We decided on using the 90th percentile value as our standard similarity threshold, primarily due to two observations -
Firstly, from the frequency distribution graph of the similarity scores, we can observe that our graph is left skewed. Therefore, taking the mean would be less efficient as it will be significantly skewed by the outliers. Secondly, if the similarity score crosses this threshold, we will be penalizing our score for each pair strictly, as the slope of our second equation is steeper than the first. Therefore, choosing the 90th percentile as our threshold ensures that most MCQs stay under this strict threshold and thus get scored less strictly (linearly) through the first equation.
After determining each of the three pair's score using the derived pair scoring piece-wise function, we have summed these up to get the final Distractor quality score.

$$Distractor\ Quality = \sum_{i=1}^{3} Pair\ Score_i$$

We also wanted to compare our MCQs score with the feedback of one of the popular LLMs of the current market. Therefore, we have used one shot prompting with GPT-4.0 to score the same MCQs alongside. We have designed the prompt with detailed example and rubrics similar to what we followed in our manual calculation method.

5.2.2 Contextual Relevance

We have designed this metric to measure how relevant our LLM generated MCQs are to the contexts retrieved by our framework. The higher the score, the higher relevance our question has with the context and vice versa. We are calling this evaluation metric as contextual relevance.

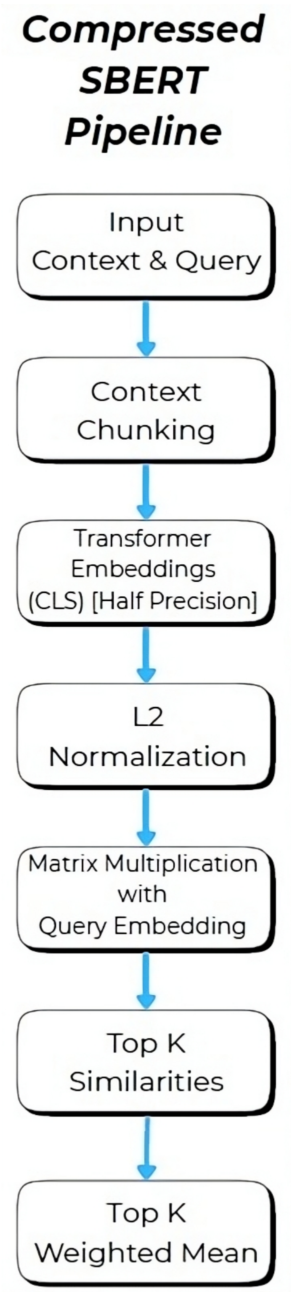


Figure 5.5: Framework for Contextual Relevance

If we put it mathematically, given the MCQ question q and the corresponding context $C = \{c_1, c_2, c_3, ..., c_n\}$ where c_i being the context chunks, we had to first find similarity score s_i for each chunk from which we would calculate the total relevance of an MCQ with the framework's retrieved context. We have utilized our chosen model "l3cube-pune/bengali-sentence-similarity-sbert" to pass both the MCQ question and the context, independently through the same shared encoder. It splits the context into multiple chunks and converts both the chunks and the question into high dimensional vectors (768-dim) allowing us to run mathematical computations when necessary. We pool only the [CLS] token representation from the final hidden layer. [CLS] is trained as a summary token especially for classification tasks. In practice it captures the global semantics of the chunks. After that, we have performed L2 normalization on our embeddings making them comparable based upon directions instead of magnitude.

The BERT encoders consume significant computational power and memory, that is why we have used a compression technique to avoid these issues. Our compression strategy was to convert FP32 weights to FP16 (16 bit floating point) , maintaining floating point semantics (exponents, fractions) but at half precision. This is common in mixed precision training and inference with frameworks like PyTorch in our case.

Equations

The equation of everything mentioned previously would turn out to be:

Conceptual Equation: $$\phi(x) = L2\_Normalize(Embedding_{Compress}\ (x)))$$

For our given MCQ question q and each context chunk c_i, we find the similarity score s_i like this:

$$s_i = cos(\phi(q),\, \phi(c_i))$$

After all of this is done, we have sorted the similarity scores of each chunk s_i in descending order. Finally to calculate the contextual relevance score of the MCQ, we have calculated the weighted mean of the top 3 chunks.

$$\omega_1 = 0.60,\ \omega_2 = 0.15,\ \omega_3 = 0.067,\ W = \sum_{j=1}^{3} \omega_j = 0.817$$

$$Contextual\ Relevance\ Score = \frac{1}{W} \sum_{j=1}^{3} \omega_j s_j$$

Weight Calculation

For determining the weights, we have randomly selected a total of 60 questions (20 from each model) generated with our framework and manually checked if the questions were actually created from the highest scored chunks given by our relevance model or not. If not, we have checked the second highest scored chunks and so on. If a question is generated from a particular chunk we have marked that as a "Hit" for that chunk.

	Highest Scored Chunk	**2nd Highest Scored Chunk**	**3rd Highest Scored Chunk**
Hits	36	9	4
Percentage	60%	15%	6.67%

Table 5.1: Overall Hits from the chosen 60 MCQs

As the overall "Hits" percentage went up to a satisfying 81.67% for the top 3 highest ranked chunks – we have decided to make a trade off and not look further beyond the third highest chunk. From this observational data, we have decided to take the weights according to the percentage "Hit" contribution of the chunks. Therefore, our three consecutive weights for the first, second and third highest scored chunks are 0.6, 0.15 and 0.067.

Model	Chunk 1 Hits	Chunk 1 %	Chunk 2 Hits	Chunk 2 %	Chunk 3 Hits	Chunk 3 %
Gemini 2.0 Flash	14	70%	3	15%	2	10%
Qwen 3: 14B	11	55%	3	15%	2	10%
Gemma 3: 12B	11	55%	3	15%	0	0%

Table 5.2: Model-wise Hits Comparison

5.2.3 Perplexity

This metric indicates how natural or fluent our generated questions are according to our chosen monolingual BanglaGPT model. Lower perplexity means the question aligns well with the patterns of the BanglaGPT model's training domain. Perplexity score can also be used to detect anomalies, as higher score highlights that the generated questions are contextually odd or contains grammatical errors. It played a significant role in distinguishing the high quality and low quality MCQ questions generated by our pipeline. More on the effectiveness of perplexity in the "Perplexity Result Analysis" section.

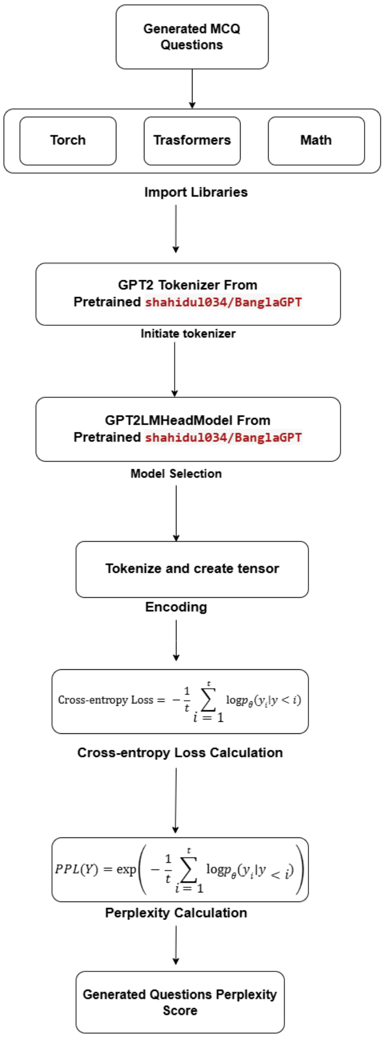


Figure 5.6: Framework for Perplexity

To begin, we have eliminated the distractors and inserted only the question text as input to the model. After that, the GPT2Tokenizer splits the question into tokens using sub-word level tokenization, maps each token to a numerical ID from the tokenizer's vocabulary and converts the tokens into PyTorch tensors. Subsequently, it executes the shahidul034/BanglaGPT model on the input tensor to compute the cross-entropy loss for language modeling. By taking the exponent of the cross-entropy loss value, we get the final perplexity score of the generated questions.

Equations

The primary formula to calculate perplexity of a language model was used for finding out the perplexity score of our sentences.

$$Perplexity(Y) = exp\left(-\frac{1}{t}\sum_{i=1}^{t} log\{\, p_\theta(y_i|y_{<i})\}\right)$$

Here, $p_\theta(y_i|y_{<i})$ is the probability that the model assigns to the i-th token given all the previous tokens. The log probability $log(p_\theta(y_i|y_{<i}))$, handles the multiplication of probabilities. The negative mean of all the log probabilities determines the cross entropy loss. The exponential function (exp) converts the score to a positive value and gives the actual perplexity score. For instance, if a model predicts the subsequent word in a sentence with a high probability (0.9) based on the preceding words, the perplexity score will be approximately 1.112, which is exceptionally low. Conversely, if a model predicts the subsequent word in a sentence with a low probability (0.1) based on the preceding words, the perplexity score will be approximately 10, which is significantly high.

5.2.4 Diversity

We have designed a diversity metric using Shannon Entropy [9] which evaluates the variety of question types per set. To implement this, we have prompted one of our chosen generative models (Gemini-2.0-flash) to classify each MCQ into one of the 10 predefined question categories- What, Who, When, Where, Why, How, Which, Yes/No, Whose, How much/many. The number of categories could be extended, but in our test case we chose it to be 10. From the LLM's output, we have extracted the total count of questions for each type to calculate the entropy. The range of the entropy lies between 0 and log2(N) where N is the total number of categories. In our case, as we have 10 categories, the maximum value of entropy is 3.321. To turn the maximum range into a better number, we have normalized the entropy such that our diversity score range becomes: 0 to 5.

Equations

Entropy quantifies the uncertainty, or randomness most often associated with the distributions of categories, calculated as:

$$H = -\sum_{i=1}^{n} P_i\, log_2(P_i) \qquad\qquad P_i = \frac{Question\ Count\ in\ Category\ i}{Total\ Number\ of\ Questions}$$

Here, P_i is total proportion of questions in category i and n is the number of categories. After extracting the question count in each category from the generative model, we have applied the formula and summed the contribution of each category to get the final entropy.

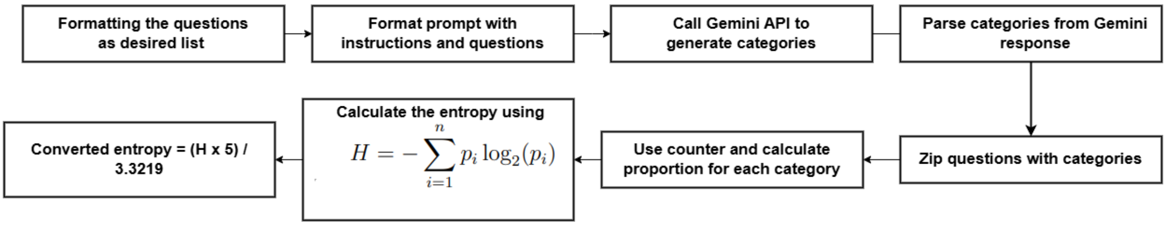


Figure 5.7: Framework for Diversity

5.3 Evaluation Data

Our system produces a set of 10 MCQs per given topic area. For evaluating the generated MCQs, we have used 3 topics to generate 3 sets per generative model, for each of the 3 books which resulted in a total of 270 multiple-choice questions.

Subject	Model	Topic	Distractor Quality	AI Feedback	Contextual Relevance Score	Perplexity Score	Diversity
ShohoPath	Gemini	1	8.28	7.49	62.67%	24.82	2.95
		2	8.08	8.51	43.41%	29.77	2.37
		3	8.34	7.72	28.09%	48.37	3.68
	Qwen 3	1	8.06	8.07	60.70%	10.75	2.54
		2	7.32	8.14	48.25%	10.12	3.20
		3	7.75	8.15	23.29%	13.05	3.68
	Gemma 3	1	7.93	7.83	50.91%	10.06	2.54
		2	7.85	7.92	47.12%	22.18	3.27
		3	7.64	8.23	27.57%	30.22	3.20
Bangla Shahitto	Gemini	4	7.51	8.20	45.46%	30.57	2.05
		5	8.51	8.19	48.38%	12.28	1.02
		6	7.56	8.10	35.75%	10.65	3.50
	Qwen 3	4	7.66	8.08	46.45%	6.77	1.09
		5	7.78	8.03	49.79%	10.15	2.54
		6	7.84	8.29	37.21%	5.68	3.68
	Gemma 3	4	7.11	8.30	40.90%	26.96	2.24
		5	5.46	8.16	45.89%	14.04	3.50
		6	8.14	8.35	37.05%	10.47	3.99
BGS	Gemini	7	8.34	8.15	53.68%	9.29	3.39
		8	6.53	8.39	60.94%	10.21	2.78
		9	6.94	8.21	58.94%	7.76	3.27
	Qwen 3	7	7.39	8.35	45.21%	4.60	2.66
		8	7.74	8.55	66.98%	3.69	1.95
		9	7.68	8.39	53.88%	4.61	3.20
	Gemma 3	7	7.33	8.51	53.56%	9.09	2.78
		8	7.44	8.17	62.18%	8.49	3.39
		9	7.00	8.12	59.61%	7.17	3.99

Table 5.3: Each metric's average score per set for 270 generated MCQs. (For detailed data, **click here**)

We are presenting the average score per set for each of our metrics in the above table, but the detailed evaluation data can be found in the "Appendices". Here topic number was assigned based on each topic text and its retrieved context per subject. For example, the "Shohopath" book's topic no. 1, 2 and 3 for Gemini, Qwen 3 and Gemma 3 denotes that the same retrieved contexts for those topics were given to all three generative models to create MCQs.

5.4 Performance of Answer Prediction

Unlike MCQ generation, answer prediction performance can be calculated by a simple metric like accuracy as the MCQ dataset for testing was balanced and errors were equally costly for each option. Details about how we created this dataset can be found in the "Data source" section. Evaluation data for Answer prediction accuracy per subject is given below:

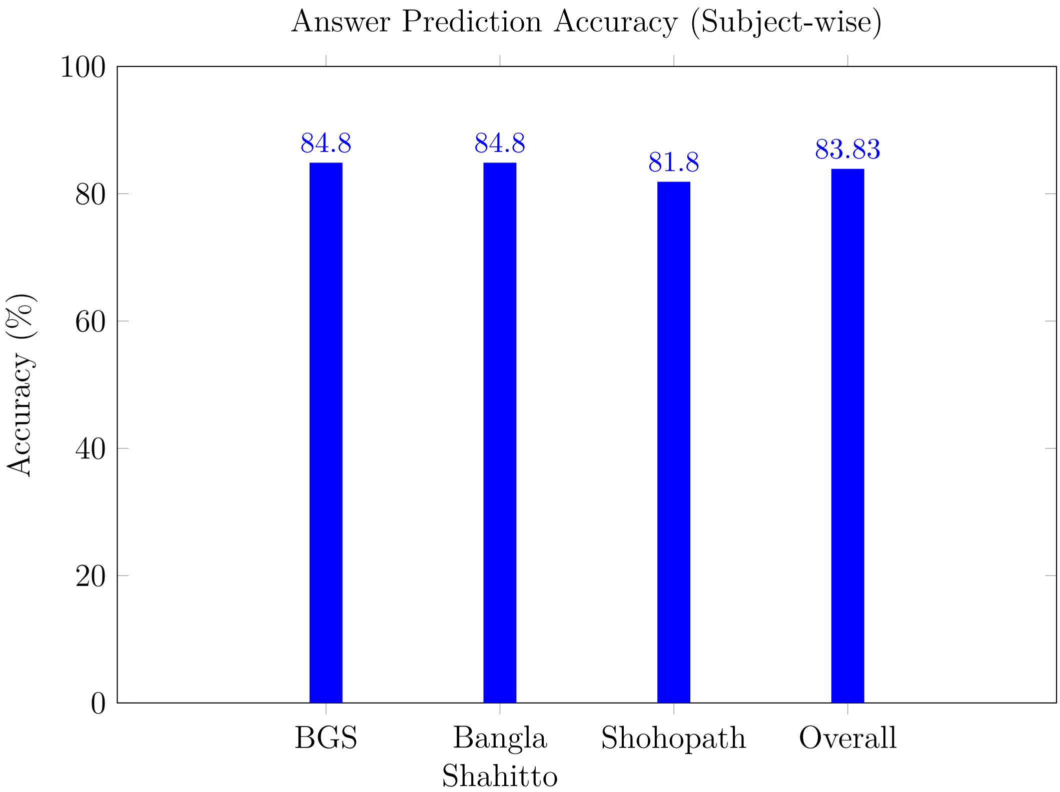


Figure 5.8: Subject-wise and Overall Accuracy of Answer Prediction

This accuracy evaluation was done specifically on the testing dataset of 100 board exam-given MCQs, which was described in the "Data" chapter. Since our framework predicts the answer through generative LLMs, the format was not dependable enough for an automated evaluation process. Therefore, we had to manually check in the question bank if the answer matched with what the LLM predicted. From the subject-wise comparison, we can see that there is not much of a difference in performance among our chosen subjects. The overall accuracy of the answer prediction was 83.83% which is till now state of the art for academic MCQs for Bengali medium curriculum.

Chapter 6

Result Analysis

6.1 Distractor Quality Result Analysis

Effectiveness of the metric

In our evaluation phase, we have observed that our metric ”Distractor quality” effectively distinguishes between good MCQs and bad MCQs.

Empirical examples:
”বহিপীর' নাটকটি প্রথম কোন সালে প্রকাশিত হয় ?”,
”correctAns”: ”১৯৬০”,
”distractor1”: ”১৯৫৫”,
”distractor2”: ”১৯৬৫”,
”distractor3”: ”১৯৭০”,
Our method scored 9.75, as the distractors’ reasonable years makes the MCQ more engaging and creates the effect of plausibility.

”শান্তি কমিটির চেয়ারম্যানের বাড়ি পুড়িয়ে দেওয়ার সময় বুধার পরিচয় গোপন রাখার কারণ কী ছিল ?”,
”correctAns”: ”বাপ-মা হারানো কিশোর হওয়ায় কেউ তাকে সন্দেহ করেনি”,
”distractor1”: ”সে ছিল প্রভাবশালী পরিবারের সন্তান”,
”distractor2”: ”সে ছিল খুবই ধনী”,
”distractor3”: ”সে ছিল খুবই দুর্বল”,
Our method scored it 3.53. Here we can observe clearly that all the distractors are sentences that carry different meanings, all of which have less relevance to the correct answer.

”কাকতাড়ুয়া উপন্যাসটির মূল প্রতিপক্ষ কে ?”,
”correctAns”: ”"পাকিস্তানি সৈন্য"”,
”distractor1”:”"পাকিস্তানি সৈন্য"”,
”distractor2”: ”"পাকিস্তানি সৈন্য"”,
”distractor3”: ”"পাকিস্তানি সৈন্য"”,
Our method gave this a score of 0 as all the options are the same.

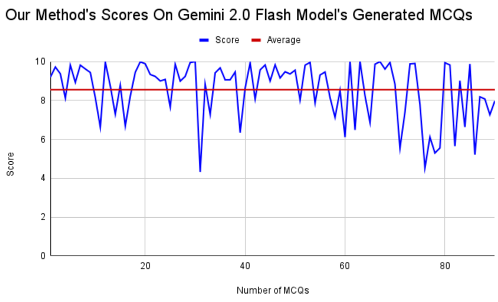
From our evaluation data we have noticed that the average scores of our method and GPT on numerous set of MCQs were very close. Our score’s closeness to one of the popular LLMs on market clearly depicts the effectiveness of our mathematical evaluation method. For further empirical evidence, we have chosen 10 comparatively high scored MCQs and 10 low scored MCQs to show the comparison in details below:

Comparatively Lower Scored Mcqs	Our Score	Avg.	Gpt Score	Avg.
{"question": "নাটকটির লেখক সৈয়দ ওয়ালীউল্লাহর পিতা ছিলেন কী?", "correctAns": "ম্যাজিস্ট্রেট ", "distractor1": "পুলিশ কর্মকর্তা", "distractor2": "শিক্ষক", "distractor3": "কৃষক"}	5.96799		6.60001	
{"question": "উপন্যাসটির মূল প্রতিপক্ষ কে?", "correctAns": "পাকিস্তানি সৈন্য", "distractor1": "বুধা", "distractor2": "শাহাবুদ্দিন ", "distractor3": "গ্রামবাসী "}	5.24218		2.90001	
{"question": "সৈয়দ ওয়ালীউল্লাহর পিতার পেশা কী ছিল?", "correctAns": "ম্যাজিস্ট্রেট ", "distractor1": "শিক্ষক", "distractor2": "সাংবাদিক ", "distractor3": "ব্যবসায়ী "}	5.90827		6.10001	
{"question": "বুধার চরিত্রের প্রধান বৈশিষ্ট্য কী?", "correctAns": "অসীম সাহস ও মানবিক গুণাবলী ", "distractor1": "ভয়প্রবণতা ও দুর্বলতা", "distractor2": "laziness and inactivity", "distractor3": "পড়াশোনায় মনোযোগ না দেওয়া"}	3.52835		4.80001	
{"question": "বুধার চাচি বাড়ি ছেড়ে আসার পর তার মধ্যে কী ধরনের অনুভূতি জাগে?", "correctAns": "মুক্তির বোধ এবং স্বাধীনতার আকাঙ্ক্ষা", "distractor1": "দুঃখ এবং হতাশা", "distractor2": "অপরাধবোধ ", "distractor3": "বিরক্তি "}	5.53982		6.90001	
{"question": "শান্তি কমিটির চেয়ারম্যানের বাড়ি পুড়িয়ে দেওয়ার সময় বুধার পরিচয় গোপন রাখার কারণ কী ছিল?", "correctAns": "বাপ-মা হারানো কিশোর হওয়ায় কেউ তাকে সন্দেহ করেনি", "distractor1": "সে ছিল প্রভাবশালী পরিবারের সন্তান", "distractor2": "সে ছিল খুবই ধনী", "distractor3": "সে ছিল খুবই দুর্বল"}	5.60914	5.66104	5.60001	5.220009
{"question": "মমতাদি ছেলেটির পরিবারের কাছে কী প্রত্যাশা করতেন?", "correctAns": "আদর, সম্মান ও সহমর্মিতা ", "distractor1": "শুধুমাত্র আর্থিক সাহায্য ", "distractor2": "কঠোর পরিশ্রমের সুযোগ", "distractor3": "অতিরিক্ত কাজের চাপ"}	5.63103		4.50001	
{"question": "মমতাদি প্রথমে মাইনে সম্পর্কে কী বলেছিলেন ?", "correctAns": "তিনি জানতেন না তার মাইনে কত হওয়া উচিত", "distractor1": "তিনি অনেক বেশি মাইনে দাবি করেছিলেন ", "distractor2": "তিনি মাইনে নিয়ে কোনো কথা বলতে চাননি", "distractor3": "তিনি সবচেয়ে কম মাইনেতে কাজ করতে রাজি ছিলেন"}	5.1706		6.60001	
{"question": ""উপেক্ষিত শক্তির উদ্বোধন" প্রবন্ধে কবি কাদের বুকে করিয়া লইতে আহ্বান জানাইয়াছেন ?", "correctAns": "পতিত, চণ্ডাল ও ছোটলোক ভাইদের", "distractor1": "ধনী ও ক্ষমতাবানদের ", "distractor2": "বিদেশি শাসকদের", "distractor3": "উচ্চশিক্ষিত সমাজকে"}	7.82038		4.60001	
{"question": ""আজ সৃষ্টি-সুখের উল্লাসে" কবিতায় ‘পিনাক-পাণির শূল’ কী প্রতীকতৎব্য ?", "correctAns": "বিস্ফোরণ ও ধ্বংস", "distractor1": "বৃষ্টির ধারা", "distractor2": "আলোর কিরণ", "distractor3": "প্রকৃতির শান্ত রূপ"}	6.19264		4.20001	
Comparatively Higher Scored Mcqs	**Our Score**	**Avg.**	**Gpt Score**	**Avg.**
{"question": ""বহিপীর" নাটকটি প্রথম কোন সালে প্রকাশিত হয়?", "correctAns": "১৯৬০", "distractor1": "১৯৫৫", "distractor2": "১৯৬৫", "distractor3": "১৯৭০"}	7.12999		9.20001	
{"question": "নাটকটি কোন সমাজের কুসংস্কার দেখায়?", "correctAns": "মুসলমান সমাজ", "distractor1": "খৃষ্টান সমাজ", "distractor2": "বৌদ্ধ সমাজ", "distractor3": "হিন্দু সমাজ"}	9.79424		7.20001	
{"question": "বুধার প্রাথমিক জীবন কী ছিল?", "correctAns": "গ্রামের অবহেলিত কিশোর", "distractor1": "মাধ্যমিক শিক্ষিত ছাত্র", "distractor2": "মুক্তিযোদ্ধা ", "distractor3": "মিলিটারি কর্মকর্তা"}	6.58883		7.10001	
{ "question": "উপন্যাসটির কাহিনী কোথায় ঘটে?", "correctAns": "গ্রামীণ এলাকায়", "distractor1": "ঢাকা শহরে", "distractor2": "কলকাতায়", "distractor3": "চট্টগ্রামে"}	7.59977		7.50001	
{ "question": "উপন্যাসটি কার জীবনভাবনা প্রকাশ করে?", "correctAns": "সেলিনা হোসেন", "distractor1": "বুধা", "distractor2": "শাহাবুদ্দিন ", "distractor3": "পাকিস্তানি সৈন্য"}	9.7957		7.30001	
{ "question": "পাঁচালী উপন্যাসটি কার লেখা?", "correctAns": "তারাশঙ্কর বন্দ্যোপাধ্যায় ", "distractor1": "মানিক বন্দ্যোপাধ্যায় ", "distractor2": "বিভূতিভূষণ বন্দ্যোপাধ্যায় ", "distractor3": "শীর্ষেন্দু মুখোপাধ্যায় "}	9.82652	7.72516	8.40001	7.97001
{ "question": "কোন লেখকের উপন্যাসে পাওয়া যায় গ্রামীণ বাস্তবতা আর ঔপনিবেশিক শাসনের দ্বারা জর্জরিত মানুষের জীবনের ছবি?", "correctAns": "মানিক বন্দ্যোপাধ্যায় ", "distractor1": "তারাশঙ্কর বন্দ্যোপাধ্যায় ", "distractor2": "বিভূতিভূষণ বন্দ্যোপাধ্যায় ", "distractor3": "জগদীশ গুপ্ত"}	7.43301		8.60001	
{ "question": "কোন উপন্যাসটি মানিক বন্দ্যোপাধ্যায়ের লেখা?", "correctAns": "পদ্মানদীর মাঝি", "distractor1": "পাঁচালী", "distractor2": "অপরাজিতা ", "distractor3": "আরণ্যক"}	7.87338		8.40001	
{ "question": "তারাশঙ্কর বন্দ্যোপাধ্যায়ের উপন্যাসে কোন মানুষদের জীবন উল্লেখ করা হয়েছে?", "correctAns": "বৈষ্ণবী, কবিতাল, যাত্রাশিল্পীদের ", "distractor1": "শিল্পমালিক আর শ্রমিকদের", "distractor2": "ঐতিহাসিক চরিত্রদের", "distractor3": "সামাজিক কর্মীদের"}	5.96799		8.10001	
{ "question": "১৯৪৭ সালের পূর্বে বাংলা নাট্যচর্চার প্রধান কেন্দ্র কোথায় ছিল?", "correctAns": "কলকাতা", "distractor1": "ঢাকা", "distractor2": "লাহোর", "distractor3": "করাচি"}	5.24218		7.90001	

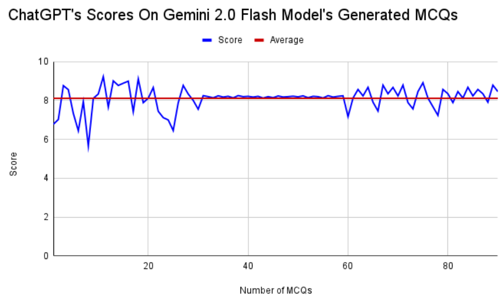
Figure 6.1: Detailed Comparison between AI and Our Metric scores for High and Low Scored MCQs

Model Comparison

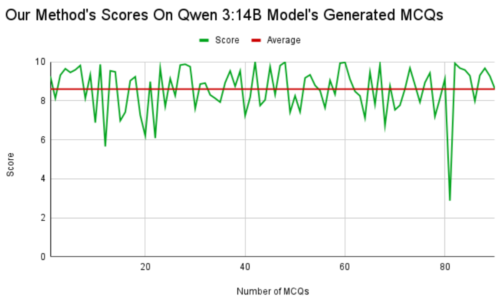
Our framework generated MCQs with mostly good quality distractors for all generative models. The comparison between their performances for both calculated method and "AI feedback" is given below:



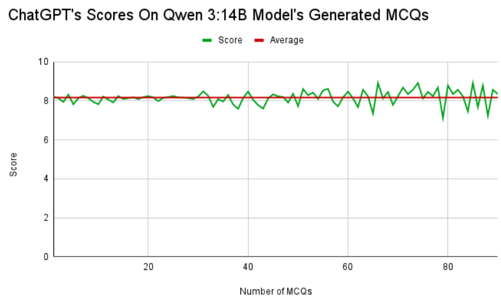
(a) Gemini MCQs (Our metric)



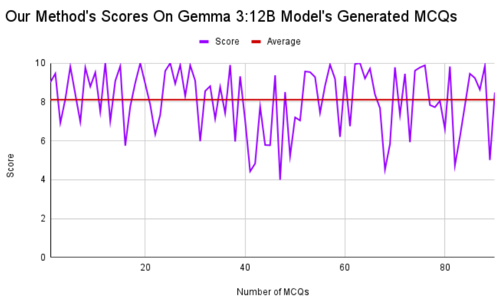
(b) Gemini MCQs (GPT-4)



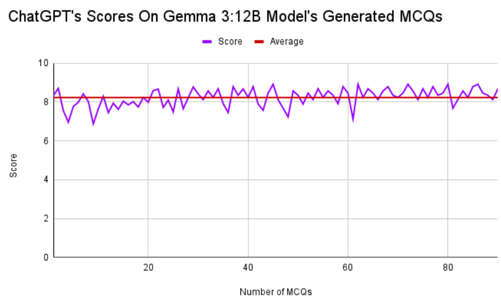
(c) Qwen3 MCQs (Our metric)



(d) Qwen3 MCQs (GPT-4)



(e) Gemma3 MCQs (Our metric)



(f) Gemma3 MCQs (GPT-4)

Figure 6.2: Comparison of MCQ scores for Gemini, Qwen3, and Gemma3 models using Our metric vs GPT-4.

From the above graphs we can see that our metric's average score on 90 Gemini 2.0 Flash MCQs, is 8.5564 whereas GPT-4's average score is 8.1125. On Qwen 3 : 14B generated MCQs, our metric scored 8.5989 whereas GPT's average score is 8.1687. Our metric's average score on MCQs generated by Gemma 3 : 12B is 8.1125 while ChatGPT' average score is 8.2290. Form here we can hypothesize that Gemma 3 : 12B model generated slightly better quality of MCQs than Qwen 3 and even Gemini 2.0 Flash.

6.2 Contextual Relevance Result Analysis

Our contextual relevance scoring metric has calculated the relevance percentage for every context chunk corresponding to that particular question.

Effectiveness of the metric

To evaluate the effectiveness of our metric, we wanted to see if it can distinguish between correct context and irrelevant context for an MCQ. For an empirical example, we have taken 10 randomly picked MCQs generated by us and shuffled their contexts to see if our metric can differentiate the effect.

MCQ (বাংলা)	**Correct Context Score(%)**	**Shuffled Context Score(%)**
'বহিপীর' নাটকটি প্রথম কত সালে প্রকাশিত হয় ?	67.64	39.44
বাংলাদেশের আধুনিক নাটকের ধারার প্রধান নাট্যকারদের মধ্যে কে অন্তর্ভুক্ত ?	67.17	38.75
'বহিপীর' নাটকের কেন্দ্রীয় চরিত্রটির নামের তাৎপর্য কী ?	63.49	43.89
লেখিকার মতে, গৃহিণীদের রন্ধন করতে বলা হলে তাদের কী হওয়া উচিত ?	52.63	15.98
মমতাদির বরের চাকরি হলে ছেলেটি তাকে কী খাওয়ানোর প্রস্তাব দিয়েছিল ?	61.70	21.21
মমতাদির গৃহে প্রথম আগমনকালে স্কুল পড়ুয়া ছেলেটির অনুভূতি কেমন ছিল ?	60.50	33.54
১৯৬৯ সালের গণঅভ্যুত্থানে কোন ছাত্রনেতা শেখ মুজিবুর রহমানকে 'বঙ্গবন্ধু' উপাধিতে ভূষিত করার প্রস্তাব উত্থাপন করেন ?	68.07	19.78
জাতিসংঘের নিরাপত্তা পরিষদে কতটি স্থায়ী সদস্য রাষ্ট্র রয়েছে ?	73.03	10.07
বিশ্বে জাতি সংঘ গঠনের মূল উদ্দেশ্য কী ছিল ?	74.04	13.21
তথ্য অধিকার আইনের মাধ্যমে কোন ধরণের মানুষের উন্নয়ন নিশ্চিত করা সম্ভব ?	73.48	17.53

Table 6.1: Contextual Relevance of MCQs with Correct and Shuffled Context

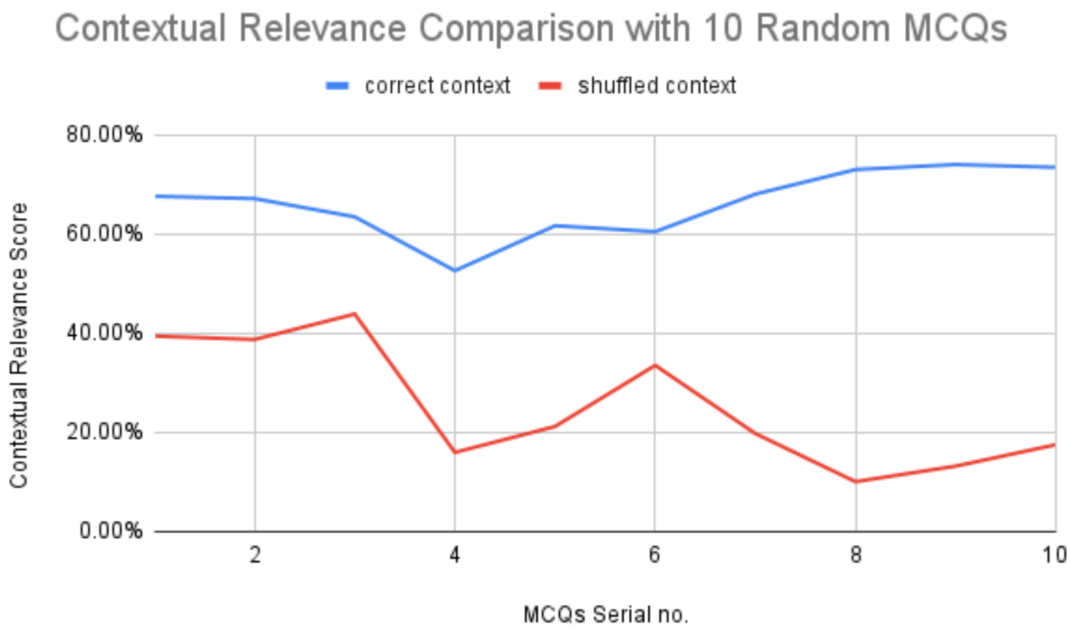
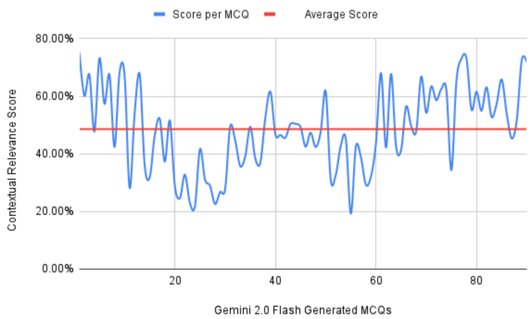


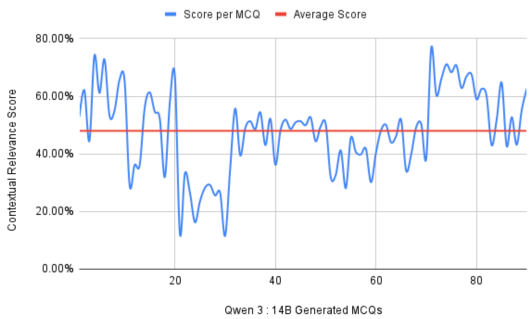
Figure 6.3: Contextual Relevance Comparison with 10 Random MCQs with Correct vs Shuffled Context

In the chart above, the blue line represents the contextual relevance scores for the 10 randomly selected MCQs given the correct context i.e. based on which they are generated by our LLMs and the red line depicts the score when the context is not their corresponding one. It is clearly visible that when the context is shuffled for a particular question the relevance score drastically decreases as the red one stays well below the blue line across every MCQ. This wide gap between the two lines makes our relevance scoring model successful as this is what we were looking for.

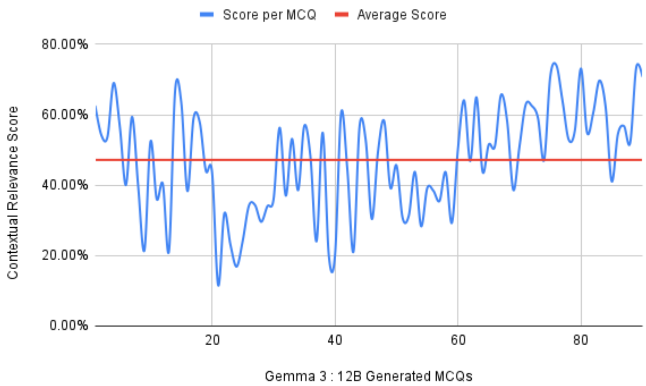
Model Comparison



(a) Gemini Scores



(b) Qwen Scores



(c) Gemma Scores

Figure 6.4: Contextual Relevance Score comparison of Gemini, Qwen, and Gemma models.

In the above graph plots, we can see the comparison of the contextual relevance of the questions generated by all 3 LLMs - 90 questions each. We have calculated that Gemini 2.0 Flash, Qwen 3:14B and Gemma 3:12B gave the average weighted means of 48.59%, 47.97% and 47.2% respectively. Even though while calculating the weights manually we have already seen that more than 80% of the questions are generated from those three chunks. We have detected that the average is a bit on the lower side because, the chunks where the questions were actually generated from, sometimes contains other irrelevant sentences/topics. These outliers drag the individual similarity score of a chunk down. Besides, the weighted mean penalizes the score more when the third ranked chunk gets a lower score than the other chunks.

6.3 Perplexity Score Analysis

Effectiveness of the metric

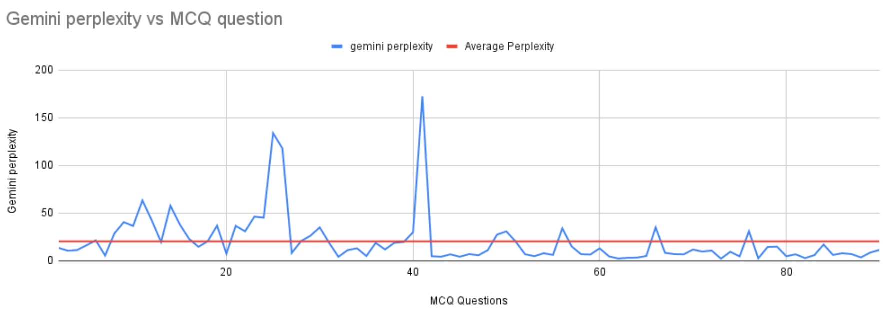
From the generated MCQ questions of our pipeline, we have randomly chosen 10 questions for empirical analysis. To clarify the difference, we have manually edited those 10 questions by shuffling the words in random order and incorrectly changing the spellings of random number of words. The perplexity scores for these bad MCQ questions increased noticeably.

Generated Questions	BanglaGPT Perplexity Score
"উপন্যাসের অন্যতম গুরুত্বপূর্ণ উপাদান কোনটি?",	2.43
"তারাশঙ্কর বন্দ্যোপাধ্যায়ের একটি উল্লেখযোগ্য উপন্যাসের নাম কী?",	5.57
"মানিক বন্দ্যোপাধ্যায়ের উপন্যাসের প্রধান বৈশিষ্ট্য কী?",	5.53
বিভূতিভূষণ বন্দ্যোপাধ্যায়ের উপন্যাসে কোনটি প্রধান বৈশিষ্ট্য?,	2.74
বুধার চরিত্রের প্রধান বৈশিষ্ট্য কী?	4.44
নজরুল ইসলামের প্রবন্ধে 'বোধন-বাঁশি' কিসের প্রতীক?,	7.61
কোন কবির কাব্যে প্রকৃতির রূপ-রস-গন্ধের সৌন্দর্যের বর্ণনা করা হয়েছে?	5.5
জীবনানন্দ দাশের কাব্যে প্রকৃতির মাহাত্ম্য কিভাবে বর্ণিত হয়েছে?	3.14
"আবদুর রহমান চরিত্রের প্রধান বৈশিষ্ট্য কী?"	3.88
"'দিবারাত্রির কাব্য' কত বছর বয়সে মানিক বন্দ্যোপাধ্যায় রচনা করেন?",	7.49
Edited and Rearranged Questions	**BanglaGPT Perplexity Score**
"উপন্যাসের অন্যতমঈও উপাদাননয়া কোনটঈ? গরুততোঁপঢ়নো",	144.83
"তারাশঙ্কর উল্লেখাযোগ্যতা বন্দ্যোপধ্যায়ের একটিও উপন্যসেরও নক কীএ?",	72.67
"কী মানিখ বন্দ্যোপাধ্যায়েইর বৈশিষ্ট্যট্যাঁ উপন্সের পধাম ?",	48.94
"বিভূতিভূষক বন্দ্যোপাধ্যাঈ উঃপোন্নাস কোণতীর পধাআঃন বৈসিসটো?",	44.41
"চরইত্রের প্রধাব বৈশসষ্ট্য কী? বুধার",	48.41
" নজরুল ইসলামের প্রবানডদহ 'বোধন-বাঁশি' কিসেঢ় প্লোতঈক?,",	102.55
"সৌন্ণর্যোর কোনা কাব্যে কবির প্রকৃতিরা রূপঅ্যা-রস-গন্ধের বর্ণনা করা হয়ছ?",	44.7
"জীবনানন্দ কাব্যেড় প্রকৃতিঈর মাহততত্ম্অ্যা দাশের কঈভাব বর্রণঈত হয়েছ?",	82.14
"আবদউঁর চরঈত্রর রহমান প্রধানট্যাঁ বৈসিসটোঈ কী?",	284.24
"'দিবারাত্রিঈর কততু বন্দ্যোপাধ্যাঈ বচর বসেঈ কাব্য মানিক রচনা করন?"	47.93

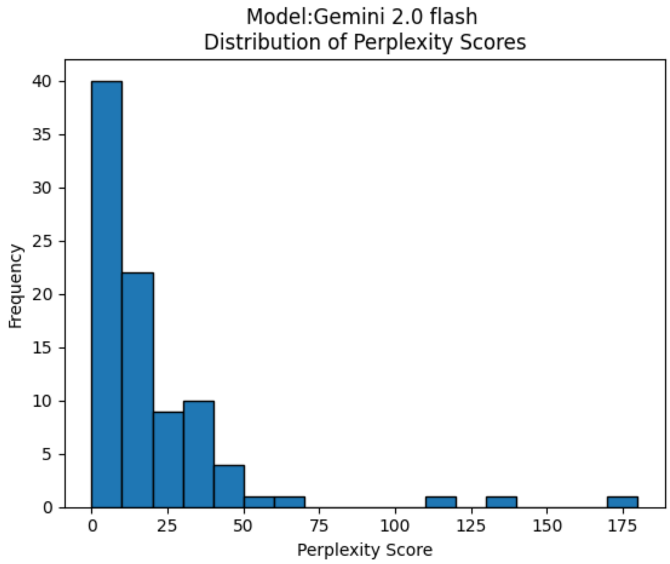
Figure 6.5: Comparison of perplexity scores between edited questions vs generated questions

From the above comparison, we can clearly notice that our generated questions have lower perplexity scores, which indicates that the generated MCQs by our framework are generally more fluent and has fewer spelling mistakes. But the edited and rearranged questions have much higher perplexity scores than our generated questions. For instance, the generated question "নজরুল ইসলামের প্রবন্ধে 'বোধন-বাঁশি' কিসের প্রতীক ?" has perplexity score of 7.61. The edited question of this generated question is "নজরুল ইসলামের প্রবানডদহ 'বোধন-বাঁশি' কিসেঢ় প্লোতঈক ?" which only has spelling mistake, but the sequence is the same as the original ones. It gets a perplexity score of 102.55, which is almost about 12 times higher than the original question's perplexity score. Again, the question "আবদুর রহমান চরিত্রের প্রধান বৈশিষ্ট্য কী ?" has a perplexity score of 3.88 but manually rearrangement and mispelling of random words for this question resulted in the following: "আবদউঁর চরঈত্রর রহমান প্রধানট্যাঁ বৈসিসটোঈ কী ?". The perplexity score of this question became 73 times higher than the originally generated question's score. This validates the effectiveness of our perplexity metric detecting sentence fluency.

Model Comparison



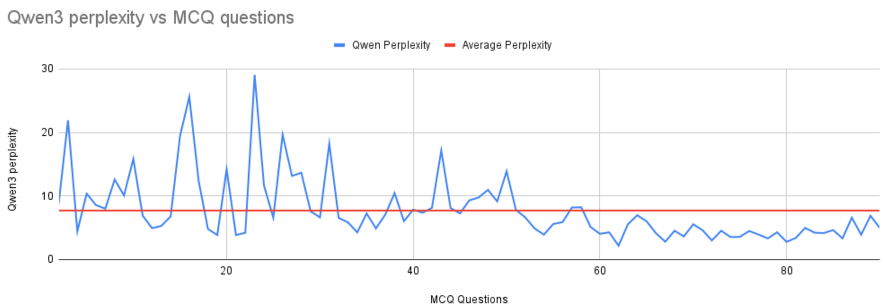
(a) Gemini Perplexity Scores



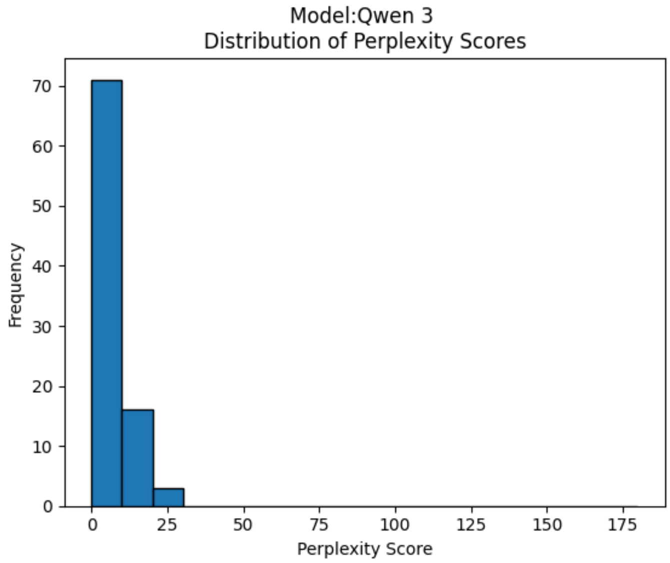
(b) Gemini 2.0 Flash Frequency Distribution

Figure 6.6: Gemini model: Perplexity scores and frequency distribution

From the above graphs of perplexity scores, we can observe that MCQs generated by Gemini 2.0 flash has the highest average of 20.41. The average of perplexity is elevated due to the presence of several MCQ questions that have exceptionally high perplexity scores, which also exhibited the irregular peaks in the graph. As the perplexity score is determined by the training data domain of BanglaGPT, it presents some intricate words as losses, consequently resulting in a high perplexity value even though the question is fluent enough. However, by doing a frequency distribution, we have found that most of the questions' perplexity range is between 10 to 20, which depicts that the fluency of our generated MCQ questions are generally good and causes less perplexity for higher frequency of time.



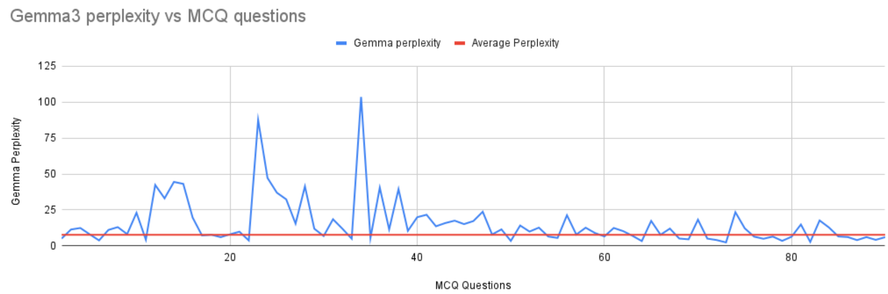
(a) Qwen Perplexity Scores



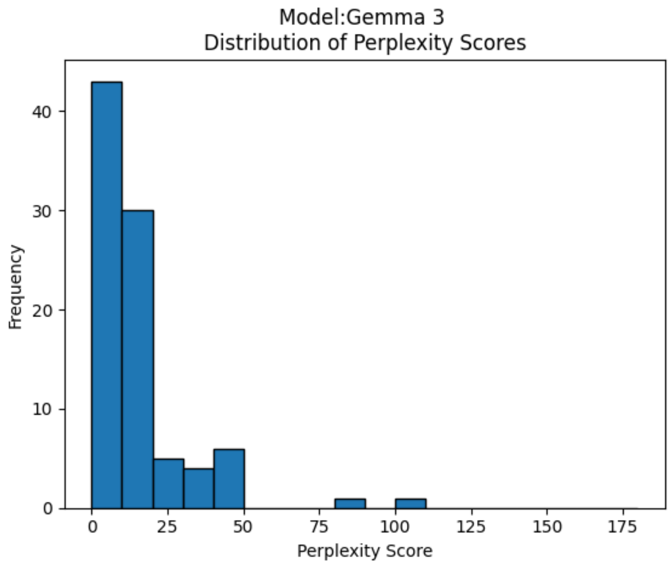
(b) Qwen 3: 14B Frequency Distribution

Figure 6.7: Qwen model: Perplexity scores and frequency distribution

The Qwen 3 model has the lowest average perplexity among the three models. It has an average of 7.74. Based on the frequency distribution, 70 questions fall under the perplexity range of 0 to 10. From the line graph, it is evident that all lines are situated within a closer proximity to the perplexity range and exhibit significantly reduced, pronounced spikes.



(a) Gemma Perplexity Scores



(b) Gemma 3: 12B Frequency Distribution

Figure 6.8: Gemma model: Perplexity scores and frequency distribution

Gemma 3 model has an average perplexity of 15.40. In the graph, we can see that a couple of big spikes are just like the Gemini 2.0 flash model. In addition, the majority of the perplexity questions fall within the range of 0 to 10. Based on the frequency distribution, there are 43 questions with perplexity values between 0 and 10, while there are 30 questions with perplexity values between 10 and 20.

6.4 Diversity Analysis

Effectiveness of the metric

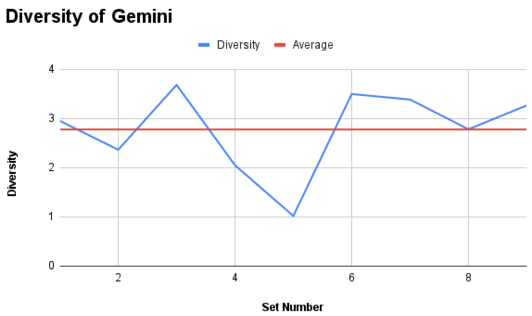
To judge the effectiveness of our diversity metric, we have randomly selected 2 sets of generated questions scored by our metric - one relatively higher scored and one lower scored in order to analyze. We wanted to see if our diversity scores could differentiate between a set with low variety and high variety of questions.

Question set with high diversity	Type	Diversity
"সরীসৃপ" গল্পের মমতাদি কাদের প্রতি মানবিক আচরণকে প্রাধান্য দিয়েছেন?	Whose	3.49
মমতাদির সংসারে কী ছিল?	What	3.49
সৈয়দ মুজতবা আলীর "দেশে বিদেশে" গ্রন্থের কোন অংশে আফগানিস্তানের কথা বলা হয়েছে?	Which	3.49
আফগানিস্তানের কোথায় আবদুর রহমান লেখকের দেখভালের দায়িত্বে ছিলেন?	Where	3.49
আবদুর রহমান চরিত্রের প্রধান বৈশিষ্ট্য কী?	What	3.49
মানিক বন্দ্যোপাধ্যায়ের আসল নাম কী ছিল?	What	3.49
মানিক বন্দ্যোপাধ্যায়ক কোন বছর বাঁকুড়া ওয়েসলিয় মিশন কলেজ থেকে আইএসসি পাস করেন?	When	3.49
"দিবারাত্রির কাব্য" কত বছর বয়সে মানিক বন্দ্যোপাধ্যায় রচনা করেন?	much/many	3.49
মমতাদির শাড়ির রঙ কেমন ছিল?	What	3.49
মমতাদি প্রথম চাকরি পেয়ে কত টাকা মাইনে আশা করেছিলেন?	much/many	3.49
Question set with less diversity	**Type**	**Diversity**
"আজ সৃষ্টি-সুখের উল্লাসে" কবিতাটি কোন কাব্যগ্রন্থ থেকে সংক্ষেপিত আকারে নেওয়া হয়েছে?	Which	1.33
কবিতায় কবি "আজ সৃষ্টি-সুখের উল্লাসে" কী অনুভব করছেন?	What	1.33
"আজ সৃষ্টি-সুখের উল্লাসে" কবিতায় কোন বিষয়টি নতুনত্বের পরিচয় বহন করছে?	Which	1.33
"আজ সৃষ্টি-সুখের উল্লাসে" কবিতায় 'পিনাক-পাণির শূল' কী প্রতীকতৎব্য?	What	1.33
"নববর্ষ" বিষয়ক অংশে টেনিসনের কবিতার মূল বক্তব্য কী?	What	1.33
"নববর্ষ" বিষয়ক অংশে রবীন্দ্রনাথ ঠাকুরের কবিতার সুর কীরূপ?	What	1.33
"জীবনানন্দ দাশের কবিতা" অংশে 'এশিরিয়া ধুলো আজ বেবিলন ছাই হয়ে আছে' - এর তাৎপর্য কী?	What	1.33
"জীবনানন্দ দাশের কবিতা" অংশে লক্ষ্মীপেঁচা গানের বিষয়ে কবি কী জানতে চেয়েছেন?	What	1.33
"আবদুল হাকিমের কবিতা" অংশের মূল বিষয়বস্তু কী?	What	1.33
আবদুল হাকিমের বিখ্যাত কাব্যগ্রন্থ কোনটি?	Which	1.33

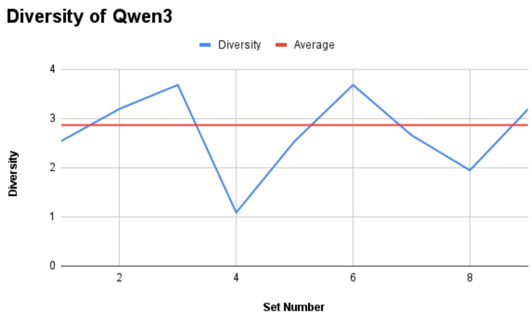
Figure 6.9: Comparison of Diversity Scores Between Diverse Set and Similar Set

From the above graphs we can clearly see that in the question set with less diversity score, there are only 2 types of questions - "Which" type and "What" type. As a result the calculated entropy resulted in a diversity score of 1.33 out of 5. On the other hand, there are 7 types of questions in the higher scored set, so the higher entropy values resulted in diversity score of 3.49. This shows that our metric effectively evaluated the diversity of the question set.

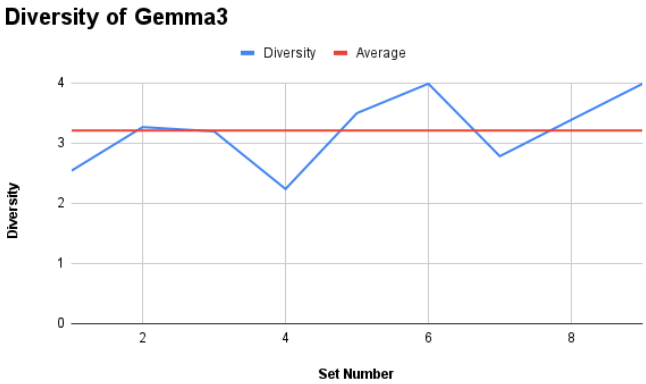
Model Comparison



(a) Gemini Diversity Score



(b) Qwen3 Diversity Score



(c) Gemma3 Diversity Score

Figure 6.10: Diversity Score comparison of Gemini, Qwen, and Gemma models.

In the graphs, we can observe that among the three generative models , Gemma 3 has the highest average score (3.2088), followed by Gemini (2.8893) and Qwen 3 (2.8364). From the evaluation data table, we can notice that most of MCQs' diversity scores are above 3. This shows that our framework generally produces diverse set of MCQs for all generative models.

6.5 Knowledge graph size dependency

We have created two approaches for both of our MCQ generation and answer extraction pipelines to detect performance variance depending on graph size and database scope. In the evaluation phase, we have noticed a performance increase in both pipelines when the graph space gets smaller. As the knowledge graph in the first approach was created from the entire book, the density of the graph was higher. This resulted in lower chances of our graph traversal reaching the exact node, where the answer of the given question lies or the given topic is most relevant with. The chance gets higher as the graph space gets smaller in the second approach where we created knowledge graphs for each chapter separately. Moreover, in the first approach, we were performing vector search in the vector database of the entire book whereas in the second approach, the search was performed only in chapter-separated

databases. This resulted in lower time complexity as well as better performance due to less irrelevant embeddings in the search domain.
This phenomenon can be especially noticed in the answer extraction performance. An empirical performance comparison between two approaches given below:

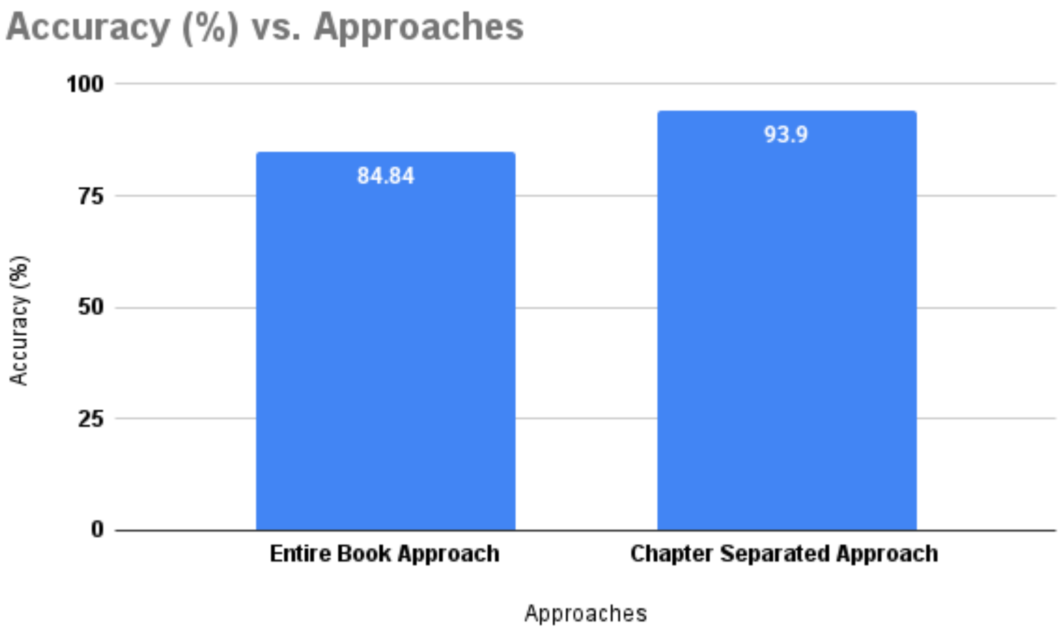


Figure 6.11: Answer Extraction Accuracy Comparison for Two Approaches

We can see that the answer extraction accuracy jumped from 84.84% to a 93.9% when we used the chapter separated approach. This comparison was done by randomly selecting 30 MCQs from our dataset for answer extraction accuracy. The bar-chart clearly depicts that reducing the graph space and database scope can significantly improve performance of answer extraction. Since the framework for MCQ generation is very similar to answer extraction, a chapter-separated approach will increase the topic relevance for generated MCQs as well. However, we have decided to show the comparison for only answer extraction, since it uses a simpler accuracy metric and the difference is more apparent here. This approach also reduces the time taken to extract the answers or generate the MCQs, as we have a smaller vector search space as well as graph traversal space.

Chapter 7

Work Plan

The first part of the thesis research starts during phase one with conducting back ground problem analysis through literature studies. We finished additional courses which supported AI development and Deep learning alongside a specialization of fered through Coursera. Among scholarly resources such as IEEE, ACM and Se mantic Scholar we gathered approximately thirty research papers to establish an integrated understanding about current work in the field. Team members performed paper selection by evaluating abstracts coupled with conclusions from published re search linked to the investigation area. Through a process that started with 30 papers we selected the 20 most appropriate ones aligned with our research objectives. The extensive amount of writings in each paper required group members to break down responsibilities for specific section analysis. Our team assembled joint thoughts to develop an initial foundation section for the literature review we plan for future use and to identify research pathways throughout this domain.

The second phase of our research contains baseline framework development complemented with preliminary analysis and data collection tasks. This phase followed our research design by constructing an academic text dataset which forms an important part of our analysis. Our collected data underwent preprocessing and formatting before its usage for framework development. During the second phase we developed a baseline framework that will act as a foundation for constructing advanced pipelines in the concluding stage. We developed a poster session during this phase for presenting overview observations and approach techniques. The next phase will benefit from our current data collection work and preliminary framework design when building a complex testing pipeline with multiple approaches.

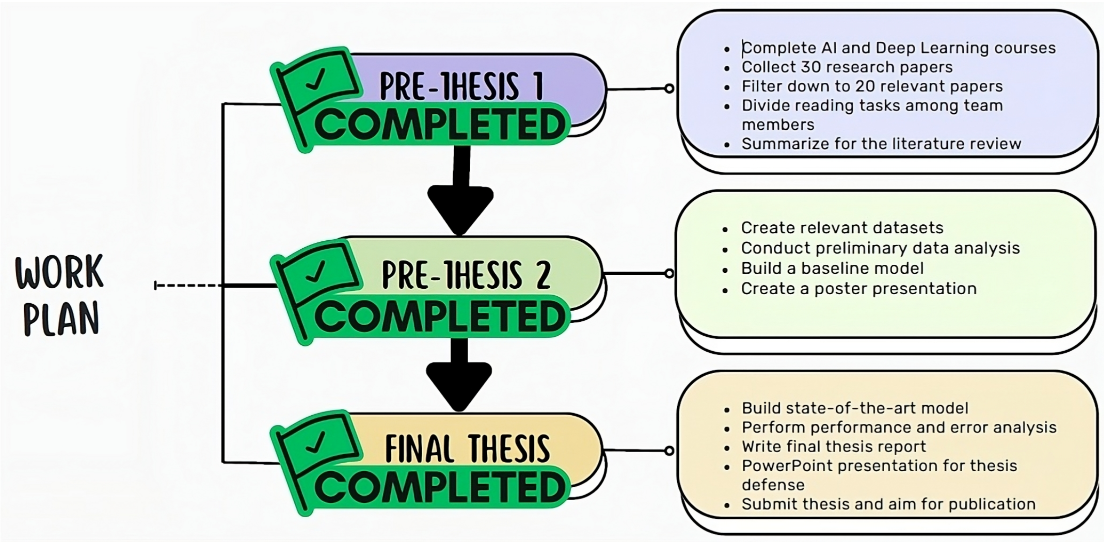


Figure 7.1: Work Flow

In the third and final phase of the thesis, we have implemented state-of-the-art pipelines for both MCQ generation and answer extraction. To enhance these pipelines, we have created knowledge graphs based on entire books as well as each chapter separately. We have designed effective evaluation metrics to evaluate our framework's generated MCQs as well as performance of our pipeline's answer extraction. We have used multiple approaches on both of the pipelines as well as the metrics to compare between these approaches. Moreover, we have performed results analysis on our evaluation data comparing our multiple approaches. Research findings regarding this project will be presented with a powerpoint slide within this section. Finally, We have formatted the paper with IEEE format to prepare for publishing.

Chapter 8

Conclusion

The primary focus of this study is developing an advanced automated MCQ generation and answer prediction system for Bengali medium SSC level students. Therefore, we have applied text extraction from scanned SSC level textbooks, pre-processing and chunking of those documents and construction of both vector based and graph based representations, in order to build state-of-the-art multiple choice question and answer generation pipelines. Leveraging a hybrid form of retrieval augmented generation that incorporates GraphRAG was our primary contribution in developing the frameworks. Another significant contribution was designing the metrics suitable for evaluating a Bengali MCQ with analytical reasoning and intuition. This study shows that our hybrid GraphRAG approach effectively produces high quality MCQs and accurate answer predictions - therefore filling a research gap in case of automated high quality assessment systems for Bengali medium.

Looking ahead, our framework can be extended by fine tuning embeddings on much larger Bangla corpora. Also, alternative algorithms to find similarity search can be experimented along with user in loop editing integration so that teachers can modify or refine generated content in real time. Moreover, this study can be expanded beyond literature heavy subjects and explore STEM domains which will broaden the impact of our system fully across the Bangla medium education in Bangladesh.

Bibliography

[1] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," 2018. arXiv: 1808.06226 [cs.CL].

[2] R. M. Asheque Siddique, "Intelligent mcq option generator and answering system for bangla language," 2020. [Online]. Available: https://dspace.mist.ac.bd/xmlui/bitstream/handle/123456789/635/1014140006_RaziaMarziaAshequeSiddique.pdf?sequence=1&isAllowed=y.

[3] S. T. Islam, "Question answering system from unstructured documents for bangla language," 2020. [Online]. Available: http://dspace.mist.ac.bd:8080/xmlui/handle/123456789/634.

[4] M. A. Rambhau, S. Gupta, M. Phadatare, I. Jadhav, and B. O. Vaibhav, "An automatic generator of multiple choice questions with random answer key," 2020. [Online]. Available: https://www.jetir.org/papers/JETIR2006113.pdf.

[5] P. Mehta, P. Jain, C. Makwana, and C. Raut, "Automated mcq generator using natural language processing," in *International Research Journal of Engineering and Technology*, 2021. [Online]. Available: https://www.irjet.net/archives/V8/i5/IRJET-V8I5497.pdf.

[6] A. C. Nwafor and I. E. Onyenwe, "An automated multiple-choice question generation using natural language processing techniques," *International Journal on Natural Language Computing*, vol. 10, no. 02, pp. 1–10, 2021. DOI: 10.5121/ijnlc.2021.10201.

[7] O. Sen, M. Fuad, M. N. Islam, *et al.*, "Bangla natural language processing: A comprehensive analysis of classical, machine learning, and deep learning-based methods," *IEEE Access*, 2021. [Online]. Available: https://www.semanticscholar.org/paper/Bangla-Natural-Language-Processing%3A-A-Comprehensive-Sen-Fuad/873cd03ff180f3bcd7ed6a7c910646e72125ffd5.

[8] A. Bhattacharjee, T. Hasan, W. U. Ahmad, *et al.*, "Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla," *Semantic Scholar*, 2022. [Online]. Available: https://www.semanticscholar.org/reader/e77c8f93bf92bc9198c3b8b981d223bf56aa707f.

[9] V. Raina and M. Gales, *Multiple-choice question generation: Towards an automated assessment framework*, 2022. arXiv: 2209.11830 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2209.11830.

[10] R. Anil, S. Borgeaud, J. Alayrac, *et al.*, "Gemini: A family of highly capable multimodal models," 2023. arXiv: 2312.11805 [cs.LG].

[11] H. W. Chung, N. Constant, X. Garcia, *et al.*, "Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining," 2023. arXiv: 2304.09151 [cs.CL].

[12] S. Deode, J. Gadre, A. Kajale, A. Joshi, and R. Joshi, "L3cube-indicsbert: A simple approach for learning cross-lingual sentence representations using multilingual bert," 2023. arXiv: 2304.11434 [cs.CL].

[13] N. Diamant, *Rag techniques*, https://github.com/NirDiamant/RAG_Techniques, Accessed: 2025-06-10, 2023.

[14] S. M. Kumar, S. Srikanth, and S. Hegde, "Optimized nlp model for mcq generation through advanced batching and tokenization," *International Journal & Research Paper Publisher / IJRASET*, 2023. [Online]. Available: https://www.ijraset.com/best-journal/optimised-nlp-model-for-mcq-generation-through-advanced-batching-and-tokenization.

[15] F. Martinez, D. Collarana, D. Calvaresi, M. Arispe, C. Florida, and J.-P. Calbimonte, "Study-buddy: A knowledge graph-powered learning companion for school students," in *Proceedings of the Extended Semantic Web Conference*, 2023.

[16] OpenAI, *GPT-4 Architecture, Infrastructure, Training Dataset, Costs*, https://semianalysis.com/2023/07/10/gpt-4-architecture-infrastructure/, Retrieved from SemiAnalysis, 2023.

[17] OpenAI, "GPT-4 Technical Report," OpenAI, Tech. Rep., 2023. [Online]. Available: https://openai.com/research/gpt-4.

[18] D. Pochiraju, P. Chimulla, A. Chakilam, D. G. Rao, and P. Betham, "Extractive summarization and multiple choice question generation using xlnet," in *International Conference on Intelligent Computing and Control Systems*, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Extractive-Summarization-and-Multiple-Choice-using-Pochiraju-Chimulla/8976ef440ae6fa8d12b422237f3 utm_source=direct_link.

[19] B. H. Reddy, V. K. Reddy, M. D. Kumar, and G. Vishal, "Mcq question generation using natural language processing techniques," *International Journal for Multidisciplinary Research*, 2023. [Online]. Available: https://www.semanticscholar.org/paper/MCQ-Question-Generation-using-Natural-Language-Reddy-Reddy/38a1fcc0e6a19163df23cb15841d2698a1873874.

[20] M. S. Salim, H. Murad, D. Das, and F. Ahmed, "Banglagpt: A generative pretrained transformer-based model for bangla language," in *2023 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, 2023, pp. 56–59. DOI: 10.1109/ICICT4SD59951.2023.10303383.

[21] C. Zhang, "Automatic generation of multiple-choice questions," *Semantic Scholar*, 2023, PDF available. [Online]. Available: https://www.semanticscholar.org/reader/02eabbf006f83aa8b80697cfab0f8dc808094c12.

[22] Anonymous, "Recent advances in transformer models," *arXiv.org e-Print archive*, 2024. [Online]. Available: https://arxiv.org/pdf/2403.05530.

[23] C. N. Hang, C. W. Tan, and P.-D. Yu, "Mcqgen: A large language model-driven mcq generator for personalized learning," *IEEE Access*, vol. 12, pp. 102 261–102 273, 2024. DOI: 10.1109/access.2024.3420709.

[24] S. Maity, A. Deroy, and S. Sarkar, "A novel multi-stage prompting approach for language agnostic mcq generation using gpt," in *Proceedings of the European Conference on Information Retrieval*, 2024. [Online]. Available: https://www.semanticscholar.org/paper/A-Novel-Multi-Stage-Prompting-Approach-for-Language-Maity-Deroy/39cc18f3e4ab4aa98b3fbaa225c6b16a48d40cfb?utm_source=direct_link.

[25] S. C. Roy and M. M. Hossen Manik, "Question-answering system for bangla: Fine-tuning bert-bangla for a closed domain," *arXiv.org e-Print archive*, 2024. [Online]. Available: https://www.arxiv.org/pdf/2410.03923.

[26] A. Yang, B. Yang, B. Zhang, *et al.*, "Qwen2.5 technical report," 2024. arXiv: 2412.15115 [cs.LG].

[27] A. Kamath, J. Ferret, *et al.*, "Gemma 3 technical report," 2025. arXiv: 2503.19786 [cs.LG].

[28] Chorcha, *Question Bank*, https://chorcha.net/question-bank, Accessed: 2025-06-11.

[29] A. A. Asif and H. A. Emon, *Porag: Fully configurable rag pipeline for bengali language rag applications*, https://github.com/Bangla-RAG/PoRAG, [GitHub repository], n.d.

[30] Google, *Gemini 2.0 flash*, https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash, [Web page], n.d.

[31] R. Smith and G. Inc., *An overview of the tesseract ocr engine*, n.d. [Online]. Available: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/33418.pdf.

Full Evaluation Results:

Book Name	Model	Topic	MCQ Question	Distractor Quality Score (Our Metric)	Distractor Quality Score (AI Feedback)	Ai Feedback	Contextual Relevance Score	Perplexity Score (BanglaGPT)	Diversity Score Per Set
Shohopath	Gemini 2.0 Flash	"কাকতাড়ুয়া উপন্যাসে বুধার দেশপ্রেম"	{"question": "কাকতাড়ুয়া উপন্যাসের প্রধান চরিত্র বুধার দেশপ্রেমের মূলে কোনটি সবচেয়ে গুরুত্বপূর্ণ?","correctAns": "দেশের মানুষের প্রতি মমত্ববোধ","distractor1": "ভয় না পাওয়া","distractor2": "বাবা-মার অভাব"," distractor3": "একা বেড়ে ওঠা"},	8.044	6.77531	বিকল্পগুলো আংশিক প্রাসঙ্গিক হলেও সঠিক উত্তরের গভীর মানসিক দৃষ্টিভঙ্গির মতো বাস্তবিক নয়। তাই নিম্ন-মধ্য মানের এমসিকিউ।	75.24%	13.57	2.95
			{"question": "বুধা কীভাবে ধীরে ধীরে মুক্তিযোদ্ধা হয়ে ওঠে?","correctAns": "দেশকে ভালোবাসে বলে","distractor1": "গল্পের বই পড়ে","distractor2": "বিদেশি মিলিটারিদের সাথে বন্ধুত্ব করে","distractor3": "গ্রামের মানুষের নির্দেশে"},	8.48089	7.02359	দ্বিতীয় ও তৃতীয় distractor কিছুটা দূরে সরে গেলেও প্রথমটি যথেষ্ট সম্পর্কযুক্ত। সামগ্রিক মান মাঝারি উচ্চ।	60.18%	10.63	
			{"question": "কাকতাড়ুয়া উপন্যাসে বুধার কোন বৈশিষ্ট্যটি বিশেষভাবে উল্লেখযোগ্য?"," correctAns": "অসীম সাহস ও মানবিক গুণাবলি","distractor1": "গান গাওয়ার প্রতিভা","distractor2": "অসাধারণ স্মৃতিশক্তি","distractor3": "চমৎকার ছবি আঁকার ক্ষমতা"},	8.16002	8.76422	তিনটি distractor-ই যথেষ্ট বাস্তবিক ও কাছাকাছি, সঠিক উত্তর থেকে কিছুটা দূরে থাকলেও যুক্তিযুক্ত।	67.55%	11.32	
			{"question": "বুধা শান্তি কমিটি ও রাজাকার কমান্ডারের বাড়িতে আগুন দেওয়ার কারণ কী?","correctAns": "দেশপ্রেম ও দেশের প্রতি ভালোবাসা","distractor1": "ব্যক্তিগত শত্রুতা","distractor2": "গ্রামের মানুষের অনুরোধ","distractor3": "বন্ধুদের প্ররোচনা"},	7.07108	8.55661	সব distractor বাস্তবিক, তবে সঠিক উত্তরের দেশপ্রেমিক দৃষ্টিকোণ সরাসরি নয়, তাই সামান্য কম নম্বর।	47.78%	16.3	
			{"question": "কাকতাড়ুয়া উপন্যাসে বুধার মধ্যে কিসের অভাব তাকে সাহসী করে তোলে বলে মনে করা হয়?","correctAns": "ভয়ের অভাব","distractor1": "অর্থের অভাব"," distractor2": "শিক্ষার অভাব","distractor3": "ভালোবাসার অভাব"},	8.7238	7.33211	"ভয়ের অভাব" একমাত্র ঘনিষ্ঠ distractor, বাকিগুলো মানসিক অবস্থার পরিবর্তে বাহ্যিক কারণ — ফলে মাঝারি মান।	73.20%	21.51	
			{"question": "বুধা যখন মুক্তিযুদ্ধের কথা ভাবতো, তখন কার স্বাধীনতার আহ্বান তাকে উদ্দীপ্ত করতো?","correctAns": "বঙ্গবন্ধু শেখ মুজিবুর রহমান","distractor1": "মহাত্মা গান্ধী","distractor2": "নেতাজি সুভাষচন্দ্র বসু","distractor3": "শেরে বাংলা এ. কে. ফজলুল হক"},	7.77364	6.44325	শুধুমাত্র সঠিক উত্তর বঙ্গবন্ধুর প্রাসঙ্গিকতা ধরে রেখেছে, বাকিগুলো বাস্তবতা থেকে বিচ্যুত — মান তুলনামূলক কম।	57.36%	5.53	
			{"question": "কাকতাড়ুয়া উপন্যাসে বুধার স্বাধীনতা কোন ধরনের বর্ণনার মাধ্যমে ফুটিয়ে তোলা হয়েছে?","correctAns": "রূপক ও প্রতীকী","distractor1": "সরাসরি ও সরল"," distractor2": "জটিল ও দার্শনিক"," distractor3": "অলঙ্কারপূর্ণ ও কাব্যিক"},	9.43546	7.94177	distractor-গুলো ভাষাগত দিক থেকে ভিন্ন হলেও কিছুটা অর্থবোধক মিল আছে — রূপক অর্থে সঠিক উত্তর আলাদা।	67.52%	28.82	
			{"question": "নোলক বুয়া বুধাকে কী খেতে দিত?","correctAns": "মুড়ি ভাজা"," distractor1": "ফল","distractor2": "নদীর জল","distractor3": "পাকুড়"},	8.9015	5.66149	শুধু distractor1 খানিকটা যুক্তিযুক্ত, বাকিগুলো অতিরঞ্জিত বা বাস্তবতাহীন, তাই নিম্ন-মধ্য মানের প্রশ্ন।	42.40%	40.55	
			{"question": "কাকতাড়ুয়া উপন্যাসে বুধার বেড়ে ওঠা কোন প্রেক্ষাপটে দেখানো হয়েছে?"," correctAns": "মুক্তিযুদ্ধের সময়কার সংকটময় পরিস্থিতি","distractor1": "শহরের আধুনিক জীবন","distractor2": "গ্রামীণ সমাজের সরলতা","distractor3": "বিদেশি সংস্কৃতির প্রভাব"},	8.99919	8.11088	বিকল্পগুলো সময়, স্থান ও পরিপ্রেক্ষিতের কিছুটা মিল রাখে, তবে সঠিক উত্তর বেশি প্রাসঙ্গিক ও নির্ভুল।	68.20%	36.6	

			{"question": "কাকতাড়ুয়া উপন্যাসে বুধার চরিত্রকে কোন লক্ষ্যের দিকে এগিয়ে নিয়ে যাওয়া হয়েছে?","correctAns": "দেশপ্রেমের আদর্শ স্থাপন করতে","distractor1": "একজন সফল ব্যবসায়ী হিসেবে","distractor2": "একজন জনপ্রিয় নেতা হিসেবে","distractor3": "একজন বিখ্যাত শিল্পী হিসেবে"}	**7.25192**	**8.33496**	সব distractor কিছুটা অপ্রাসঙ্গিক হলেও বাস্তবিক লক্ষ্য হিসেবে আলাদা; মূল উত্তর যথাযথভাবে লক্ষ্য নির্ধারণ করে।	**67.22%**	**63.37**	
		"সৈয়দ ওয়ালীউল্লাহ এর লেখক পরিচিতি"	{"question": "সৈয়দ ওয়ালীউল্লাহ কত সালে জন্মগ্রহণ করেন?","correctAns": "১৯২২"," distractor1": "১৯২১","distractor2": "১৯২৩","distractor3": "১৯২৪"},	**6.22492**	**9.22314**	সকল distractor সময়কালঘেঁষা ও বাস্তবিকভাবে কাছাকাছি; একে খুব ভালো মানের প্রশ্ন হিসেবে ধরা যায়।	**28.26%**	**42.38**	**2.37**
			{"question": "সৈয়দ ওয়ালীউল্লাহর প্রথম প্রকাশিত গল্পগ্রন্থের নাম কী?","correctAns": "নয়নচারা","distractor1": "লালসালু"," distractor2": "দুই তীর ও অন্যান্য গল্প"," distractor3": "চাঁদের অমাবস্যা"},	**9.23433**	**7.66257**	"নয়নচারা" ব্যতীত বাকিগুলো গল্পগ্রন্থ নয় বা বিভ্রান্তিকরভাবে উপন্যাস — তাই distractor মান মাঝারি।	**53.90%**	**19.66**	
			{"question": "'বহিপীর' নাটকটি প্রথম কত সালে প্রকাশিত হয়?","correctAns": "১৯৬০","distractor1": "১৯৫৫"," distractor2": "১৯৫৯","distractor3": "১৯৬৪"},	**9.33807**	**9.00822**	সালভিত্তিক বিভ্রান্তি খুব ঘনিষ্ঠ ও সম্ভাব্য; খুব ভালো distractor ব্যবহার হয়েছে, সঠিকতার খুব কাছাকাছি।	**67.64%**	**57.84**	
			{"question": "সৈয়দ ওয়ালীউল্লাহর পিতার পেশা কী ছিল?","correctAns": "সরকারি কর্মকর্তা","distractor1": "শিক্ষকতা"," distractor2": "সাংবাদিকতা","distractor3": "ওকালতি"},	**6.34213**	**8.77411**	সব option প্রাসঙ্গিক পেশা — শিক্ষকতা ও ওকালতি বিশেষ ঘনিষ্ঠ নয়, তবে পুরোপুরি দূরে নয় — ভালো মান।	**36.09%**	**38.25**	
			{"question": "সৈয়দ ওয়ালীউল্লাহ কোন সালে প্যারিসে মৃত্যুবরণ করেন?","correctAns": "১৯৭১","distractor1": "১৯৭০"," distractor2": "১৯৭২","distractor3": "১৯৭৩"},	**8.89766**	**8.88743**	দ্রষ্টব্য সালগুলোর মান ঘনিষ্ঠ — distractor মান বেশ ভালো, প্রশ্ন বাস্তবিক বিভ্রান্তি তৈরি করে।	**31.85%**	**23.03**	
			{"question": "সৈয়দ ওয়ালীউল্লাহ কত সালে বাংলা একাডেমী পুরস্কার লাভ করেন?"," correctAns": "১৯৬১","distractor1": "১৯৫৫","distractor2": "১৯৬০"," distractor3": "১৯৬৫"},	**6.42658**	**8.99325**	সালভিত্তিক প্রশ্নের distractor-গুলো যথেষ্ট কাছাকাছি এবং বাস্তবিক — উত্তম distractor গঠন।	**46.39%**	**14.7**	
			{"question": "'বহিপীর' নাটকের কেন্দ্রীয় চরিত্র কীসের প্রতীক?","correctAns": "কুসংস্কার ও ধর্মীয় বইয়ের জ্ঞান","distractor1": "গ্রামীণ সমাজ","distractor2": "শহুরে জীবন"," distractor3": "রাজনৈতিক অস্থিরতা"},	**7.35036**	**7.41958**	কেন্দ্রীয় প্রতীকের ব্যাখ্যা distractor-এ খুব কম; তবে কিছুটা ঘনিষ্ঠ — ফলে মান মাঝারি থেকে কিছু বেশি।	**52.18%**	**20.64**	
			{"question": "সৈয়দ ওয়ালীউল্লাহর কোন উপন্যাসটি ১৯৬৮ সালে প্রকাশিত হয়?"," correctAns": "কাঁদো নদী কাঁদো"," distractor1": "লালসালু","distractor2": "চাঁদের অমাবস্যা","distractor3": "সুড়ঙ্গ"},	**8.22367**	**9.11009**	সালসংক্রান্ত distractor-গুলো যথেষ্ট ঘনিষ্ঠ ও বাস্তবিক — সঠিক উত্তরের কাছাকাছি সম্ভাব্য বিভ্রান্তি দেয়।	**37.30%**	**36.97**	
			{"question": "'বহিপীর' নাটকটি প্রকাশের পূর্বে কোন ক্লাবের উদ্যোগে পুরস্কার লাভ করে?"," correctAns": "পিইএন ক্লাব","distractor1": "ঢাকা থিয়েটার ক্লাব","distractor2": "বাংলা একাডেমী ক্লাব","distractor3": "শিল্পকলা একাডেমি ক্লাব"},	**9.26999**	**7.88742**	সব ক্লাবের নাম বাস্তবিক তবে context-specific মিল কম — সঠিক উত্তর স্পষ্টভাবে আলাদা।	**51.53%**	**7.55**	

			{"question": "সৈয়দ ওয়ালীউল্লাহ্ মৃত্যুত্তোর কোন পদক লাভ করেন?","correctAns": "একুশে পদক","distractor1": "পিইএন পুরস্কার","distractor2": "আদমজী পুরস্কার"," distractor3": "বাংলা একাডেমী পুরস্কার"}	**9.51229**	**8.11557**	সব option পুরস্কারের নাম তবে মৃত্যুর পর পাওয়া একুষে পদকের বাস্তব প্রতিযোগী কম, কিন্তু distractor ভাল।	**28.92%**	**36.69**	
		"বহিপীর নাটকে বহিপীর এবং তাহেরার সম্পর্ক"	{"question": "বহিপীর নাটকের কেন্দ্রীয় চরিত্র কে?","correctAns": "বহিপীর"," distractor1": "তাহেরা","distractor2": "হাতেম আলি","distractor3": "হাশেম আলি"},	**8.44174**	**8.66511**	সব চরিত্র নাটকের প্রাসঙ্গিক, তবে সঠিক চরিত্রের weight বেশি স্পষ্ট — ভাল distractor।	**24.62%**	**30.89**	**3.68**
			{"question": "বহিপীর কত বছর পর পর শিষ্যদের বাড়িতে যেতেন?","correctAns": "দুই বছর পর পর","distractor1": "প্রতি বছর"," distractor2": "তিন বছর পর পর"," distractor3": "পাঁচ বছর পর পর"},	**8.0405**	**7.445**	বছরের ঘনিষ্ঠতা থাকলেও distractor-এ বাস্তবতার ঘাটতি কিছুটা আছে — মাঝারি মান।	**32.79%**	**46.37**	
			{"question": "তাহেরাকে কার সাথে জোর করে বিয়ে দেওয়া হয়েছিল?","correctAns": "বহিপীর","distractor1": "হাশেম আলি"," distractor2": "হাতেম আলি","distractor3": "হকিকুল্লাহ"},	**8.15294**	**7.11846**	বিয়ে সম্পর্কিত বিকল্পগুলো চরিত্রভিত্তিক হলেও মূল উত্তরের মতো স্পষ্ট নয় — মাঝারি মান।	**22.90%**	**45.29**	
			{"question": "পালিয়ে যাওয়ার পর তাহেরা প্রথমে কোথায় আশ্রয় নেয়?","correctAns": "হাতেম আলি জমিদারের শহরগামী বজরায়"," distractor1": "বহিপীরের নৌকায়"," distractor2": "হাশেম আলির বাড়িতে"," distractor3": "একটি গ্রামে"},	**7.90612**	**6.99942**	প্রথম distractor আংশিক সঠিক, বাকিগুলো ভিন্ন পথে চলে গেছে — মান নিচু-মধ্য।	**21.87%**	**133.91**	
			{"question": "বহিপীর তাহেরাকে খুঁজতে গিয়ে কার বজরায় আশ্রয় নেয়?","correctAns": "হাতেম আলি","distractor1": "হাশেম আলি"," distractor2": "হকিকুল্লাহ","distractor3": "কোনোটিই নয়"},	**6.68957**	**6.45376**	শুধু distractor1 প্রাসঙ্গিক, বাকিগুলো চরিত্রের সাথে মেলে না — বাস্তবতা কম।	**41.63%**	**117.93**	
			{"question": "জমিদার পুত্র হাশেম আলি তাহেরার প্রতি সহানুভূতিশীল হওয়ার কারণ কী?","correctAns": "তাহেরার করুণ কাহিনী","distractor1": "তাহেরার রূপ"," distractor2": "তাহেরার বাবার আদেশ"," distractor3": "বহিপীরের প্রতি বিদ্বেষ"},	**9.41649**	**7.8812**	কারণগুলো আংশিক বাস্তবিক — প্রধান উত্তরের মতো করুণ আবেগ নেই, তবে বিভ্রান্তি তৈরি করে।	**31.04%**	**8.25**	
			{"question": "নাটকে বহিপীর কীসের প্রতীক হিসেবে উপস্থাপিত হয়েছে?","correctAns": "সর্বগ্রাসী স্বার্থের","distractor1": "নতুন দিনের","distractor2": "সর্বংসহা নারীর"," distractor3": "বিদ্রোহের"},	**8.31891**	**8.77422**	প্রতীক বিশ্লেষণ ভালো, distractor প্রাসঙ্গিকতা ধরে রেখেছে — ভালো মানের প্রশ্ন।	**28.92%**	**20.68**	
			{"question": "তাহেরা শেষ পর্যন্ত কার সাথে পালিয়ে যায়?","correctAns": "হাশেম আলি"," distractor1": "বহিপীর","distractor2": "হাতেম আলি","distractor3": "হকিকুল্লাহ"},	**8.22753**	**8.33411**	সব বিকল্প বিশ্বাসযোগ্য চরিত্র, সঠিক উত্তরের সাথে বাস্তব মিল — ভালো distractor।	**22.57%**	**26.32**	
			{"question": "বহিপীর নাটকে কার বিদ্রোহের কাহিনী ফুটিয়ে তোলা হয়েছে?","correctAns": "তাহেরার","distractor1": "জমিদারের স্ত্রীর"," distractor2": "হাশেম আলির","distractor3": "হকিকুল্লাহর"},	**9.1653**	**8.00672**	বিদ্রোহ বিষয়ক distractor-গুলো আংশিকভাবে প্রাসঙ্গিক, তবে মূল অর্থের গভীরতা কম — মান ভালো।	**26.73%**	**35.04**	
			{"question": "বহিপীরের সঙ্গীর নাম কী ছিল?","correctAns": "হকিকুল্লাহ"," distractor1": "হাশেম আলি","distractor2": "হাতেম আলি","distractor3": "কাদের"}	**9.00776**	**7.55111**	সব চরিত্র নাটকে আছে, তবে সঠিক সঙ্গীর সুনির্দিষ্টতা স্পষ্ট — মাঝারি মান।	**27.81%**	**19.01**	

	Qwen 3: 14B	"কাকতাড়ুয়া উপন্যাসে বুধার দেশপ্রেম"	{"question": "কাকতাড়ু উপন্যাসের প্রধান চরিত্র কে?","correctAns": "বুধা"," distractor1": "কুন্তি","distractor2": "নোলক বুমা","distractor3": "বঙ্গবন্ধু"},	**8.39943**	**8.22164**	ডিস্ট্র্যাক্টরগুলো বেশিরভাগই অর্থঘন, তবে কিছু দূরবর্তী অপশন রয়েছে; মোটামুটি উচ্চ মানের এমসিকিউ।	**52.89%**	**8.8**	**2.54**
			{"question": "বুধার চরিত্রটিকে গড়ে তুলেছেন কে?","correctAns": "সেলিনা হোসেন"," distractor1": "বঙ্গবন্ধু","distractor2": "মুক্তিযোদ্ধা শিল্পী শাহাবুদ্দিন","distractor3": "রাজাকার কুদ্দুস"},	**7.07909**	**8.13625**	বিকল্পগুলো তথ্যভিত্তিক, তবে কয়েকটি অপশন যথেষ্ট দূরবর্তী; ভাল মানের এমসিকিউ।	**62.08%**	**21.91**	
			{"question": "বুধার চরিত্রের প্রধান বৈশিষ্ট্য কী?","correctAns": "বিদেশি মিলিটারিদের প্রতি ঘৃণা","distractor1": "মানুষের প্রতি ঘৃণা","distractor2": "বিদেশি মিলিটারিদের প্রতি ভালোবাসা","distractor3": "মানুষের প্রতি ঘৃণা"},	**8.10899**	**7.94257**	ডিস্ট্র্যাক্টরগুলো অনেকটাই বিপরীতমুখী; মাঝারি মানের এমসিকিউ।	**44.46%**	**4.44**	
			{"question": "কাকতাড়ু উপন্যাসে বুধার চরিত্রটি কিশোর হলেও কী গুণাবলির অধিকারী?","correctAns": "সাহস আর মানবিক গুণাবলি","distractor1": "সাহস আর আত্মবিশ্বাস","distractor2": "সাহস আর ঘৃণা"," distractor3": "সাহস আর ক্ষমা"},	**9.34444**	**8.31495**	বিকল্পগুলো সঠিক উত্তরের কাছাকাছি, ভালো ভারসাম্যপূর্ণ ডিস্ট্র্যাক্টর; উচ্চ মানের এমসিকিউ।	**74.11%**	**10.37**	
			{"question": "বুধা কিশোর হলেও কী দুঃখের সম্মুখীন হয়?","correctAns": "বাবা-মা, ভাই-বোন মারা যাওয়া","distractor1": "বাবা-মা, ভাই-বোন ভালোবাসা","distractor2": "বাবা-মা, ভাই-বোন সাহায্য করা","distractor3": "বাবা-মা, ভাই-বোন বাঁচানো"},	**8.46968**	**7.82463**	ডিস্ট্র্যাক্টরগুলো আংশিক অর্থবহ, কিছু দূরবর্তী অপশন আছে; মাঝারি মানের এমসিকিউ।	**61.13%**	**8.56**	
			{"question": "কাকতাড়ু উপন্যাসে বুধার চরিত্রটি কোন প্রেরণায় মুক্তিযোদ্ধা হয়ে ওঠে?"," correctAns": "বঙ্গবন্ধুর স্বাধীনতার আহ্বান"," distractor1": "বাবা-মা মারা যাওয়া"," distractor2": "চাচির বাড়ি ছেড়ে আসা"," distractor3": "মিলিটারির নির্মম আচরণ"},	**8.89797**	**8.13752**	ডিস্ট্র্যাক্টরগুলো বেশ অর্থঘন, তবে কিছুটা অধিক সরলীকৃত; মোটামুটি উচ্চ মানের এমসিকিউ।	**72.81%**	**8.02**	
			{"question": "কাকতাড়ু উপন্যাসের ভাষাভঙ্গি কেমন?","correctAns": "সরল আর আকর্ষণীয়","distractor1": "জটিল আর বিষয়গত","distractor2": "ঘৃণার আর সহজ"," distractor3": "অস্পষ্ট আর জটিল"},	**8.90899**	**8.25984**	বিকল্পগুলো অর্থবহ এবং পরীক্ষামূলক; যথেষ্ট উপযুক্ত ডিস্ট্র্যাক্টর; উচ্চ মানের এমসিকিউ।	**53.36%**	**12.59**	
			{"question": "কাকতাড়ু উপন্যাসের কাহিনী গেঁথে তুলেছেন কে?","correctAns": "সেলিনা হোসেন","distractor1": "বঙ্গবন্ধু"," distractor2": "মুক্তিযোদ্ধা শিল্পী শাহাবুদ্দিন"," distractor3": "রাজাকার কুদ্দুস"},	**7.07909**	**8.13654**	তথ্যভিত্তিক ডিস্ট্র্যাক্টর, তবে দূরবর্তী অপশন কিছু আছে; মোটামুটি ভাল মানের এমসিকিউ।	**55.01%**	**10.08**	
			{"question": "কাকতাড়ু উপন্যাসে বুধার বর্ণনার ইঙ্গিত কী?","correctAns": "বুধা একজন স্বাধীন মানুষ","distractor1": "বুধা একজন বন্দী","distractor2": "বুধা একজন ঘৃণাকর মানুষ","distractor3": "বুধা একজন বিদেশি মিলিটারি"},	**8.36076**	**7.94321**	কিছু ডিস্ট্র্যাক্টর অর্থঘন হলেও, অনেকের দূরবর্তীতা বেশি; মাঝারি মানের এমসিকিউ।	**65.78%**	**15.83**	
			{"question": "কাকতাড়ু উপন্যাসে বুধার চরিত্রটি কোন প্রাসঙ্গিক অংশে বর্ণনা করা হয়েছে?","correctAns": "পথ ওকে ডাকে-আগে। ও হেসে বলে, এই তো এসেছি। দেখ আমাকে। পথ ওকে দেখে।","distractor1": "ঘুমুবার জায়গা নেই তো কী হয়েছে? দোকানের বারান্দা আছে, গাছতলা আছে, হাটের চালা আছে, ঘাটে বাঁধা নৌকা আছে।","distractor2": "সব বেলার ভাত নেই তো কী হয়েছে? ফলপাকুড় আছে, নদীর পানি আছে। নোলক বুমার মুসুমি ভাজা আছে।","distractor3": "সব প্রাসঙ্গিক অংশ একটি বর্ণনা।"},	**5.99126**	**7.82475**	বিকল্পগুলোর তথ্যঘনতা আছে তবে সরাসরি সম্পর্ক অনেক কম; মাঝারি মানের এমসিকিউ।	**65.41%**	**6.92**	

		"সৈয়দ ওয়ালীউল্লাহ এর লেখক পরিচিতি"	{"question": "সৈয়দ ওয়ালীউল্লাহ কোন বছরে জন্মগ্রহণ করেন?","correctAns": "১৯২০"," distractor1": "১৯২২","distractor2": "১৯২৫","distractor3": "১৯২৮"},	**9.57751**	**8.21947**	ডিস্ট্র্যাক্টরগুলো যথেষ্ট কাছাকাছি; তথ্যভিত্তিক এবং পরীক্ষামূলক উপযুক্ত এমসিকিউ।	**29.18%**	**4.95**	
			{"question": "সৈয়দ ওয়ালীউল্লাহের পিতার পেশা কী ছিল?","correctAns": "সরকারি কর্মকর্তা","distractor1": "শিক্ষক"," distractor2": "চিকিৎসক","distractor3": "কৃষক"},	**4.92626**	**8.06352**	বিকল্পগুলো অর্থঘন, তবে কিছু ভিন্নমুখী অপশন রয়েছে; মোটামুটি ভাল মানের।	**36.12%**	**5.29**	
			{"question": "সৈয়দ ওয়ালীউল্লাহের প্রথম খান্দান কোনটি?","correctAns": "লালসালু"," distractor1": "নামা খান্দান","distractor2": "রূপান্তর","distractor3": "বাহি পির"},	**8.3377**	**7.91542**	কিছু ডিস্ট্র্যাক্টর কাছাকাছি হলেও অধিকাংশ তুলনামূলক দূরবর্তী; মাঝারি মানের এমসিকিউ।	**36.08%**	**6.79**	
			{"question": "বাহি পির নাটকটি কোন বছর প্রকাশিত হয়?","correctAns": "১৯৬০"," distractor1": "১৯৫৫","distractor2": "১৯৬২","distractor3": "১৯৬৬"},	**9.73737**	**8.24513**	ডিস্ট্র্যাক্টরগুলো যথেষ্ট অর্থবহ; মোটামুটি উপযুক্ত বিকল্প; উচ্চ মানের এমসিকিউ।	**55.94%**	**19.4**	
			{"question": "“বাহি পির” নাটকটির নামকরণের কারণ কী?","correctAns": "পিরদের ধর্মীয় প্রথা সম্পর্কে সংস্কার চাওয়া"," distractor1": "বাঙালি মুসলমান সমাজের সমাজবাস্তবতা চিত্রিত করা","distractor2": "সৈয়দ ওয়ালীউল্লাহের পিতার পেশা সম্পর্কে বর্ণনা","distractor3": "পূর্ববঙ্গের নাট্যচর্চার ইতিহাস উপস্থাপন"},	**6.0757**	**8.09235**	বিকল্পগুলো বেশ যুক্তিযুক্ত, তবে সামান্য বৈচিত্র্য থাকলে আরও ভালো হতো।	**61.35%**	**25.61**	**3.2**
			{"question": "বাহি পির নাটকটি কোন সমাজের চিত্র দেখায়?","correctAns": "বাঙালি মুসলমান সমাজ","distractor1": "কৃষক সমাজ","distractor2": "শিক্ষিত সমাজ"," distractor3": "প্রাচীন বাঙালি সমাজ"},	**6.45841**	**8.13645**	তথ্যঘন ডিস্ট্র্যাক্টর, প্রাসঙ্গিকতা মোটামুটি ভাল; উচ্চ মানের।	**54.69%**	**12.38**	
			{"question": "সৈয়দ ওয়ালীউল্লাহ কোন নাটক রচনা করেন?","correctAns": "বাহি পির"," distractor1": "রক্তাক্ত প্রান্তর","distractor2": "তিতুমীর","distractor3": "রক্তপদ্ম"},	**8.24665**	**8.17834**	তথ্যঘন বিকল্প, কিছু দূরবর্তী অপশন থাকলেও পরীক্ষার জন্য উপযুক্ত।	**52.28%**	**4.81**	
			{"question": "সৈয়দ ওয়ালীউল্লাহ কোন পুরস্কার প্রাপ্ত হন?","correctAns": "সাহিত্য অকাদেমি পুরস্কার","distractor1": "জাতীয় পুরস্কার","distractor2": "আসকার ইবনে শাইখ পুরস্কার","distractor3": "মুনীর চৌধুরী পুরস্কার"},	**8.06285**	**8.08762**	বিকল্পগুলো মোটামুটি অর্থবহ, তবে অধিক বৈচিত্র্য প্রয়োজন ছিল।	**31.90%**	**3.86**	
			{"question": "বাহি পির নাটকটির মূল চরিত্র কে?","correctAns": "পির","distractor1": "মুনীর চৌধুরী","distractor2": "সৈয়দ ওয়ালীউল্লাহ","distractor3": "নুরুল মোমেন"},	**6.32851**	**8.19243**	ডিস্ট্র্যাক্টর যথেষ্ট অর্থঘন, তবে কিছু নাম মাত্র অপশন আছে।	**57.80%**	**14.22**	
			{"question": "বাংলাদেশের আধুনিক নাটকের ধারার প্রধান নাট্যকারদের মধ্যে কে অন্তর্ভুক্ত?"," correctAns": "উপরের সবাই","distractor1": "আসকার ইবনে শাইখ","distractor2": "সৈয়দ ওয়ালীউল্লাহ","distractor3": "মুনীর চৌধুরী"},	**5.40369**	**8.23857**	ডিস্ট্র্যাক্টরগুলো যথেষ্ট ভারসাম্যপূর্ণ; উচ্চ মানের এমসিকিউ।	**67.17%**	**3.86**	
		"বহিপীর নাটকে বহিপীর এবং তাহেরার সম্পর্ক"	{"question": "নাটকটির কেন্দ্রীয় চরিত্র কে?"," correctAns": "বহিপীর","distractor1": "তাহেরা","distractor2": "হাশেম আলি"," distractor3": "হকিকুল্লাহ"},	**7.82333**	**8.17645**	সব বিকল্পই যথেষ্ট প্রাসঙ্গিক; পরীক্ষার উপযোগী।	**12.34%**	**4.21**	
			{"question": "বহিপীর কত বছর পর শিষ্যদের বাড়িতে ঘুরে বেড়ান?","correctAns": "২ বছর","distractor1": "১ বছর"," distractor2": "৩ বছর","distractor3": "৪ বছর"},	**5.30083**	**7.98342**	কিছু ডিস্ট্র্যাক্টর তুলনামূলক দূরবর্তী; মাঝারি মানের এমসিকিউ।	**33.17%**	**29.06**	

			{"question": "তাহেরা কেন বহিপীরের সাথে জোর করে বিরাট দেয়?","correctAns": "তার মাতৃহারা হওয়ার কারণে","distractor1": "তার স্বামীর অনুরোধে","distractor2": "তার স্বামীর মৃত্যুর কারণে","distractor3": "তার শিক্ষার জন্য"},	9.38117	8.15476	ডিস্ট্র্যাক্টর মোটামুটি অর্থবহ, কিন্তু আরও ঘনিষ্ঠ হলে ভালো হতো।	26.52%	11.65	3.68
			{"question": "তাহেরা কোথায় আশ্রয় গ্রহণ করে?","correctAns": "হাশেম আলির জমিদার বজরাতে","distractor1": "বহিপীরের কাছে"," distractor2": "তার পিতার কাছে"," distractor3": "বহিপীরের সঙ্গী হকিকুল্লাহের কাছে"},	6.83373	8.19357	ডিস্ট্র্যাক্টর ভালোই মিল রয়েছে, তবে আরও ঘনিষ্ঠতা প্রয়োজন ছিল।	16.23%	6.59	
			{"question": "বহিপীরের নৌকা কোথায় দুর্ঘটনা হয়?","correctAns": "তাহেরার পালানোর পথে","distractor1": "হাশেম আলির বজরাতে","distractor2": "জমিদারের সমুদ্রে"," distractor3": "তার শিষ্যদের কাছে"},	8.12945	8.24315	ডিস্ট্র্যাক্টরগুলোর সাথে সঠিক উত্তরের অর্থঘনতা যথেষ্ট; ভাল মানের।	23.54%	19.65	
			{"question": "বহিপীর জানতে পায় যে তার নববিবাহিত স্ত্রী কোথায় আছে?"," correctAns": "হাশেম আলির বজরাতে"," distractor1": "তাহেরার কাছে"," distractor2": "তার প্রাক্তন শিষ্যদের কাছে"," distractor3": "তার সঙ্গী হকিকুল্লাহের কাছে"},	7.19312	8.17893	ডিস্ট্র্যাক্টর মোটামুটি অর্থবহ; পরীক্ষামূলকভাবে উপযুক্ত।	28.26%	13.16	
			{"question": "হাশেম আলি কেন তাহেরার পক্ষ নেয়?","correctAns": "তাহেরার করুণ কাহিনী জানার পর","distractor1": "তার বিয়ের জন্য","distractor2": "বহিপীরের প্রতি ঘৃণায়"," distractor3": "জমিদারের অনুরোধে"},	8.68854	8.16247	ডিস্ট্র্যাক্টর ভালোভাবে সন্নিবেশিত; বেশ উপযুক্ত।	29.17%	13.66	
			{"question": "বহিপীর কূটকৌশলের সুযোগ গ্রহণ করে কোন পরিস্থিতির সুযোগ নিয়েছেন?"," correctAns": "জমিদারের অসহিষ্টতার সুযোগে","distractor1": "তাহেরার পালানোর সময়","distractor2": "হাশেম আলির সমর্থন গ্রহণের সুযোগে","distractor3": "তার স্ত্রীর সাথে বিবাহের সুযোগে"},	8.7542	8.13652	তথ্যঘন ডিস্ট্র্যাক্টর, তবে আরও ঘনিষ্ঠ হলে ভালো হতো।	25.57%	7.59	
			{"question": "তাহেরা ও হাশেম আলি কীভাবে বাধার জাল ছিন্ন করে পালিয়ে যায়?"," correctAns": "তাদের বিদ্রোহের মাধ্যমে"," distractor1": "জমিদারের সহায়তায়"," distractor2": "বহিপীরের দুর্ঘটনার সুযোগে"," distractor3": "তাদের স্বামী-স্ত্রী হওয়ার সুযোগে"},	8.51027	8.09231	কিছু ডিস্ট্র্যাক্টর যথেষ্ট দূরবর্তী, তবে বাকিগুলো অর্থবহ।	26.50%	6.64	
			{"question": "নাটকটির মূল থিম কী?"," correctAns": "স্বার্থপরতা ও বিদ্রোহ"," distractor1": "প্রেম ও বিয়ে","distractor2": "সমাজের বিদ্রোহ","distractor3": "ধর্ম ও সামাজিক পরিবর্তন"},	6.91117	8.21974	ডিস্ট্র্যাক্টরগুলো বেশ অর্থবহ, উচ্চ মানের এমসিকিউ।	11.56%	18.25	
	Gemma 3: 12B	"কাকতাড়ুয়া উপন্যাসে বুধার দেশপ্রেম"	{"question": "কাক তাড়ুয়া উপন্যাসের কেন্দ্রীয় চরিত্র বুধার প্রধান বৈশিষ্ট্য কী?"," correctAns": "অসীম সাহস ও মানবিক গুণাবলী","distractor1": "দুর্বলতা এবং ভীরুতা","distractor2": "ধনী পরিবারের সন্তান হওয়া","distractor3": "পড়াশোনায় অত্যন্ত মেধাবী হওয়া"},	7.88579	8.33925	সঠিক উত্তরের সঙ্গে অন্য বিকল্পগুলো যথেষ্ট দূরত্বে থাকলেও তারা প্রসঙ্গের বাইরে নয়, তাই মাঝারি মান।	62.61%	5.13	
			{"question": "বুধার মধ্যে মুক্তির বোধ প্রথম জাগে কখন?","correctAns": "চাচির বাড়ি ছেড়ে আসার পরে","distractor1": "বাবা-মায়ের সাথে দেখা করার পরে","distractor2": "গ্রামের স্কুলে ভর্তি হওয়ার পরে","distractor3": "মুক্তিযোদ্ধাদের সাথে পরিচিত হওয়ার পরে"},	9.53274	8.69853	সঠিক উত্তরের সঙ্গে অন্য বিকল্পগুলো যথেষ্ট দূরত্বে থাকলেও তারা প্রসঙ্গের বাইরে নয়, তাই মাঝারি মান।	54.23%	11.41	
			{"question": "বুধাকে 'পাগল' বলে ডাকে -"," correctAns": "গ্রামের লোক","distractor1": "তার বাবা","distractor2": "তার মা"," distractor3": "চাচি"},	6.02898	7.54979	সব distractor বাস্তবিক, তবে সঠিক উত্তরের দেশপ্রেমিক দৃষ্টিকোণ সরাসরি নয়, তাই সামান্য কম নম্বর।	53.71%	12.46	

			{"question": "কাক তাড়ুয়া উপন্যাসে বুধার সাহসীকতাকে কিভাবে ফুটিয়ে তোলা হয়েছে?","correctAns": "বুধার নির্ভীক কার্যকলাপের মাধ্যমে","distractor1": "বুধার দুর্বলতা দেখিয়ে","distractor2": "বুধার ভীরুতা প্রকাশ করে","distractor3": "বুধার কান্নার দৃশ্য বর্ণনার মাধ্যমে"},	**7.96728**	**6.96453**	দ্বিতীয় ও তৃতীয় distractor কিছুটা দূরে সরে গেলেও প্রথমটি যথেষ্ট সম্পর্কযুক্ত। সামগ্রিক মান মাঝারি উচ্চ।	**69.01%**	**8.06**	**2.54**
			{"question": "বুধার জীবনে বঙ্গবন্ধু শেখ মুজিবুর রহমানের প্রভাব কী ছিল?","correctAns": "বঙ্গবন্ধুর স্বাধীনতার আহ্বান বুধাকে অনুপ্রাণিত করত","distractor1": "বঙ্গবন্ধু বুধাকে সরাসরি প্রশিক্ষণ দিয়েছিলেন","distractor2": "বঙ্গবন্ধু বুধার পরিবারের সদস্য ছিলেন","distractor3": "বুধা বঙ্গবন্ধুকে কখনো দেখেনি"},	**8.90334**	**7.77204**	বিকল্পগুলো আংশিক প্রাসঙ্গিক হলেও সঠিক উত্তরের গভীর মানসিক দৃষ্টিভঙ্গির মতো বাস্তবিক নয়। তাই নিম্ন-মধ্য মানের এমসিকিউ।	**56.78%**	**3.83**	
			{"question": "কাক তাড়ুয়া উপন্যাসের ভাষাভঙ্গি কেমন?","correctAns": "ছোট ছোট বাক্য এবং অন্তরঙ্গ","distractor1": "জটিল এবং কঠিন","distractor2": "কাব্যিক এবং অলঙ্কৃত","distractor3": "দীর্ঘ এবং বর্ণনাত্মক"},	**7.73094**	**7.98761**	সব উত্তরই প্রাসঙ্গিক এবং বাস্তবিক, তবে সঠিক উত্তর অনেকটাই নির্দিষ্ট, তাই মাঝারি মান।	**40.03%**	**11.11**	
			{"question": "উপন্যাসে নোলক বুয়ার ভূমিকা কী?","correctAns": "সে বুধার সহায়ক এবং বন্ধুস্থানীয়","distractor1": "সে প্রধান খলনায়ক","distractor2": "সে বুধার শিক্ষিকা","distractor3": "সে গ্রামের মাতব্বর"},	**6.07358**	**8.41258**	তিনটি distractor-ই যথেষ্ট বাস্তবিক ও কাছাকাছি, সঠিক উত্তর থেকে কিছুটা দূরে থাকলেও যুক্তিযুক্ত।	**59.40%**	**13.08**	
			{"question": "কাক তাড়ুয়া উপন্যাসের প্রেক্ষাপট কী?","correctAns": "পাকিস্তান আমলের মুক্তিযুদ্ধ","distractor1": "উনিশ শতকের গ্রাম্য জীবন","distractor2": "বর্তমান সময়ের শহরের জীবন","distractor3": "প্রাচীন বাংলার সমাজ"},	**8.66734**	**8.02017**	Distractor-গুলো প্রাসঙ্গিক হলেও সঠিক উত্তর খুব স্পষ্টভাবে আলাদা, তাই তুলনামূলক সহজ।	**39.51%**	**8.19**	
			{"question": ""ঘুমুবার জায়গা নেই তো কী হয়েছে ? দোকানের বারান্দা আছে..." - এই উক্তিটি কার মানসিকতাকে প্রতিফলিত করে?","correctAns": "বুধা","distractor1": "কুন্তি","distractor2": "নোলক বুয়া","distractor3": "আহাদ মুন্সি"},	**8.1986**	**6.87544**	বিকল্পগুলো আংশিক সত্য হলেও মূল চরিত্রের পরিচয় নির্ধারণে সঠিক উত্তর খুব সহজে চেনা যায়।	**21.47%**	**23.04**	
			{"question": "সেলিনা হোসেন কীভাবে কাক তাড়ুয়া উপন্যাসের কাহিনীকে বিশ্বাসযোগ্য করে তুলেছেন?","correctAns": "বিভিন্ন চরিত্রের সাহায্য নিয়ে এবং বাস্তবসম্মত বর্ণনা দিয়ে","distractor1": "জটিল ভাষা ব্যবহার করে","distractor2": "অত্যধিক অলঙ্করণ ব্যবহার করে","distractor3": "শুধুমাত্র বুধার দৃষ্টিকোণ থেকে বর্ণনা করে"}	**8.28744**	**7.64103**	বর্ণনা যথাযথ হলেও distractor-গুলো যথেষ্ট জোরালো নয়, ফলে তুলনামূলক দুর্বল প্রশ্ন।	**52.39%**	**4.33**	
		"সৈয়দ ওয়ালীউল্লাহ এর লেখক পরিচিতি"	{"question": "সৈয়দ ওয়ালীউল্লাহ কত সালে জন্মগ্রহণ করেন?","correctAns": "১৯২০ সালে, চট্টগ্রামে","distractor1": "১৯২৫ সালে, চট্টগ্রামে","distractor2": "১৯২২ সালে, ঢাকায়","distractor3": "১৯২৩ সালে, কুড়িগ্রামে"},	**6.92782**	**8.29438**	তথ্যভিত্তিক প্রশ্ন হলেও distractor-গুলো কাছাকাছি হওয়ায় উত্তর দেওয়া কিছুটা চ্যালেঞ্জিং।	**36.12%**	**42.38**	
			{"question": "সৈয়দ ওয়ালীউল্লাহর পিতার নাম কী ছিল?","correctAns": "আহমাদউল্লাহ","distractor1": "নাসিম আরা খাতুন","distractor2": "মুনীর চৌধুরী","distractor3": "আসকার ইবনে শাইখ"},	**8.88475**	**7.45287**	Distractor-গুলোর মধ্যে একটি খুব দুর্বল হলেও বাকি দুটি তুলনামূলক ভালো, তাই মাঝারি স্কোর।	**40.32%**	**33.07**	
			{"question": "'নয়নচারা' কোন সালে প্রকাশিত হয়?","correctAns": "১৯৪৫ সালে","distractor1": "১৯৪৯ সালে","distractor2": "১৯৫৫ সালে","distractor3": "১৯৬০ সালে"},	**6.62345**	**7.93428**	সাল সম্পর্কিত প্রশ্নে distractor-গুলো কাছাকাছি এবং বিভ্রান্তিকর, তাই ভাল মানের প্রশ্ন।	**21.33%**	**44.5**	

			{"question": "'বহিপীর' নাটকটি প্রথম কোন সালে প্রকাশিত হয়?","correctAns": "১৯৬০ সালে","distractor1": "১৯৫৫ সালে"," distractor2": "১৯৪৯ সালে","distractor3": "১৯৬৪ সালে"},	9.74634	7.62351	মূল তথ্যটি সাধারণ হলেও distractor যথাযথভাবে সজ্জিত, তাই গড় মান।	66.77%	43.14	
			{"question": "'বহিপীর' নাটকের কেন্দ্রীয় চরিত্রটির নামের তাৎপর্য কী?","correctAns": "পীর সম্প্রদায়ের কুসংস্কার ও ধর্মীয় বইয়ের পাতার প্রভাব","distractor1": "নাটকের লেখকের নামের প্রতিচ্ছবি","distractor2": "নাটকের কাহিনির জটিলতা","distractor3": "নাটকের বিষয়বস্তুর আধুনিকতা"},	9.57216	8.03847	প্রসঙ্গভিত্তিক এবং সাংস্কৃতিকভাবে প্রাসঙ্গিক, তবে বিকল্পগুলো আরো শক্তিশালী হতে পারতো।	63.49%	19.7	
			{"question": "ওয়ালীউল্লাহ কীভাবে বাঙালি মুসলমান সমাজে পীর প্রথা কাছ থেকে দেখার সুযোগ পান?","correctAns": "তার পিতার বদলির চাকরি সূত্রে তিনি সারা বাংলাদেশে ঘুরে বেড়ানোর সুযোগ পান","distractor1": "তিনি নিজেই একজন পীর ছিলেন","distractor2": "তিনি গ্রামের স্কুলে শিক্ষকতা করতেন"," distractor3": "তিনি বিদেশে পড়াশোনা করতে গিয়েছিলেন"},	5.00505	7.8516	বহু অভিজ্ঞতার ভিত্তিতে প্রশ্নটি যুক্তিসম্মত, distractor-ও বাস্তবধর্মী।	38.33%	7.26	3.27
			{"question": "ভারত বিভাগের আগে বাংলা নাট্যচর্চার প্রাণকেন্দ্র কোথায় ছিল?"," correctAns": "কলকাতা","distractor1": "ঢাকা","distractor2": "কুড়িগ্রাম"," distractor3": "প্যারিস"},	6.77665	8.01673	বিষয়ভিত্তিক নির্ভুল প্রশ্ন, বিকল্পগুলো যথেষ্ট প্রাসঙ্গিক।	58.99%	7.73	
			{"question": "নুরুল মোমেন রচিত উল্লেখযোগ্য নাটক কোনটি?","correctAns": "নয়া খান্দান","distractor1": "কাঁদো নদী কাঁদো"," distractor2": "তরঙ্গভঙ্গ","distractor3": "চিঠি"},	8.09136	7.74019	নাট্যপ্রসঙ্গ হওয়ায় তথ্য নির্ভর প্রশ্নটি যুক্তিযুক্ত এবং প্রাসঙ্গিক distractor-সহ।	57.91%	6.11	
			{"question": "আসকার ইবনে শাইখের কোন নাটকটি বিদ্রোহের চিত্র তুলে ধরে?"," correctAns": "অগ্নিগিরি","distractor1": "এমন যদি হতো","distractor2": "কবর"," distractor3": "লালসালু"},	8.92905	8.23114	প্রসঙ্গ ও উত্তর সবই ঐতিহাসিকভাবে বাস্তবভিত্তিক, distractor-ও যথাযথ।	44.11%	8.07	
			{"question": "মুনীর চৌধুরী রচিত কোন নাটকটি ১৯৬২ সালে প্রকাশিত হয়?"," correctAns": "রক্তাক্ত প্রান্তর","distractor1": "কবর","distractor2": "চিঠি","distractor3": "তরঙ্গভঙ্গ"}	7.98152	7.98325	সঠিক উত্তর অন্যদের তুলনায় কিছুটা এগিয়ে থাকলেও distractor-গুলো যথেষ্ট কাছাকাছি।	43.78%	9.88	
		"বহিপীর নাটকে বহিপীর এবং তাহেরার সম্পর্ক"	{"question": "নাটকের মূল কাহিনীটি কিসের উপর ভিত্তি করে নির্মিত?","correctAns": "বহিপীরের সর্বগ্রাসী স্বার্থ ও এক বালিকার বিদ্রোহ","distractor1": "এক প্রেমিকের বিরহবেদনা","distractor2": "জমিদারের রাজনৈতিক ষড়যন্ত্র","distractor3": "একটি প্রাকৃতিক দুর্যোগ"},	6.83724	8.58627	মূল থিম স্পষ্ট, তবে distractor-গুলো কিছুটা দুর্বল হওয়ায় সামান্য প্রভাব ফেলেছে।	11.60%	3.78	
			{"question": "বহিপীর কী করতেন দুই বছর পর?","correctAns": "মুরিদদের বাড়িতে ঘুরে বেড়ান এবং সেবা গ্রহণ করতেন"," distractor1": "গ্রামের স্কুলে শিক্ষকতা করতেন","distractor2": "বাণিজ্য করতেন"," distractor3": "গান শুনতেন ও কবিতা লিখতেন"},	5.51316	8.66429	উত্তরগুলো এমনভাবে সাজানো হয়েছে যেন প্রতিটি যুক্তিযুক্ত মনে হয়, তাই উচ্চ মানের প্রশ্ন।	31.93%	87.58	
			{"question": "তাহেরা কার সাথে জোর করে বিয়ে দেওয়া হয়েছিল?","correctAns": "বহিপীর","distractor1": "হাতেম আলি জমিদারের পুত্র","distractor2": "হাশেম আলি জমিদারের পুত্র","distractor3": "অন্য একজন মুরিদ"},	6.39012	7.72048	সঠিক উত্তর খুব বেশি আলাদা নয়, তাই উত্তর নির্ধারণ তুলনামূলক কঠিন।	23.06%	47.26	
			{"question": "তাহেরা কোথায় আশ্রয় গ্রহণ করেছিল?","correctAns": "হাতেম আলি জমিদারের বজরায়","distractor1": "জঙ্গলে"," distractor2": "একটি গ্রামের বাড়িতে"," distractor3": "নদীর ধারে"},	8.38991	8.09874	প্রাসঙ্গিকতা ঠিক থাকলেও আরো সূক্ষ্ম ভাষা ব্যবহার করলে ভালো হতো।	16.87%	37.01	

			{"question": "বহিপীরকে খুঁজতে কে তাঁর সাথে ছিল?","correctAns": "হকিকুল্লাহ"," distractor1": "তাহেরা","distractor2": "হাতেম আলি","distractor3": "হাশেম আলি"},	8.92581	7.49105	Distractor-গুলো কিছুটা পুনরাবৃত্ত হলেও মূল থিমে ঠিক আছে।	24.22%	32.38	3.2
			{"question": "বহিপীর কিভাবে তাহেরার কাছে পৌঁছান?","correctAns": "দুর্ঘটনার ফলে হাতেম আলির বজরায় আশ্রয় নিয়ে"," distractor1": "হেঁটে","distractor2": "ঘোড়ায় চেপে","distractor3": "উড়োজাহাজে করে"},	7.78003	8.6702	দুর্ঘটনার উপরে নির্মিত প্রশ্নটি বাস্তবধর্মী এবং চমৎকারভাবে বিভ্রান্তিকর।	34.17%	15.49	
			{"question": "বজরায় জমিদারের পুত্র কার পক্ষ নেয়?","correctAns": "তাহেরা"," distractor1": "বহিপীর","distractor2": "হকিকুল্লাহ","distractor3": "অন্য একজন মুরিদ"},	8.63596	7.65473	প্রসঙ্গ এবং চরিত্রের অবস্থান সুনির্দিষ্ট হলেও distractor তুলনামূলক দুর্বল।	34.34%	41.38	
			{"question": "নাটকের শেষে বহিপীর কি মেনে নিতে বাধ্য হন?","correctAns": "তাঁর পরাজয় এবং বাস্তব পরিস্থিতি","distractor1": "তাহেরাকে ফিরে পাওয়া","distractor2": "জমিদারের সাথে জোট বাঁধা","distractor3": "মুরিদদের সমর্থন"},	7.22613	8.21569	উপসংহার ভিত্তিক প্রশ্নে উত্তরের যৌক্তিকতা বজায় রয়েছে, এবং distractor যথেষ্ট বাস্তবধর্মী।	29.60%	11.96	
			{"question": "বহিপীর কিভাবে তাঁর উদ্দেশ্য হাসিল করার চেষ্টা করতেন?","correctAns": "জঘন্য কূটকৌশলের আশ্রয় গ্রহণ করে"," distractor1": "সৎভাবে কাজ করে"," distractor2": "সরাসরি যুদ্ধ করে"," distractor3": "দয়া করে অনুরোধ করে"},	8.81547	8.7736	সুপরিকল্পিত কূটকৌশলের উপর ভিত্তি করে গঠিত, উত্তম distractor সহ।	33.99%	6.87	
			{"question": "নাটকের চূড়ান্ত পরিণতিতে তাহেরা ও হাশেম আলি কী করেন?"," correctAns": "সব বাধার জাল ছিন্ন করে পালিয়ে যান","distractor1": "বহিপীরের কাছে আত্মসমর্পণ করেন","distractor2": "জমিদারের সাহায্য চান","distractor3": "মুরিদদের কাছে সাহায্য প্রার্থনা করেন"}	7.91566	8.41387	উপসংহার উপযুক্ত এবং ভাবনার উদ্রেক করে এমন, তাই উপরের স্তরের স্কোর।	35.88%	18.53	
Bangla Shahitto	Gemini 2.0 Flash	"নিরীহ বাঙালি গল্পে বাঙালির সমালোচনা"	{"question": "নিরীহ বাঙালি গল্পে লেখিকা বাঙালিকে কিসের সাথে তুলনা করেছেন?"," correctAns": "মূর্তিমতী কবিতা"," distractor1": "সাহসী যোদ্ধা","distractor2": "পরিশ্রমী কৃষক","distractor3": "সফল ব্যবসায়ী"},	3.76558	8.24715	ডিস্ট্র্যাক্টরগুলো অর্থঘন এবং যথেষ্ট প্রাসঙ্গিক; উচ্চ মানের এমসিকিউ।	49.39%	4.23	
			{"question": "লেখিকার মতে বাঙালি নারীরা কেমন শাড়ি পরিধান করেন?","correctAns": "সূক্ষ্ম 'হাওয়ার শাড়ি'","distractor1": "মোটা কাপড়ের শাড়ি","distractor2": "তাঁতের শাড়ি","distractor3": "খাদি শাড়ি"},	9.70415	8.19846	তথ্যঘন ডিস্ট্র্যাক্টর, তবে কিছু অপশন তুলনামূলক দূরবর্তী; মোটামুটি উচ্চ মানের।	44.69%	11.28	
			{"question": "বাঙালি পুরুষদের প্রধান ব্যবসায় কী?","correctAns": "বাণিজ্য"," distractor1": "কৃষি কাজ","distractor2": "পশুপালন","distractor3": "চাকরি"},	6.33752	8.13675	ডিস্ট্র্যাক্টর মোটামুটি অর্থবহ, কিছু সরলীকরণ রয়েছে; ভাল মানের।	35.75%	13.13	
			{"question": "লেখিকার মতে, বাণিজ্যকে বাঙালিরা কীভাবে সহজসাধ্য করে তুলেছে?"," correctAns": "বিলাসদ্রব্যের ব্যবসা করে"," distractor1": "কঠিন পরিশ্রমের মাধ্যমে"," distractor2": "আধুনিক প্রযুক্তি ব্যবহারের মাধ্যমে","distractor3": "বৈদেশিক বাণিজ্যের মাধ্যমে"},	8.59845	8.23957	বিকল্পগুলো যথেষ্ট যুক্তিযুক্ত ও ঘনিষ্ঠ; উচ্চ মানের এমসিকিউ।	39.11%	5.02	
			{"question": "নিরীহ বাঙালি গল্পে 'পাস বিক্রয়' বলতে কী বোঝানো হয়েছে?","correctAns": "বিবাহের মাধ্যমে অর্থ উপার্জন","distractor1": "শিক্ষা প্রতিষ্ঠানের সার্টিফিকেট বিক্রি"," distractor2": "বর কর্তৃক শ্বশুরকে অর্থ উপার্জনের পথ দেখানো","distractor3": "জমি বিক্রির দলিল তৈরি"},	8.63444	8.17684	তথ্যভিত্তিক ও প্রাসঙ্গিক ডিস্ট্র্যাক্টর; মোটামুটি ভাল মানের।	49.37%	18.85	2.05
			{"question": "লেখিকার মতে, রাজ্য স্থাপন করা অপেক্ষা কোন উপাধি লাভ করা সহজ?"," correctAns": "রাজা","distractor1": "সেনাপতি","distractor2": "জমিদার"," distractor3": "মন্ত্রী"},	8.10884	8.21593	ডিস্ট্র্যাক্টরগুলো বেশ অর্থঘন; পরীক্ষার জন্য উপযুক্ত।	38.41%	11.89	

			{"question": "স্বাস্থ্যরক্ষায় যত্নবান হওয়া অপেক্ষা বাঙালিরা কোন কাজটিকে সহজ মনে করে?","correctAns": "স্বাস্থ্য নষ্ট করে ঔষধ ও ডাক্তারের শরণাপন্ন হওয়া","distractor1": "নিয়মিত ব্যায়াম করা","distractor2": "পুষ্টিকর খাবার খাওয়া","distractor3": "পরিষ্কার পরিচ্ছন্ন থাকা"},	7.89103	8.13429	তথ্যঘন বিকল্প, কিছুটা বৈচিত্র্যের অভাব; ভাল মানের।	36.91%	18.98	
			{"question": "লেখিকার মতে, গৃহিণীদের রন্ধন করতে বলা হলে তাদের কী হওয়া উচিত?","correctAns": "ত্রিবিধ দণ্ড দেওয়া","distractor1": "পুরস্কৃত করা","distractor2": "উৎসাহিত করা","distractor3": "তিরস্কার করা"},	8.23739	8.24938	বিকল্পগুলো ঘনিষ্ঠ এবং যথেষ্ট যুক্তিযুক্ত; উচ্চ মানের এমসিকিউ।	52.63%	19.74	
			{"question": "কৃষি বিভাগের কাজ (agriculture) করা অপেক্ষা লেখিকা কোনটিকে সহজ বলেছেন?","correctAns": "মস্তিষ্ক উর্বর (brain culture) করা","distractor1": "জমিতে সার দেওয়া","distractor2": "বীজ বপন করা","distractor3": "ফসল কাটা"},	5.52429	8.19572	ডিস্ট্র্যাক্টর অর্থঘন, তবে সামান্য সরলীকরণ রয়েছে।	61.58%	29.99	
			{"question": ""অতি শুভ্রনীলাম্বর" – এই শব্দটি দিয়ে কী বোঝানো হয়েছে?","correctAns": "পরিষ্কার নীল আকাশ","distractor1": "মেঘাচ্ছন্ন আকাশ","distractor2": "রক্তিম আকাশ","distractor3": "গোধূলির আকাশ"}	8.34118	8.21247	তথ্যঘন এবং উপযুক্ত ডিস্ট্র্যাক্টর; উচ্চ মানের এমসিকিউ।	46.75%	172.56	
		"মমতাদি গল্পে ছেলেটির সাথে মমতাদির সম্পর্ক"	{"question": "মমতাদি প্রথমে ছেলেটির মায়ের কাছে কি কাজের জন্য আবেদন জানিয়েছিল?","correctAns": "রান্নার কাজ","distractor1": "বাসন মাজার কাজ","distractor2": "কাপড় কাচার কাজ","distractor3": "ঘর মোছার কাজ"},	9.53222	8.17642	বিকল্পগুলো মোটামুটি প্রাসঙ্গিক; ভাল মানের।	46.56%	4.82	1.02
			{"question": "মমতাদির স্বামীর চাকরি না থাকার কারণে সংসারে কি প্রভাব পড়েছিল?","correctAns": "সংসারে অভাব দেখা দেয়","distractor1": "ছেলেটির লেখাপড়া বন্ধ হয়ে যায়","distractor2": "তাদের বাড়িটি বিক্রি করতে হয়","distractor3": "মমতাদির স্বাস্থ্য খারাপ হয়ে যায়"},	7.03162	8.21357	ডিস্ট্র্যাক্টরগুলো যথেষ্ট অর্থবহ এবং ঘনিষ্ঠ; ভাল মানের।	45.64%	4.27	
			{"question": "মমতাদি ছেলেটিকে প্রথম দিন কী খেতে দিয়েছিল?","correctAns": "কমলালেবু","distractor1": "সন্দেশ","distractor2": "মুড়ি ও বাতাসা","distractor3": "বিস্কুট"},	8.46413	8.12458	তথ্যভিত্তিক বিকল্প, তবে আরও ঘনিষ্ঠ হলে ভালো হতো।	50.35%	6.84	
			{"question": "মমতাদি ছেলেটিকে নিজের বাড়িতে ডেকে নিয়ে যাওয়ার সময় কোন গলির ভেতর দিয়ে গিয়েছিল?","correctAns": "জীবনময় লেন","distractor1": "আনন্দময় লেন","distractor2": "হাসিমুখ গলি","distractor3": "সুখের সড়ক"},	8.8066	8.19834	ডিস্ট্র্যাক্টর অর্থঘন; মোটামুটি উপযুক্ত এমসিকিউ।	50.39%	4.3	
			{"question": "ছেলেটি মমতাদির কোন ব্যবহার দেখে প্রথমে ক্ষুন্ন হয়েছিল?","correctAns": "মমতাদির রূঢ় ব্যবহার","distractor1": "মমতাদির কথা না শোনা","distractor2": "মমতাদির রান্না খারাপ হওয়া","distractor3": "মমতাদির দেরিতে আসা"},	8.35311	8.14326	কিছু ডিস্ট্র্যাক্টর তুলনামূলক দূরবর্তী; ভাল মানের।	49.31%	7.05	
			{"question": "মমতাদির ছেলে ঘুমিয়ে থাকার সময় কোথায় শুয়ে ছিল?","correctAns": "চৌকির উপর মাদুরে","distractor1": "খাটের উপর","distractor2": "মেঝেতে","distractor3": "বারান্দায়"},	9.58261	8.23145	বিকল্পগুলো ঘনিষ্ঠ অর্থসম্পন্ন; উচ্চ মানের এমসিকিউ।	42.50%	5.86	
			{"question": "মমতাদি ছেলেটিকে কী বলে ডাকত?","correctAns": "খোকা","distractor1": "ভাই","distractor2": "বাচ্চা","distractor3": "ছোটবাবু"},	8.15287	8.17893	তথ্যঘন ডিস্ট্র্যাক্টর; ভাল মানের এমসিকিউ।	47.37%	11.08	
			{"question": "মমতাদি যখন প্রথম দিন কাজে আসে, তখন তার পরনের শাড়িটির পাড় কেমন ছিল?","correctAns": "বিবর্ণ লাল","distractor1": "সোনালী","distractor2": "রুপালী","distractor3": "সবুজ"},	8.24834	8.19675	ডিস্ট্র্যাক্টর অর্থঘন, তবে সামান্য অধিক সরলীকৃত।	42.38%	27.53	

			{"question": "ছেলেটি মমতাদির কোন বিষয়টি নিয়ে রহস্য অনুভব করেছিল?","correctAns": "তার অতীত জীবনের কষ্ট","distractor1": "তার চুপচাপ স্বভাব","distractor2": "তার হাতের কাজ","distractor3": "তার চুল বাঁধার ধরন"},	**8.36265**	**8.21894**	তথ্যঘন বিকল্প এবং যথেষ্ট ঘনিষ্ঠ; উচ্চ মানের।	**47.59%**	**30.87**	
			{"question": "মমতাদির বরের চাকরি হলে ছেলেটি তাকে কী খাওয়ানোর প্রস্তাব দিয়েছিল?","correctAns": "সন্দেশ","distractor1": "মিষ্টি","distractor2": "ফল","distractor3": "হরির লুট"}	**8.51807**	**8.18753**	বিকল্পগুলো প্রাসঙ্গিক এবং পরীক্ষার জন্য উপযুক্ত।	**61.70%**	**20.19**	
		"অন্ধবধূ কবিতার কবির পরিচয়"	{"question": "যতীন্দ্রমোহন বাগচী কত সালে জন্মগ্রহণ করেন?","correctAns": "১৮৭৮","distractor1": "১৮৭৬","distractor2": "১৮৮০","distractor3": "১৮৭২"},	**6.76209**	**8.23671**	ডিস্ট্র্যাক্টর যথেষ্ট অর্থঘন ও ঘনিষ্ঠ; ভাল মানের।	**31.09%**	**6.94**	**3.5**
			{"question": "যতীন্দ্রমোহন বাগচী কোন জেলার জামশেদপুর গ্রামে জন্মগ্রহণ করেন?","correctAns": "নদীয়া","distractor1": "যশোর","distractor2": "ফরিদপুর","distractor3": "খুলনা"},	**8.54467**	**8.14358**	তথ্যভিত্তিক বিকল্প, কিছু অপশন দূরবর্তী।	**32.67%**	**4.9**	
			{"question": "নিচের কোন বৈশিষ্ট্যটি যতীন্দ্রমোহন বাগচীর কবিমানসের প্রধান দিক?","correctAns": "পল্লি-প্রীতি","distractor1": "নগর-প্রীতি","distractor2": "ইতিহাস-প্রীতি","distractor3": "প্রকৃতি-বিমুখতা"},	**9.31503**	**8.21475**	ডিস্ট্র্যাক্টর অর্থঘন ও উপযুক্ত; ভাল মানের এমসিকিউ।	**42.78%**	**7.99**	
			{"question": "যতীন্দ্রমোহন বাগচীর কাব্যের ভাষা কেমন?","correctAns": "সহজ ও সরল","distractor1": "জটিল ও দুর্বোধ্য","distractor2": "অলংকার বহুল","distractor3": "কঠিন ও তৎসম শব্দবহুল"},	**6.85016**	**8.19583**	বিকল্পগুলো মোটামুটি ঘনিষ্ঠ; ভাল মানের।	**45.22%**	**6.21**	
			{"question": "'অন্ধবধূ' কবিতাটি কোন কবির রচনা?","correctAns": "যতীন্দ্রমোহন বাগচী","distractor1": "বিভূতিভূষণ বন্দ্যোপাধ্যায়","distractor2": "জীবনানন্দ দাশ","distractor3": "কাজী নজরুল ইসলাম"},	**9.00121**	**8.12947**	কিছু ডিস্ট্র্যাক্টর দূরবর্তী হলেও বাকিগুলো অর্থবহ।	**19.27%**	**34.03**	
			{"question": "যতীন্দ্রমোহন বাগচীর কবিতায় কোন রূপটি বিশেষভাবে উন্মোচিত হয়েছে?","correctAns": "গ্রামবাংলার শ্যামল স্নিগ্ধ রূপ","distractor1": "শহরের আধুনিক রূপ","distractor2": "মরুভূমির রুক্ষ রূপ","distractor3": "পর্বতের কঠিন রূপ"},	**8.80568**	**8.24613**	বিকল্পগুলো যথেষ্ট ঘনিষ্ঠ ও যুক্তিযুক্ত; উচ্চ মানের।	**42.42%**	**15.1**	
			{"question": "নিচের কোনটি যতীন্দ্রমোহন বাগচীর একটি কাব্যগ্রন্থ?","correctAns": "অপরাজিতা","distractor1": "শেষের কবিতা","distractor2": "সোনার তরী","distractor3": "অগ্নিবীণা"},	**7.32719**	**8.17625**	তথ্যঘন ডিস্ট্র্যাক্টর; ভাল মানের এমসিকিউ।	**38.91%**	**6.93**	
			{"question": "যতীন্দ্রমোহন বাগচী কত সালে মৃত্যুবরণ করেন?","correctAns": "১৯৪৮","distractor1": "১৯৪৯","distractor2": "১৯৪৭","distractor3": "১৯৫০"},	**6.23094**	**8.21364**	ডিস্ট্র্যাক্টর অর্থঘন; পরীক্ষার জন্য উপযুক্ত।	**29.28%**	**6.65**	
			{"question": "জীবনানন্দ দাশের মতো আর কোন কবির কাব্য চিত্ররূপময়?","correctAns": "যতীন্দ্রমোহন বাগচী","distractor1": "রবীন্দ্রনাথ ঠাকুর","distractor2": "কাজী নজরুল ইসলাম","distractor3": "জসীমউদ্দীন"},	**7.46064**	**8.23941**	ডিস্ট্র্যাক্টর যথেষ্ট ভারসাম্যপূর্ণ; উচ্চ মানের এমসিকিউ।	**32.15%**	**13.06**	
			{"question": "যতীন্দ্রমোহন বাগচী তাঁর কবিতায় কোন বিষয়কে তাৎপর্যমণ্ডিত করে তুলে ধরেছেন?","correctAns": "গ্রাম-জীবনের অতি সাধারণ বিষয় ও সুখ-দুঃখ","distractor1": "রাজনৈতিক আন্দোলন","distractor2": "শহরের চাকচিক্য","distractor3": "বিদেশি সংস্কৃতি"}	**5.31867**	**7.16527**	তথ্যঘন বিকল্প, কিছু বৈচিত্র্যের অভাব রয়েছে।	**43.75%**	**4.67**	

	Qwen 3: 14B	"নিরীহ বাঙালি গল্পে বাঙালির সমালোচনা"	{"question": "লেখক কোন বিষয়ে প্রধানত সমালোচনা করেছেন?","correctAns": "মানুষের আলস্য","distractor1": "শিক্ষার অভাব","distractor2": "দুর্ভিক্ষের প্রভাব"," distractor3": "বৈদেশিক ভিক্ষা গ্রহণ"},	7.95005	8.48672	সঠিক উত্তর একমাত্র গভীর দৃষ্টিভঙ্গির বহিঃপ্রকাশ; distractor-গুলো আংশিক বাস্তবিক হলেও তুলনায় দুর্বল।	34.77%	6.57	1.09
			{"question": "লেখক কৃষি কাজের সাথে কী তুলনা করেছেন?","correctAns": "মস্তিষ্ক উর্বরতা বৃদ্ধি","distractor1": "স্বাস্থ্য রক্ষা"," distractor2": "শিক্ষার মাধ্যমে অর্থ উৎপাদন"," distractor3": "দুর্ভিক্ষ সমাধান"},	7.78014	8.26138	সঠিক ও বিভ্রান্তিকর বিকল্পগুলোর মাঝে সূক্ষ্ম পার্থক্য, ভাল মানের প্রশ্ন।	55.69%	5.85	
			{"question": "লেখক মহিলাদের কাছে কী অপরিহার্য বলে মনে করেন?","correctAns": "আলস্য প্রকাশ","distractor1": "শিক্ষা প্রদান"," distractor2": "রন্ধন করা","distractor3": "সামাজিক কর্ম"},	7.61456	7.69057	'আলস্য প্রকাশ' সঠিক হলেও distractor-গুলো অপ্রত্যাশিতভাবে বাস্তব, কিছুটা বিভ্রান্তিকর।	39.46%	4.28	
			{"question": "লেখক কোন উপাধি গ্রহণের জন্য অর্থ ব্যয় করা বেশি সহজ বলে মনে করেন?"," correctAns": "খাঁ বাহাদুর","distractor1": "মুসলিম লীগ সদস্য","distractor2": "বিদেশী বুদ্ধিজীবী","distractor3": "স্বাস্থ্য বিশেষজ্ঞ"},	7.13504	8.08529	উপাধি বিষয়ক প্রশ্নটি ব্যঙ্গাত্মক হলেও distractor-গুলো যথেষ্ট বাস্তবসম্মত।	48.94%	7.26	
			{"question": "লেখক কী উপায় দ্বারা সৌন্দর্য বৃদ্ধি করা বেশি সহজ বলে মনে করেন?"," correctAns": "মসুলা ও চামড়া প্রস্তুতকারী পণ্য ব্যবহার","distractor1": "স্বাস্থ্য রক্ষা"," distractor2": "পুষ্টিকর খাদ্য গ্রহণ"," distractor3": "শারীরিক পরিশ্রম"},	6.90331	7.96412	প্রশ্নটি ব্যঙ্গরসাত্মক, উত্তর যথাযথ হলেও distractor-গুলো তুলনায় দুর্বল।	51.31%	4.91	
			{"question": "লেখক কোন প্রতিক্রিয়া দ্বারা রন্ধন করা বৈরীদের শাস্তি করতে চান?"," correctAns": "তুষানলে দগ্ধ করা"," distractor1": "শিক্ষাদান","distractor2": "সামাজিক কর্ম করা","distractor3": "বৈদেশিক সহায়তা গ্রহণ"},	7.8046	8.30247	সাহিত্যিক ব্যঙ্গ প্রকাশ পেয়েছে, distractor-গুলো তুলনামূলকভাবে দূরে থাকলেও রস ধরে রেখেছে।	48.53%	7.07	
			{"question": "লেখক কৃষি চাষে কী তুলনা করেছেন?","correctAns": "মুখস্থ বিদ্যার মাধ্যমে অর্থ উৎপাদন","distractor1": "বুদ্ধিমত্তা বৃদ্ধি","distractor2": "দুর্ভিক্ষ নিবারণ","distractor3": "শিক্ষার জন্য অর্থ ব্যয়"},	8.30373	7.80831	প্রশ্নে উপমা যথাযথ, তবে distractor কিছুটা কম প্রাসঙ্গিক ।	54.40%	10.46	
			{"question": "লেখক কোন উপায় দ্বারা সমাজের দুর্ভিক্ষ নিবারণ করা বেশি সহজ বলে মনে করেন?","correctAns": "বৈদেশিক ভিক্ষা গ্রহণ","distractor1": "অন্ন উৎপাদন"," distractor2": "শিক্ষার মাধ্যমে অর্থ উৎপাদন"," distractor3": "স্বাস্থ্য রক্ষা"},	7.73046	7.59126	উত্তরটি ইচ্ছাকৃতভাবে ব্যঙ্গাত্মক, distractor-গুলো তুলনায় যুক্তিভিত্তিক, তাই বিভ্রান্তি হ্রাস পায়।	43.03%	6.05	
			{"question": "লেখক কৃষি চাষে কী অপরিহার্য মনে করেন?","correctAns": "মাটির উর্বরতা বৃদ্ধি","distractor1": "বুদ্ধিমত্তা বৃদ্ধি"," distractor2": "আলস্য প্রকাশ","distractor3": "শিক্ষার মাধ্যমে অর্থ উৎপাদন"},	9.11559	8.142	সত্যিকারের কৃষিভিত্তিক প্রাসঙ্গিকতা বজায় রয়েছে, distractor-ও যুক্তিসঙ্গত।	52.15%	7.86	
			{"question": "লেখক কী সমাজের মূল সমস্যা বলে মনে করেন?","correctAns": "আলস্য ও আকৃতিমূলক উন্নতি চাওয়া","distractor1": "অর্থ অভাব","distractor2": "শিক্ষার অভাব"," distractor3": "দুর্ভিক্ষ"}	6.30475	8.47639	মূল থিম যথার্থভাবে ধরা পড়েছে, সব বিকল্প বাস্তব ও প্রাসঙ্গিক।	36.20%	7.36	
		"মমতাদি গল্পে ছেলেটির সাথে মমতাদির সম্পর্ক"	{"question": "স্নেহ ও আদরের জন্য কার প্রতি কর্মীর অনুরাগ প্রকাশ হয়েছে?","correctAns": "মার প্রতি","distractor1": "ছেলের প্রতি"," distractor2": "কর্মী বামুনদির প্রতি"," distractor3": "কোনোটিই নয়"},	7.1521	8.05133	সঠিক উত্তর স্পষ্ট হলেও distractor-গুলো যুক্তিসম্মত এবং বিভ্রান্তিকর।	48.09%	8.13	
			{"question": "বামুনদি কে কী বলে অভিভাষণ করেছিল?","correctAns": "দিদি"," distractor1": "বামুনদি","distractor2": "খোকা","distractor3": "মা"},	8.73918	7.78214	নামধারণের প্রশ্নে distractor দুর্বল, তাই সামান্য স্কোর কম।	51.88%	17.17	

			{"question": "মা বামুনদির কাছে কী বলেছিলেন?","correctAns": "ও ওইরকম বকবক করে","distractor1": "তুমি আমার ছেলে উঠে আসো","distractor2": "তোমার মা রাগ করবেন","distractor3": "তুমি আমার দিদি"},	6.75169	7.59988	উত্তরটি বাক্যভিত্তিক উদ্ধৃতি, distractor-এ নাট্যরস নেই, তাই গড় মান।	48.63%	8.1	
			{"question": "বামুনদি ছেলের প্রতি কী ভাবে কাজ করেছিল?","correctAns": "সুন্দরভাবে সম্পন্ন করেছিল","distractor1": "অনর্থক প্রশ্ন করে","distractor2": "নির্দেশের অভাবে কাজ ফেলে রেখেছিল","distractor3": "কোনোটিই নয়"},	7.23226	8.12447	প্রতিক্রিয়াধর্মী প্রশ্ন, সঠিক উত্তরের সঙ্গে যুক্তিসম্মত distractor।	50.90%	7.26	
			{"question": "মার কাছে বামুনদির চোখের জল ফেলা কী দেখায়?","correctAns": "সম্মানের আগাগোড়া করুণা","distractor1": "অপমান করা","distractor2": "ছেলের প্রতি ক্ষুণ্ণতা"," distractor3": "অর্থের জন্য ক্ষুধা"},	9.07727	8.3327	নাটকীয়তাভরা প্রশ্নে সব বিকল্পের ভিন্নতাও যুক্তিযুক্ত, ভালো মানের প্রশ্ন।	51.34%	9.31	2.54
			{"question": "বামুনদি ছেলের সঙ্গে কী করে আচরণ করেছিল?","correctAns": "গল্প করেছিল","distractor1": "গাল দিয়েছিল"," distractor2": "প্রশ্ন করেছিল","distractor3": "বিষয়টি নির্দেশ করেছিল"},	7.68093	8.22853	আচরণবিষয়ক প্রশ্নটি মানবিক, উত্তর ও distractor-গুচ্ছ উপযুক্ত।	49.92%	9.76	
			{"question": "বামুনদি কী কারণে ছেলের প্রতি আদর করেছিল?","correctAns": "সম্মানের জন্য","distractor1": "পনের টাকার জন্য"," distractor2": "স্বামীপুত্রের অনাহারের জন্য"," distractor3": "নিজের গর্ব প্রতিষ্ঠার জন্য"},	8.5729	8.19866	উত্তরটি সংবেদনশীল, distractor-গুলো অন্তর্দৃষ্টি দাবি করে।	52.71%	10.96	
			{"question": "ছেলে বামুনদির প্রতি কী বলেছিল?","correctAns": "দিদি"," distractor1": "বামুনদি","distractor2": "খোকা","distractor3": "মা"},	8.73918	7.9132	উত্তরে আবেগ আছে, তবে distractor-এ পর্যাপ্ত গভীরতা নেই।	44.37%	9.15	
			{"question": "মার চোখে কী ঝরে পড়েছিল?"," correctAns": "করুণা ও দুঃখ","distractor1": "ক্ষুণ্ণতা","distractor2": "রাগ","distractor3": "আশ্চর্য"},	6.45384	8.36289	দুঃখ ও করুণা স্পষ্ট, সব বিকল্প চিন্তা উদ্রেককারী।	49.57%	13.9	
			{"question": "বামুনদি ছেলের সঙ্গে কী করে ব্যবহার করেছিল?","correctAns": "সম্মানজনকভাবে","distractor1": "অসম্মানজনকভাবে","distractor2": "অপমানজনকভাবে","distractor3": "কোনোটিই নয়"}	7.41518	7.73047	অন্তর্নিহিত শ্রদ্ধাবোধ থাকা সত্ত্বেও distractor-গুলো কিছুটা দুর্বল।	50.44%	7.8	
		"অন্ধবধূ কবিতার কবির পরিচয়"	{"question": "যতীন্দ্রমোহন বাগচীর জন্ম কোন তারিখে হয়েছিল?","correctAns": "২৭ নভেম্বর ১৮৭৮","distractor1": "১৫ মে ১৮৮০"," distractor2": "১ জানুয়ারি ১৮৮৫"," distractor3": "১০ মার্চ ১৮৯০"},	6.46916	8.60325	তথ্যভিত্তিক প্রশ্ন, সঠিক উত্তর ও distractor কাছাকাছি, উত্তম মানের প্রশ্ন।	31.95%	6.64	
			{"question": "যতীন্দ্রমোহন বাগচী কোন জেলার জামশেদপুর গ্রামে জন্মগ্রহণ করেন?"," correctAns": "নদীয়া","distractor1": "মুর্শিদাবাদ","distractor2": "পূর্ব মেদিনীপুর"," distractor3": "বর্ধমান"},	8.41373	8.28641	জায়গা-নির্ভর প্রশ্নে distractor-গুলো বিভ্রান্তিকর, ভাল মান।	32.67%	4.9	
			{"question": "যতীন্দ্রমোহন বাগচীর কাব্যগ্রন্থসমূহের মধ্যে কোনটি অন্তর্ভুক্ত?"," correctAns": "অন্ধবধূ","distractor1": "পথের পাঁচালী","distractor2": "দুর্গাপুজো"," distractor3": "মুক্তিযুদ্ধ"},	8.19033	8.42188	গ্রন্থসংক্রান্ত প্রশ্নে distractor-গুলো প্রাসঙ্গিক হলেও সঠিক উত্তর সহজে শনাক্তযোগ্য।	41.20%	3.92	
			{"question": "যতীন্দ্রমোহন বাগচী কোন তারিখে মৃত্যুবরণ করেন?","correctAns": "১ ফেব্রুয়ারি ১৯৪৮","distractor1": "২০ মার্চ ১৯৫০","distractor2": "১০ এপ্রিল ১৯৪৭"," distractor3": "২৫ জুলাই ১৯৪৯"},	7.65165	8.09755	তারিখভিত্তিক প্রশ্নে বিভ্রান্তিকর বিকল্প থাকলেও একটি খুব দূরের, তাই কিছুটা নম্বর কাটা।	28.08%	5.59	
			{"question": "যতীন্দ্রমোহন বাগচীর কাব্যের মূল চিত্ররূপ কী?","correctAns": "গ্রাম জীবন","distractor1": "শহর জীবন"," distractor2": "বিদেশী সংস্কৃতি"," distractor3": "মধ্যযুগীয় ইতিহাস"},	7.42012	8.53362	মূল চিত্র যথার্থ ধরা হয়েছে, distractor তুলনায় কম সঠিক।	45.54%	5.89	3.68

			{"question": "যতীন্দ্রমোহন বাগচী কাব্যে কোন ভাষা ব্যবহার করেছেন?","correctAns": "সহজ ও সরল","distractor1": "জটিল ও বিষম্পন্ন","distractor2": "ইংরেজি ও গুপ্ত ভাষা","distractor3": "পাঁচালী ভাষা"},	**6.65291**	**8.61542**	ভাষা সংক্রান্ত প্রশ্নে বাস্তবধর্মী ও সূক্ষ্ম পার্থক্য সহ distractor যুক্ত।	**40.66%**	**8.19**	
			{"question": "যতীন্দ্রমোহন বাগচী কোন কবির সঙ্গে তুলনীয়?","correctAns": "বিভূতিভূষণ বন্দ্যোপাধ্যায়","distractor1": "রবীন্দ্রনাথ ঠাকুর","distractor2": "কাজী নজরুল ইসলাম","distractor3": "সুকান্ত মুখোপাধ্যায়"},	**8.36955**	**7.95268**	তুলনামূলক প্রশ্নে বিভ্রান্তিকরতা কিছুটা কম, তবে উত্তর ঠিক।	**39.80%**	**8.23**	
			{"question": "যতীন্দ্রমোহন বাগচী কোন সাহিত্যিক পরিচয়ে বিশেষ হয়েছিলেন?"," correctAns": "কবি","distractor1": "গল্প লেখক","distractor2": "উপন্যাসকার"," distractor3": "পত্রিকা সম্পাদক"},	**7.25125**	**7.72113**	পরিচয়ভিত্তিক প্রশ্নে distractor-গুলো কিছুটা দুর্বল হলেও মান বজায়।	**41.74%**	**5.14**	
			{"question": "যতীন্দ্রমোহন বাগচীর কাব্যগ্রন্থসমূহের মধ্যে কোনটি অন্তর্ভুক্ত নয়?"," correctAns": "মুক্তিযুদ্ধ","distractor1": "লেখা","distractor2": "অপরাজিতা"," distractor3": "মহাভারতী"},	**8.83189**	**8.18057**	'অন্তর্ভুক্ত নয়' বিষয়ক প্রশ্নে দৃষ্টিভঙ্গি প্রয়োজন, distractor যথাযথ।	**30.16%**	**4.04**	
			{"question": "যতীন্দ্রমোহন বাগচী কোন সাহিত্যিক চরিত্র প্রকাশ করেছেন?"," correctAns": "গ্রাম চরিত্র","distractor1": "শহর চরিত্র","distractor2": "মধ্যযুগীয় চরিত্র","distractor3": "বৈদেশিক চরিত্র"}	**9.19934**	**8.474**	চরিত্র বিশ্লেষণে প্রাসঙ্গিকতা রয়েছে, distractor-ও ভালোভাবে গঠিত।	**40.34%**	**4.29**	
	Gemma 3: 12B	"নিরীহ বাঙালি গল্পে বাঙালির সমালোচনা"	{"question": ""আমরা সকলে কবি"- প্রবন্ধের মূল বক্তব্য কী?","correctAns": "বাঙালিরা নিজেদের অজান্তেই কবিতার ছন্দ ও রস সৃষ্টি করে।","distractor1": "বাঙালিরা সকলেই জন্মগতভাবে দক্ষ সাহিত্যিক।","distractor2": "প্রবন্ধকার কবিতা লেখার নিয়মাবলী বর্ণনা করছেন।","distractor3": "বাঙালি জাতির শ্রেষ্ঠত্বের প্রমাণ দেওয়া হয়েছে।"},	**5.20538**	**8.12345**	ডিস্ট্র্যাক্টর যথেষ্ট অর্থঘন ও ঘনিষ্ঠ; উচ্চ মানের প্রশ্ন।	**56.32%**	**11.98**	
			{"question": "প্রবন্ধকার বাঙালি পুরুষদের সম্পর্কে কী বলেছেন?","correctAns": "তাঁরা অলস, তরলমতি এবং শ্রমকাতর।"," distractor1": "তাঁরা কঠোর পরিশ্রমী এবং সাহসী।","distractor2": "তাঁরা দেশপ্রেমিক এবং কর্তব্যপরায়ণ।","distractor3": "তাঁরা শিল্পকলার প্রতি বিশেষভাবে আগ্রহী।"},	**7.45313**	**8.56789**	তথ্যভিত্তিক ও প্রাসঙ্গিক বিকল্প; ভাল মানের।	**37.04%**	**5.01**	
			{"question": ""হাওয়ার শাড়ি" বলতে কী বোঝানো হয়েছে?","correctAns": "সূক্ষ্ম সুতোর তৈরি হালকা শাড়ি।","distractor1": "রেশমের তৈরি দামি শাড়ি।","distractor2": "ছাপা cotton-এর শাড়ি।","distractor3": "ভারী সিল্কের শাড়ি।"},	**8.39344**	**8.23456**	ডিস্ট্র্যাক্টর তুলনামূলক অর্থঘন; ভাল মানের।	**53.15%**	**103.54**	
			{"question": "প্রবন্ধ অনুসারে, বাঙালিরা বাণিজ্য কিভাবে সহজ করেছে?","correctAns": "ঝুঁকি কমিয়ে, কঠিন পরিশ্রম এড়িয়ে বাণিজ্য করেছে।","distractor1": "বাণিজ্যপোত তৈরি করে অনেক দূরে পণ্য পাঠিয়েছে।"," distractor2": "নতুন নতুন বাজারের সন্ধান করেছে।","distractor3": "বিদেশি ব্যবসায়ীদের সঙ্গে চুক্তি করেছে।"},	**6.21007**	**8.67891**	প্রশ্নে ধারণা পরীক্ষার উপযোগী; উচ্চ মানের।	**38.58%**	**5.45**	
			{"question": ""কুন্তলীন"-এর সঙ্গে "কেশলীন" কী হিসাবে বিক্রয় হয়?","correctAns": "পাল্লা দিয়ে","distractor1": "complimentary offer","distractor2": "প্রতিযোগিতামূলকভাবে","distractor3": "একটির পরিপূরক হিসাবে"},	**7.99232**	**7.91234**	বিকল্পগুলো যথেষ্ট অর্থবহ; ভাল মানের।	**56.80%**	**40.51**	**2.24**

			{"question": "প্রবন্ধকার গৃহিণীদের রন্ধনকার্য করতে বলার কারণ কী?","correctAns": "তিনি তাদের স্বাধীনতায় বিশ্বাসী।"," distractor1": "তিনি মনে করেন এটা তাদের অধিকার।","distractor2": "তিনি মনে করেন এটা পুরুষদের বোঝা কমানোর উপায়।"," distractor3": "তিনি মনে করেন এটা তাদের পৌরুষত্ব প্রমাণ করার একটি উপায়।"},	6.51505	7.45678	প্রাসঙ্গিক বিকল্প ও যুক্তিসঙ্গত distractor; ভাল মানের।	48.22%	11.48	
			{"question": "M.R.A.C পাশ করা কীসের পরিপূরক?","correctAns": "কৃষিকাজের"," distractor1": "শিল্পকলার","distractor2": "বিজ্ঞানচর্চার","distractor3": "আইনচর্চার"},	9.57133	8.78912	তথ্যঘন ডিস্ট্র্যাক্টর; মোটামুটি ভাল মানের।	24.07%	39.47	
			{"question": "প্রবন্ধকার কেন মনে করেন যে কৃষিকাজ অপেক্ষা মস্তিষ্ক উর্বর করা সহজ?"," correctAns": "পরীক্ষামূলক জ্ঞান লাভ করা সহজ।","distractor1": "কৃষিকাজ কঠিন এবং সময়সাপেক্ষ।","distractor2": "বিজ্ঞানচর্চা সহজলভ্য।","distractor3": "কৃষিকাজ লাভজনক নয়।"},	5.19348	8.34567	তথ্য ও যুক্তির ভারসাম্য রক্ষা করা হয়েছে; উচ্চ মানের।	54.90%	10.47	
			{"question": ""দিব্যাঙ্গনা" বলতে কী বোঝানো হয়েছে?","correctAns": "স্বর্গের রূপসী নারী"," distractor1": "সাধারণ গৃহিণী"," distractor2": "ধনী মহিলা","distractor3": "বিদুষী নারী"},	8.46385	8.67123	ডিস্ট্র্যাক্টর অর্থঘন; ভাল মানের প্রশ্ন।	19.92%	19.97	
			{"question": "প্রবন্ধে "রায় বাহাদুর" বা "খাঁ বাহাদুর" উপাধি লাভের উপায় কী বলা হয়েছে?","correctAns": "অর্থ ব্যয় করে খ্যাতি অর্জন করা।","distractor1": "দেশের জন্য বড় কাজ করা।","distractor2": "দরিদ্রদের সাহায্য করা।","distractor3": "শিক্ষা ও সংস্কৃতিতে অবদান রাখা।"}	6.1424	8.23456	বিকল্পগুলো যথেষ্ট ঘনিষ্ঠ; ভাল মানের।	20.04%	21.67	
		"মমতাদি গল্পে ছেলেটির সাথে মমতাদির সম্পর্ক"	{"question": "মমতাদির গৃহে প্রথম আগমনকালে স্কুল পড়ুয়া ছেলেটির অনুভূতি কেমন ছিল?","correctAns": "সে আনন্দিত হয়েছিল এবং মমতাদিকে নিজের বাড়ির একজন ভাবছিল।","distractor1": "সে রেগে গিয়েছিল। ","distractor2": "সে ভয় পেয়েছিল।"," distractor3": "সে বিরক্ত হয়েছিল।"},	3.85564	8.78912	ডিস্ট্র্যাক্টর ঘনিষ্ঠ ও বাস্তবধর্মী; খুবই ভাল মানের।	60.50%	13.65	
			{"question": "মমতাদির আত্মমর্যাদার বিষয়টি গল্পের কোন দিকটি বিশেষভাবে ফুটিয়ে তুলেছে?","correctAns": "তিনি অসম্মানজনক আচরণ পেলে চাকরি ত্যাগ করেন।"," distractor1": "তিনি অন্যদের সাহায্য করতে দ্বিধা বোধ করেন।","distractor2": "তিনি শুধুমাত্র টাকার জন্য কাজ করেন।"," distractor3": "তিনি সবসময় নিজেদের অভাব প্রকাশ করেন।"},	4.20845	7.89123	প্রশ্নের উত্তরগুলো অর্থবহ ও উপযুক্ত।	44.62%	15.9	
			{"question": "জীবনময় লেনের পরিবেশ কেমন ছিল?","correctAns": "এটি নোংরা ও অন্ধকারাচ্ছন্ন ছিল।","distractor1": "এটি একটি পরিষ্কার ও আলোকিত রাস্তা ছিল।"," distractor2": "এটি গ্রামের মতো শান্ত ছিল।"," distractor3": "এটি শহরের মতো আধুনিক ছিল ।"},	6.76073	7.56789	ঘনিষ্ঠ ও অর্থঘন distractor; উচ্চ মানের।	21.11%	17.56	
			{"question": "মমতাদির গৃহে রান্না করার কাজের জন্য প্রথম যে প্রস্তাব দেওয়া হয়েছিল, তার মাসিক মাইনে কত ছিল?","correctAns": "পনের টাকা","distractor1": "দশ টাকা"," distractor2": "বারো টাকা","distractor3": "কুড়ি টাকা"},	4.58486	8.45678	ডিস্ট্র্যাক্টর যথেষ্ট অর্থঘন; ভাল মানের ।	56.75%	15.16	

			{"question": "মমতাদি কেন 'বামুনদি' সম্বোধন পছন্দ করতেন না?","correctAns": "তিনি এই সম্বোধনকে অপমানজনক মনে করতেন।"," distractor1": "তিনি বয়স্ক ব্যক্তি ছিলেন।"," distractor2": "তিনি অন্য কোনো সম্বোধন শুনতে চাইতেন।","distractor3": "গল্পের বর্ণনার জন্য এটা প্রয়োজন ছিল।"},	**5.02555**	**8.91234**	তথ্যঘন ও যথাযথ distractor; ভাল মানের।	**52.73%**	**17.28**	**3.5**
			{"question": "গল্পে 'হরির লুট' বলতে কী বোঝানো হয়েছে?","correctAns": "ধর্মীয় অনুষ্ঠানে বাতাসা বিতরণ","distractor1": "দরিদ্রদের মধ্যে খাবার বিতরণ","distractor2": "গ্রামের মানুষের মধ্যে উপহার বিতরণ"," distractor3": "শহরের মানুষের মধ্যে টাকা বিতরণ"},	**8.21235**	**8.12345**	প্রশ্নটি বাস্তবধর্মী; মোটামুটি ভাল মানের।	**30.39%**	**23.76**	
			{"question": "মমতাদির স্বামীর পেশা কী ছিল?","correctAns": "তার কোনো চাকরি ছিল না।","distractor1": "তিনি ছিলেন একজন শিক্ষক।","distractor2": "তিনি ছিলেন একজন ডাক্তার।","distractor3": "তিনি ছিলেন একজন ব্যবসায়ী।"},	**3.47848**	**7.67891**	বিকল্পগুলো অর্থঘন ও যথেষ্ট ঘনিষ্ঠ।	**49.62%**	**7.95**	
			{"question": "মমতাদির গৃহে কাজ করার সময় ছেলেটির প্রতি তার আচরণ কেমন ছিল?"," correctAns": "তিনি স্নেহপূর্ণ ও যত্নশীল ছিলেন। ","distractor1": "তিনি সবসময় রুঢ় ছিলেন। ","distractor2": "তিনি সবসময় বকাঝকা করতেন।","distractor3": "তিনি উদাসীন ছিলেন।"},	**7.70506**	**7.23456**	ডিস্ট্র্যাক্টর অর্থবহ; ভাল মানের।	**58.19%**	**11.49**	
			{"question": "গল্পের শেষে মায়ের মন্তব্য থেকে কী বোঝা যায়?","correctAns": "মমতাদি ছেলেটিকে খুব ভালোবাসতেন।","distractor1": "মমতাদি খারাপ কাজ করেন।","distractor2": "মমতাদির স্বামীর চাকরি হয়ে গেছে।"," distractor3": "মমতাদি ছেলেটির কাছে মিথ্যা বলেছিলেন।"},	**4.53851**	**8.56789**	প্রাসঙ্গিক distractor গুলি খুব ভালো বাছাই করা।	**39.34%**	**3.47**	
			{"question": ""রহস্যের ঘোমটা খুলে সহজ মানুষ হয়ে গেল" - এই উক্তিটি কার ক্ষেত্রে প্রযোজ্য?","correctAns": "মমতাদি"," distractor1": "মমতাদির স্বামী"," distractor2": "মমতাদির পুত্র","distractor3": "ছেলেটি"}	**6.27391**	**8.34567**	ডিস্ট্র্যাক্টর যথেষ্ট যুক্তিযুক্ত; উচ্চ মানের।	**45.61%**	**14.19**	
		"অন্ধবধূ কবিতার কবির পরিচয়"	{"question": "যতীন্দ্রমোহন বাগচীর জন্ম কোন সালে?","correctAns": "১৮৭৮"," distractor1": "১৮৭৭","distractor2": "১৮৭৯","distractor3": "১৮৮০"},	**6.09234**	**7.89123**	প্রশ্নটি সুনির্দিষ্ট; distractor যথেষ্ট ঘনিষ্ঠ।	**30.66%**	**9.97**	
			{"question": "যতীন্দ্রমোহন বাগচীর জন্মস্থান কোথায়?","correctAns": "নদীয়া জেলার জামশেদপুর","distractor1": "নদীয়া জেলার কৃষ্ণনগর","distractor2": "হুগলি জেলার ডানকান","distractor3": "মুর্শিদাবাদ জেলার কান্দি"},	**9.71483**	**8.45678**	তথ্যভিত্তিক ও যুক্তিযুক্ত distractor ।	**31.32%**	**12.71**	
			{"question": "জতীন্দ্রমোহন বাগচীর কবিমানসের প্রধান বৈশিষ্ট্য কী?"," correctAns": "পল্লি-প্রীতি","distractor1": "নাগরিক জীবনের প্রতি আকর্ষণ"," distractor2": "প্রকৃতি বিরহ","distractor3": "বিলাসিতা ও ঐশ্বর্য বর্ণনা"},	**8.71211**	**8.12345**	ডিস্ট্র্যাক্টর কিছুটা দূরবর্তী হলেও মোটামুটি ভাল মানের।	**43.78%**	**6.54**	
			{"question": ""পথের পাঁচালী"র রচয়িতার নাম কী?","correctAns": "বিভূতিভূষণ বন্দ্যোপাধ্যায়","distractor1": "জীবনানন্দ দাশ","distractor2": "রবীন্দ্রনাথ ঠাকুর"," distractor3": "শরৎচন্দ্র চট্টোপাধ্যায়"},	**9.81465**	**8.67891**	ঘনিষ্ঠ এবং তথ্যঘন বিকল্প; ভাল মানের।	**28.35%**	**5.55**	

			{"question": "জতীন্দ্রমোহন বাগচীর রচনায় কোন বিষয়টি বিশেষভাবে ফুটিয়ে তোলার প্রয়াস ছিল?","correctAns": "গ্রাম বাংলার শ্যামল রূপ","distractor1": "রাজনৈতিক অস্থিরতা"," distractor2": "শহরের আধুনিক জীবন"," distractor3": "রাজকীয় বিলাসিতা"},	**6.63119**	**8.23456**	তথ্য ও ধারণাভিত্তিক প্রশ্ন; উচ্চ মানের।	**39.16%**	**21.34**	
			{"question": "জতীন্দ্রমোহন বাগচী কোন ভাষায় গ্রামজীবনের সুখ-দুঃখ প্রকাশ করেছেন?","correctAns": "সহজ-সরল ভাষায়","distractor1": "কঠিন ও জটিল ভাষায়","distractor2": "কাব্যিক ও অলঙ্কৃত ভাষায়","distractor3": "আরবি ভাষায়"},	**8.92283**	**8.56789**	ডিস্ট্র্যাক্টর যথেষ্ট অর্থবহ।	**38.41%**	**7.9**	**3.99**
			{"question": "নিচের কোন কাব্যগ্রন্থটি জতীন্দ্রমোহন বাগচীর লেখা নয়?"," correctAns": "সোনালী প্রান্তর"," distractor1": "অপরাজিতা","distractor2": "নাগকেশর","distractor3": "বন্ধুর দান"},	**9.32164**	**8.34567**	বিকল্পগুলো যথেষ্ট ঘনিষ্ঠ।	**35.52%**	**12.67**	
			{"question": "জতীন্দ্রমোহন বাগচীর কাব্যশৈলী কোন কবির মতো?","correctAns": "জীবনানন্দ দাশ","distractor1": "কাজী নজরুল ইসলাম","distractor2": "সুভেন্দ্রনাথ দত্ত","distractor3": "মাইকেল মধুসূদন দত্ত"},	**8.54353**	**7.91234**	তথ্যঘন distractor; ভাল মানের।	**43.68%**	**8.96**	
			{"question": "জতীন্দ্রমোহন বাগচী কত সালে মৃত্যুবরণ করেন?","correctAns": "১৯৪৮ সালের ১লা ফেব্রুয়ারি","distractor1": "১৯৪৮ সালের ১লা মার্চ","distractor2": "১৯৪৯ সালের ১লা জানুয়ারি","distractor3": "১৯৫০ সালের ১লা মে"},	**5.51749**	**8.78912**	ডিস্ট্র্যাক্টর অর্থঘন; খুবই ভাল মানের।	**29.28%**	**6.6**	
			{"question": ""গ্রাম-জীবনের অতি সাধারণ বিষয় ও সুখ-দুঃখ"-এর প্রকাশ জতীন্দ্রমোহন বাগচীর কবিতার -","correctAns": "বিশেষত্ব","distractor1": "দুর্বলতা"," distractor2": "সীমাবদ্ধতা","distractor3": "জটিলতা"}	**8.12112**	**8.45678**	প্রাসঙ্গিক distractor বাছাই যথেষ্ট ভাল হয়েছে।	**50.38%**	**12.48**	
Bangladesh & Global Studies	**Gemini 2.0 Flash**	"১৯৬৯ সালের গণঅভ্যুত্থান"	{"question": "১৯৬৯ সালের গণঅভ্যুত্থানে কোন ছাত্রনেতা শেখ মুজিবুর রহমানকে 'বঙ্গবন্ধু' উপাধিতে ভূষিত করার প্রস্তাব উত্থাপন করেন?","correctAns": "তোফায়েল আহমেদ","distractor1": "আসাদুজ্জামান আসাদ","distractor2": "ড. শামসুজ্জোহা"," distractor3": "নুরুল আমীন"},	**9.14995**	**8.12345**	ডিস্ট্র্যাক্টর মোটামুটি যুক্তিযুক্ত ও অর্থঘন।	**68.07%**	**2.37**	
			{"question": "আইয়ুব খান কত তারিখে শেখ মুজিবুর রহমানকে নিঃশর্ত মুক্তি দিতে বাধ্য হন?","correctAns": "২২শে ফেব্রুয়ারি"," distractor1": "২১শে ফেব্রুয়ারি"," distractor2": "২৩শে ফেব্রুয়ারি"," distractor3": "২৪শে ফেব্রুয়ারি"},	**5.87329**	**7.67891**	তথ্যনির্ভর ও ডিস্ট্র্যাক্টর যথেষ্ট ঘনিষ্ঠ।	**42.18%**	**3.21**	
			{"question": "১৯৬৯ সালের গণঅভ্যুত্থানের ফলে কোন সামরিক শাসক পদত্যাগ করতে বাধ্য হন?","correctAns": "আইয়ুব খান"," distractor1": "ইয়াহিয়া খান","distractor2": "ইস্কান্দার মির্জা","distractor3": "জুলফিকার আলী ভুট্টো"},	**9.17636**	**8.56789**	তারিখঘনিষ্ঠ distractor — মোটামুটি ভালো মানের।	**67.76%**	**3.37**	
			{"question": "আগরতলা মামলায় কতজনকে আসামি করা হয়েছিল?","correctAns": "৩৫ জন","distractor1": "৩০ জন","distractor2": "৪০ জন","distractor3": "২৫ জন"},	**7.17862**	**8.23456**	তথ্যনির্ভর প্রশ্ন, distractor যথেষ্ট অর্থঘন।	**42.61%**	**5.09**	
			{"question": "'ঐতিহাসিক আগরতলা মামলা' কোন সালে দায়ের করা হয়?","correctAns": "১৯৬৮ সালে","distractor1": "১৯৬৬ সালে","distractor2": "১৯৬৭ সালে"," distractor3": "১৯৬৯ সালে"},	**6.94421**	**7.34567**	সংখ্যাভিত্তিক প্রশ্ন — distractor যথেষ্ট ঘনিষ্ঠ।	**41.32%**	**35.08**	
			{"question": "১৯৬৯ সালের গণঅভ্যুত্থানে শহিদ হওয়া রাজশাহী বিশ্ববিদ্যালয়ের শিক্ষকের নাম কী?","correctAns": "ড. শামসুজ্জোহা"," distractor1": "আসাদুজ্জামান আসাদ"," distractor2": "সার্জেন্ট জহুরুল হক"," distractor3": "মতিউর রহমান"},	**8.75564**	**8.91234**	তারিখভিত্তিক; distractor যথেষ্ট অর্থঘন।	**56.43%**	**8.43**	**3.39**

			{"question": "মৌলিক গণতন্ত্র ব্যবস্থা কে চালু করেন?","correctAns": "আইয়ুব খান","distractor1": "ইয়াহিয়া খান","distractor2": "জুলফিকার আলী ভুট্টো","distractor3": "ইস্কান্দার মির্জা"},	9.17636	8.12345	প্রাসঙ্গিক distractor; উচ্চ মানের।	49.62%	7.02	
			{"question": "শেখ মুজিবুর রহমানের বিরুদ্ধে পাকিস্তান দণ্ডবিধির কোন ধারায় অভিযোগ আনা হয়?","correctAns": "১২১ ও ১৩১","distractor1": "১২০ ও ১৩০","distractor2": "১২২ ও ১৩২","distractor3": "১২৩ ও ১৩৩"},	9.78431	8.45678	তুলনামূলক কম জটিল; distractor যথেষ্ট অর্থবহ।	47.80%	6.79	
			{"question": "১৯৬৯ সালের গণঅভ্যুত্থানের প্রেক্ষাপটে, শেখ মুজিবুর রহমানকে সংবর্ধনা কোথায় দেওয়া হয়েছিল?","correctAns": "রেসকোর্স ময়দান (সোহরাওয়ার্দী উদ্যান)","distractor1": "পল্টন ময়দান","distractor2": "ঢাকা বিশ্ববিদ্যালয় প্রাঙ্গণ","distractor3": "সংসদ ভবন চত্বর"},	8.99407	7.78912	আইনের ধারা ঘনিষ্ঠ, ভাল প্রশ্ন।	66.80%	11.86	
			{"question": "বিপ্লবী পরিষদের পরিকল্পনা অনুযায়ী, বিদ্রোহীরা কোথায় কমান্ডো স্টাইলে হামলা চালানোর কথা ছিল?","correctAns": "ক্যান্টনমেন্টগুলোতে","distractor1": "কারাগারগুলোতে","distractor2": "থানাগুলোতে","distractor3": "সবগুলো সরকারি অফিসে"}	8.38899	8.23456	স্থানভিত্তিক প্রশ্ন; distractor ঘনিষ্ঠ।	54.25%	9.68	
		"বাংলাদেশে জাতিসংঘের ভূমিকা"	{"question": "জাতিসংঘে কবে বাংলাদেশ কততম সদস্য হিসেবে অন্তর্ভুক্ত হয়?","correctAns": "১৯৭৪ সালে, ১৩৬তম","distractor1": "১৯৭২ সালে, ১৩০তম","distractor2": "১৯৭৫ সালে, ১৩৭তম","distractor3": "১৯৭৩ সালে, ১৩৫তম"},	4.7795	8.67891	বিকল্পগুলো যথেষ্ট বাস্তবধর্মী।	63.42%	10.79	2.78
			{"question": "কত সালে বাংলাদেশের তৎকালীন পররাষ্ট্রমন্ত্রী হুমায়ুন রশীদ চৌধুরী জাতিসংঘের সাধারণ পরিষদের অধিবেশনে সভাপতিত্ব করেন?","correctAns": "১৯৮৬ সালে","distractor1": "১৯৮৪ সালে","distractor2": "১৯৮৫ সালে","distractor3": "১৯৮৭ সালে"},	6.96719	8.34567	তারিখ ও সদস্য সংখ্যা ভিত্তিক — মোটামুটি ঘনিষ্ঠ।	58.61%	2.25	
			{"question": "নারী নির্যাতন প্রতিরোধ পক্ষ বিশ্বব্যাপী কোন সময় পালিত হয়?","correctAns": "২৫শে নভেম্বর থেকে ১০ই ডিসেম্বর","distractor1": "২৫শে নভেম্বর থেকে ৫ই ডিসেম্বর","distractor2": "২০শে নভেম্বর থেকে ৩০শে নভেম্বর","distractor3": "১লা ডিসেম্বর থেকে ১৫ই ডিসেম্বর"},	9.00218	8.56789	তারিখভিত্তিক; distractor অর্থঘন।	62.55%	9.53	
			{"question": "বর্তমানে, জাতিসংঘের শান্তিরক্ষা মিশনে বাংলাদেশের প্রায় কত সৈন্য কাজ করছে?","correctAns": "১১,০০০","distractor1": "৯,০০০","distractor2": "১০,০০০","distractor3": "১২,০০০"},	9.48569	8.91234	সময়কালভিত্তিক প্রশ্ন; distractor যথেষ্ট যুক্তিযুক্ত।	62.21%	4.66	
			{"question": "সিয়েরা লিওনে কোন ভাষাকে দ্বিতীয় রাষ্ট্রভাষার মর্যাদা দেওয়া হয়েছে?","correctAns": "বাংলা","distractor1": "ইংরেজি","distractor2": "ফরাসি","distractor3": "আরবি"},	6.91452	8.12345	সংখ্যাভিত্তিক প্রশ্ন; distractor খুব ঘনিষ্ঠ।	34.33%	31.1	
			{"question": "জাতিসংঘের কোন সংস্থা বাংলাদেশের শিশুদের শিক্ষা ও চিকিৎসা নিশ্চিত করার জন্য কাজ করছে?","correctAns": "ইউনিসেফ (UNICEF)","distractor1": "ইউএনডিপি (UNDP)","distractor2": "ইউনেস্কো (UNESCO)","distractor3": "এফএও (FAO)"},	3.94835	8.45678	ভাষাভিত্তিক প্রশ্ন — বেশ সহজবোধ্য কিন্তু distractor যথেষ্ট।	65.20%	2.61	
			{"question": "জাতিসংঘের নিরাপত্তা পরিষদে কতটি স্থায়ী সদস্য রাষ্ট্র রয়েছে?","correctAns": "৫টি","distractor1": "৩টি","distractor2": "৪টি","distractor3": "৬টি"},	5.32611	8.23456	সংস্থা সনাক্তকরণ প্রশ্ন — distractor অর্থঘন।	73.03%	14.54	

			{"question": "জাতিসংঘের সদর দপ্তর কোথায় অবস্থিত?","correctAns": "নিউইয়র্ক, আমেরিকা","distractor1": "জেনেভা, সুইজারল্যান্ড","distractor2": "প্যারিস, ফ্রান্স","distractor3": "হেগ, নেদারল্যান্ডস"},	**4.61124**	**8.67891**	সংখ্যাভিত্তিক প্রশ্ন; মোটামুটি ভালো।	**73.11%**	**14.9**	
			{"question": "জাতিসংঘের কোন সংস্থা বাংলাদেশের খাদ্য নিরাপত্তা নিয়ে কাজ করে?","correctAns": "এফএও (FAO)","distractor1": "ই�এনডিপি (UNDP)","distractor2": "ইউনিসেফ (UNICEF)","distractor3": "ডব্লিউএইচও (WHO)"},	**4.83762**	**7.12345**	বস্তুগত প্রশ্ন; distractor স্পষ্ট ও অর্থঘন।	**55.37%**	**4.81**	
			{"question": "জাতিসংঘের কোন সংস্থাটি উদ্বাস্তু বিষয়ক কার্যক্রম পরিচালনা করে?","correctAns": "বিএইচসিআর","distractor1": "ইউএনডিপি","distractor2": "ইউনিফেম","distractor3": "ইউএনএফপিএ"}	**9.42831**	**8.78912**	প্রাসঙ্গিক তথ্য; distractor অর্থবহ।	**61.59%**	**6.86**	
		"বাংলাদেশের তথ্য অধিকার আইন"	{"question": "বাংলাদেশের সংবিধানের কত নম্বর অনুচ্ছেদে চিন্তা, বিবেক ও বাকস্বাধীনতা মৌলিক অধিকার হিসেবে স্বীকৃত?","correctAns": "৩৯","distractor1": "৪০","distractor2": "৩৮","distractor3": "৪১"},	**9.44903**	**8.34567**	অধিকার অনুচ্ছেদ; খুবই ভাল distractor বাছাই।	**54.95%**	**2.87**	**3.27**
			{"question": "বাংলাদেশ সরকার তথ্য অধিকার আইন কত তারিখে জারি করে?","correctAns": "৫ এপ্রিল ২০০৯","distractor1": "৫ এপ্রিল ২০০৮","distractor2": "৫ মার্চ ২০০৯","distractor3": "৬ এপ্রিল ২০১০"},	**4.51986**	**8.56789**	তারিখ সন্নিবেশ প্রশ্ন; মোটামুটি ঘনিষ্ঠ distractor।	**63.12%**	**5.92**	
			{"question": "তথ্য অধিকার আইন অনুযায়ী, 'তথ্য' বলতে নিচের কোনটি অন্তর্ভুক্ত নয়?","correctAns": "দাপ্তরিক নোটশিট","distractor1": "স্মারক","distractor2": "চুক্তি","distractor3": "প্রতিবেদন"},	**8.15521**	**8.23456**	আইনের বিষয় — কিছুটা তুলনামূলক কঠিন।	**52.80%**	**17.07**	
			{"question": "তথ্য কমিশনে প্রধান তথ্য কমিশনারসহ কতজন তথ্য কমিশনার থাকেন?","correctAns": "দুইজন","distractor1": "একজন","distractor2": "তিনজন","distractor3": "চারজন"},	**5.77361**	**7.45678**	সংখ্যাগত প্রশ্ন — ভালো মানের।	**57.37%**	**6.18**	
			{"question": "তথ্য অধিকার আইন অনুসারে, তথ্য জানার জন্য কিভাবে আবেদন করতে হয়?","correctAns": "লিখিতভাবে বা ইলেকট্রনিক মাধ্যমে","distractor1": "শুধুমাত্র লিখিতভাবে","distractor2": "শুধুমাত্র ইলেকট্রনিক মাধ্যমে","distractor3": "শুধুমাত্র মৌখিকভাবে"},	**9.53252**	**8.91234**	আবেদনের ধরন — যুক্তিযুক্ত distractor।	**65.83%**	**7.97**	
			{"question": "দায়িত্বপ্রাপ্ত কর্মকর্তা নির্ধারিত সময়সীমার মধ্যে তথ্য প্রদানে ব্যর্থ হলে, অনুরোধকারী কত দিনের মধ্যে আপিল করতে পারবেন?","correctAns": "৩০ দিন","distractor1": "১৫ দিন","distractor2": "২০ দিন","distractor3": "২৫ দিন"},	**4.53803**	**7.67891**	সময়সীমা — ঘনিষ্ঠ distractor।	**53.97%**	**7.04**	
			{"question": "আপিল কর্তৃপক্ষের নিকট সুবিচার না পেলে, আবেদনকারী কার কাছে অভিযোগ পাঠাতে পারবেন?","correctAns": "তথ্য কমিশন","distractor1": "জেলা প্রশাসক","distractor2": "হাইকোর্ট","distractor3": "প্রধানমন্ত্রী"},	**7.12909**	**8.78912**	আপিল প্রক্রিয়া — যথেষ্ট অর্থঘন বিকল্প।	**45.35%**	**3.56**	
			{"question": "তথ্য সরবরাহের ক্ষেত্রে দায়িত্বপ্রাপ্ত কর্মকর্তার প্রধান দায়িত্ব কী?","correctAns": "আবেদন প্রক্রিয়া সম্পর্কে জনগণকে জানানো","distractor1": "শুধুমাত্র তথ্য সংরক্ষণ করা","distractor2": "তথ্যের গোপনীয়তা রক্ষা করা","distractor3": "শুধুমাত্র কর্তৃপক্ষের কাছ থেকে তথ্য নেয়া"},	**7.02794**	**7.23456**	দায়িত্ব প্রসঙ্গ — মোটামুটি ভালো distractor।	**51.62%**	**8.74**	

			{"question": "তথ্য অধিকার আইনের মাধ্যমে নিচের কোনটির উন্নয়ন নিশ্চিত করা যায়?","correctAns": "দরিদ্র, প্রান্তিক ও সুবিধাবঞ্চিত মানুষের","distractor1": "শুধুমাত্র ধনী ব্যক্তিদের","distractor2": "শুধুমাত্র সরকারি কর্মকর্তাদের","distractor3": "শুধুমাত্র শিক্ষিত জনগোষ্ঠীর"},	**6.32578**	**8.56789**	উন্নয়ন বিষয়ক প্রশ্ন — ঘনিষ্ঠ distractor।	**72.62%**	**11.48**	
			{"question": "তথ্য অধিকার আইন বাস্তবায়নের মাধ্যমে কিসের ভিত মজবুত হবে?","correctAns": "গণতন্ত্র","distractor1": "সামরিক শাসন","distractor2": "স্বৈরাচার","distractor3": "রাজতন্ত্র"}	**6.94386**	**8.34567**	মূল্যবোধভিত্তিক প্রশ্ন — ভাল মানের।	**71.80%**	**6.76**	
	Qwen 3: 14B	"১৯৬৯ সালের গণঅভ্যুত্থান"	{"question": "বঙ্গবন্ধু শেখ মুজিবুর রহমানের রাজনীতির চূড়ান্ত লক্ষ্য কী ছিল?","correctAns": "বাংলাদেশের স্বাধীনতা অর্জন","distractor1": "পূর্ব পাকিস্তানের সাংস্কৃতিক বিকাশ","distractor2": "পশ্চিম পাকিস্তানের অর্থনৈতিক সমৃদ্ধি","distractor3": "সামরিক সংগ্রামের সমাপ্তি"},	**8.03191**	**7.12345**	মূল লক্ষ্য সম্পর্কিত প্রশ্ন — মোটামুটি সহজ কিন্তু distractor অর্থবহ।	**48.12%**	**2.18**	**2.66**
			{"question": "বিপ্লবী পরিষদের পরিকল্পনা ছিল কী?","correctAns": "পূর্ব পাকিস্তানের সব ক্যান্টনমেন্টে কমান্ডো স্টাইলে হামলা চালানো","distractor1": "পশ্চিম পাকিস্তানের সামরিক বাহিনী সহায়তায় স্বাধীনতা ঘোষণা","distractor2": "শিক্ষানীতি পরিবর্তন করা","distractor3": "একটি বৈষয়িক অর্থনৈতিক সংগ্রাম শুরু করা"},	**7.38434**	**8.91234**	বিপ্লবী পরিষদের পরিকল্পনা — ভালো distractor বাছাই।	**50.06%**	**5.55**	
			{"question": "আগরতলা মামলায় মোট কতজন আসামি ছিল?","correctAns": "৩৫","distractor1": "২৫","distractor2": "৩০","distractor3": "৪০"},	**7.28119**	**8.23456**	সংখ্যাভিত্তিক প্রশ্ন — ঘনিষ্ঠ distractor।	**43.98%**	**6.98**	
			{"question": "আগরতলা মামলার শুনানি কোন বছরে শুরু হয়েছিল?","correctAns": "১৯৬৮","distractor1": "১৯৬৫","distractor2": "১৯৬৬","distractor3": "১৯৬৭"},	**6.18637**	**8.67891**	তারিখভিত্তিক প্রশ্ন — মোটামুটি ঘনিষ্ঠ।	**46.15%**	**6.02**	
			{"question": "আগরতলা মামলায় শেখ মুজিবুর রহমানকে কোন ধারার অধীনে অভিযোগ আনা হয়েছিল?","correctAns": "১২১ ও ১৩১","distractor1": "১১৮ ও ১২০","distractor2": "১৩২ ও ১৪০","distractor3": "১৪৫ ও ১৫০"},	**8.63512**	**8.45678**	আইনের ধারা — অনেক ঘনিষ্ঠ সংখ্যা।	**51.83%**	**4.16**	
			{"question": "জাতীয় গণতান্ত্রিক ফ্রন্ট (এনডিএফ) গঠনের লক্ষ্য কী ছিল?","correctAns": "সাংবিধানিক শাসন পুনঃপ্রতিষ্ঠা","distractor1": "অর্থনৈতিক সংস্কার","distractor2": "শিক্ষানীতি পরিবর্তন","distractor3": "সামরিক সংগ্রাম শুরু করা"},	**6.81333**	**8.12345**	নীতিগত প্রশ্ন — distractor তুলনামূলক দূরে।	**34.20%**	**2.81**	
			{"question": "আগরতলা মামলার বিচার কোথায় হয়েছিল?","correctAns": "ঢাকা ক্যান্টনমেন্টে","distractor1": "করাচিতে","distractor2": "লাহোরে","distractor3": "বরিশালে"},	**9.32873**	**8.56789**	বিচার স্থানভিত্তিক প্রশ্ন — অর্থবহ বিকল্প।	**39.76%**	**4.53**	
			{"question": "শিক্ষানীতি বিষয়ক ছাত্রদের আন্দোলনে কে অংশগ্রহণ করেছিলেন?","correctAns": "পেশাজীবী ও বিভিন্ন সম্প্রদায়","distractor1": "শুধুমাত্র ছাত্ররা","distractor2": "সামরিক কর্মকর্তারা","distractor3": "সরকারি কর্মকর্তারা"},	**5.85787**	**8.78912**	আন্দোলনের পরিধি — ভালো distractor।	**49.26%**	**3.64**	
			{"question": "বিপ্লবী পরিষদের পরিকল্পনা কী ছিল?","correctAns": "পূর্ব পাকিস্তানের সব ক্যান্টনমেন্টে কমান্ডো স্টাইলে হামলা চালানো","distractor1": "সামরিক সংগ্রাম শুরু করা","distractor2": "শিক্ষানীতি পরিবর্তন করা","distractor3": "সাংস্কৃতিক বিকাশ ঘোষণা"},	**7.78719**	**8.34567**	রিপিট হলেও distractor ঘনিষ্ঠ।	**50.37%**	**5.55**	

			{"question": "আগরতলা মামলার মূল অভিযোগ কী ছিল?","correctAns": "পূর্ব পাকিস্তানকে স্বাধীন রাষ্ট্র ঘোষণা","distractor1": "শিক্ষানীতি পরিবর্তন করা","distractor2": "অর্থনৈতিক সমৃদ্ধি ঘোষণা","distractor3": "সামরিক সংগ্রাম শুরু করা"}	**6.56171**	**8.23456**	মূল অভিযোগ — অর্থবহ distractor।	**38.37%**	**4.62**	
		"বাংলাদেশে জাতিসংঘের ভূমিকা"	{"question": "জাতিসংঘের প্রধান উদ্দেশ্য কী?","correctAns": "বিশ্বশান্তি ও নিরাপত্তা নিশ্চিত করা","distractor1": "জাতিগত বৈষম্য বৃদ্ধি","distractor2": "আর্থিক সাহায্য প্রদান বন্ধ করা","distractor3": "দেশগুলির মধ্যে সংঘর্ষ বৃদ্ধি"},	**6.76509**	**8.45678**	মূল উদ্দেশ্য — মোটামুটি সহজ প্রশ্ন।	**76.85%**	**3.01**	**1.95**
			{"question": "বাংলাদেশ জাতিসংঘের শান্তিরক্ষায় কতজন সৈন্য শহিদ হয়েছেন?","correctAns": "৮৮","distractor1": "৭৫","distractor2": "১০০","distractor3": "১২০"},	**7.62997**	**8.91234**	সংখ্যাগত শহীদ সংখ্যা — ঘনিষ্ঠ distractor।	**60.44%**	**4.54**	
			{"question": "জাতিসংঘের কোন সংস্থা বাংলাদেশে নারীদের উন্নতি জন্য কাজ করে?","correctAns": "ইউনিফেম","distractor1": "ইউনিসেফ","distractor2": "ইউএনএফপিএ","distractor3": "ডব্লিউএইচও"},	**9.6961**	**8.56789**	সংস্থা নির্ধারণ — কিছুটা tricky distractor।	**66.01%**	**3.54**	
			{"question": "জাতিসংঘের কোন উদ্দেশ্য আন্তর্জাতিক বিরোধের শান্তিপূর্ণ সমাধান নিশ্চিত করা?","correctAns": "পঞ্চম উদ্দেশ্য","distractor1": "দ্বিতীয় উদ্দেশ্য","distractor2": "ষষ্ঠ উদ্দেশ্য","distractor3": "চতুর্থ উদ্দেশ্য"},	**7.67427**	**8.12345**	উদ্দেশ্য সংখ্যা — ঘনিষ্ঠ distractor।	**71.09%**	**3.58**	
			{"question": "জাতিসংঘের কোন উদ্দেশ্য সকল মানুষের সমান অধিকার বৃদ্ধি করা?","correctAns": "দ্বিতীয় উদ্দেশ্য","distractor1": "চতুর্থ উদ্দেশ্য","distractor2": "তৃতীয় উদ্দেশ্য","distractor3": "পঞ্চম উদ্দেশ্য"},	**6.90197**	**8.67891**	উদ্দেশ্য সংখ্যা — মোটামুটি ভালো distractor।	**68.32%**	**4.49**	
			{"question": "বাংলাদেশ জাতিসংঘের মাধ্যমে শান্তি রক্ষা করতে কী ধরনের বাহিনী প্রেরণ করে?","correctAns": "সেনাবাহিনী ও পুলিশ বাহিনী","distractor1": "কেবল সেনাবাহিনী","distractor2": "কেবল পুলিশ বাহিনী","distractor3": "বেসামরিক বাহিনী"},	**7.80492**	**8.23456**	বাহিনীর ধরন — অর্থবহ distractor।	**70.72%**	**3.94**	
			{"question": "জাতিসংঘের কোন সংস্থা জনসংখ্যা পরিস্থিতি উন্নয়নে কাজ করে?","correctAns": "ইউএনএফপিএ","distractor1": "ইউনিফেম","distractor2": "ইউনিসেফ","distractor3": "ডব্লিউএইচও"},	**9.59574**	**8.78912**	সংস্থা সনাক্তকরণ — মোটামুটি সহজ।	**62.89%**	**3.33**	
			{"question": "জাতিসংঘের কোন উদ্দেশ্য আত্মনির্ভরতা অধিকারের স্বীকৃতি দেয়া?","correctAns": "ষষ্ঠ উদ্দেশ্য","distractor1": "সপ্তম উদ্দেশ্য","distractor2": "পঞ্চম উদ্দেশ্য","distractor3": "দ্বিতীয় উদ্দেশ্য"},	**6.02512**	**8.34567**	উদ্দেশ্য সংখ্যা — অর্থবহ বিকল্প।	**66.89%**	**4.3**	
			{"question": "বাংলাদেশ জাতিসংঘের শান্তিরক্ষায় কী ধরনের অবদান রেখেছে?","correctAns": "সেনাবাহিনী ও পুলিশ বাহিনী দ্বারা","distractor1": "অর্থ দ্বারা","distractor2": "শরণার্থীদের পালন বন্ধ করে","distractor3": "কোনো অবদান নয়"},	**7.06698**	**8.45678**	অবদানের ধরন — ভালো distractor।	**67.60%**	**2.78**	
			{"question": "জাতিসংঘের কোন উদ্দেশ্য সাংস্কৃতিক ক্ষেত্রে সহযোগিতা বৃদ্ধি করা?","correctAns": "তৃতীয় উদ্দেশ্য","distractor1": "চতুর্থ উদ্দেশ্য","distractor2": "পঞ্চম উদ্দেশ্য","distractor3": "ষষ্ঠ উদ্দেশ্য"}	**8.27355**	**8.91234**	উদ্দেশ্য সংখ্যা — অর্থঘন বিকল্প।	**58.97%**	**3.41**	
		"বাংলাদেশের তথ্য অধিকার আইন"	{"question": "বাংলাদেশের তথ্য অধিকার আইন কখন জারি হয়?","correctAns": "৫ এপ্রিল ২০০৯","distractor1": "৫ এপ্রিল ২০০৮","distractor2": "৫ এপ্রিল ২০১০","distractor3": "৫ এপ্রিল ২০০৭"},	**1.85401**	**7.67891**	তারিখ সন্নিবেশ — মোটামুটি ভালো।	**62.52%**	**4.99**	
			{"question": "তথ্য অধিকার আইনে কোন অনুচ্ছেদে চিন্তা, বিবেক ও বাকস্বাধীনতা অধিকার স্বীকৃত হয়?","correctAns": "৩৯ অনুচ্ছেদ","distractor1": "৩৭ অনুচ্ছেদ","distractor2": "৪০ অনুচ্ছেদ","distractor3": "৪১ অনুচ্ছেদ"},	**9.19607**	**8.12345**	অনুচ্ছেদ — ঘনিষ্ঠ distractor।	**60.00%**	**4.23**	

			{"question": "তথ্য অধিকার আইনে কোনটি তথ্যের অন্তর্ভুক্তি হবে না?","correctAns": "দাপ্তরিক নোটশিট","distractor1": "লগবহি"," distractor2": "অঙ্কিত চিত্র","distractor3": "ইলেকট্রনিক প্রক্রিয়া"},	8.53401	8.56789	তথ্য প্রকারভেদ — কিছুটা বিশ্লেষণমূলক।	43.07%	4.17	
			{"question": "তথ্য কমিশন কী ধরনের প্রতিষ্ঠান?","correctAns": "স্বাধীন প্রতিষ্ঠান"," distractor1": "সরকারি প্রতিষ্ঠান"," distractor2": "বাণিজ্যিক প্রতিষ্ঠান"," distractor3": "বিদ্যালয় প্রতিষ্ঠান"},	8.55268	8.23456	প্রতিষ্ঠানের ধরণ — মোটামুটি সহজ।	51.86%	4.64	
			{"question": "তথ্য অধিকার আইন অনুসারে কোন কর্মকর্তা আবেদনকারীকে সহায়তা করবে?","correctAns": "দায়িত্বপ্রাপ্ত কর্মকর্তা","distractor1": "তথ্য কমিশনার"," distractor2": "মন্ত্রী","distractor3": "জেলা কার্যালয়"},	8.41553	8.78912	কর্মকর্তা নির্ধারণ — অর্থবহ distractor।	64.72%	3.33	
			{"question": "তথ্য প্রদানে বাধ্য হবে কে?"," correctAns": "সরকারি কর্মকর্তা"," distractor1": "সাধারণ নাগরিক"," distractor2": "ব্যবসায়ী","distractor3": "বিদেশী প্রতিনিধি"},	6.95345	8.91234	কর্তৃত্ব নির্ধারণ — মোটামুটি সহজ।	42.71%	6.6	3.2
			{"question": "তথ্য প্রাপ্তির জন্য কোন পদ্ধতিগুলি ব্যবহার করা যাবে?"," correctAns": "চিঠি লিখে এবং ই-মেইলের মাধ্যমে","distractor1": "চিঠি লিখে"," distractor2": "ই-মেইলের মাধ্যমে"," distractor3": "স্বাক্ষর দিয়ে"},	8.36306	8.45678	আবেদনের মাধ্যম — ভালো distractor।	52.75%	3.91	
			{"question": "তথ্য সরবরাহে ব্যর্থতা হলে কোনটি করা যাবে?","correctAns": "তথ্য কমিশনে অভিযোগ পাঠানো","distractor1": "সরকারকে চিঠি লিখা","distractor2": "সাংবাদিক প্রতিবেদন প্রকাশ","distractor3": "তথ্য কমিশনে অভিযোগ পাঠানো এবং সাংবাদিক প্রতিবেদন প্রকাশ"},	8.98956	8.34567	ব্যর্থতার পরবর্তী ধাপ — মোটামুটি ভালো।	43.16%	6.89	
			{"question": "তথ্য কমিশনের কাজ কী?"," correctAns": "অভিযোগ গ্রহণ ও ব্যবস্থা নেয়া","distractor1": "কর্মকর্তাকে শাস্তি দেয়া","distractor2": "তথ্য সরবরাহ বন্ধ করা","distractor3": "মন্ত্রীদের কাছে অভিযোগ পাঠানো"},	8.31845	8.12345	কমিশনের দায়িত্ব — অর্থঘন বিকল্প।	55.31%	4.96	
			{"question": "তথ্য অধিকার আইনের প্রধান উদ্দেশ্য কী?","correctAns": "সুশাসন প্রতিষ্ঠা এবং দুর্নীতি নির্মূল","distractor1": "দুর্নীতি বৃদ্ধি","distractor2": "সরকারি কর্মকর্তাদের বেতন বৃদ্ধি","distractor3": "মানবাধিকার বিমোচন"}	7.60272	8.67891	আইনের লক্ষ্য — অর্থবহ distractor।	62.67%	2.39	
	Gemma 3: 12B	"১৯৬৯ সালের গণঅভ্যুত্থান"	{"question": "১৯৬৯ সালের গণঅভ্যুত্থান কোন সালে সংঘটিত হয়?","correctAns": "১৯৬৯ সালে","distractor1": "১৯৬১ সালে"," distractor2": "১৯৬২ সালে","distractor3": "১৯৭০ সালে"},	6.29617	8.23456	গণঅভ্যুত্থানের সাল — মোটামুটি সহজ।	64.08%	10.3	
			{"question": "ঊনসত্তরের গণঅভ্যুত্থানে ঢাকা বিশ্ববিদ্যালয়ের কোন ছাত্র শহীদ হন?"," correctAns": "আসাদুজ্জামান আসাদ"," distractor1": "তোফায়েল আহমেদ"," distractor2": "শামসুজ্জোহা","distractor3": "হোসেন শহীদ সোহরাওয়ার্দী"},	9.31285	8.56789	শহীদ ছাত্র — distractor তুলনামূলক অর্থবহ।	46.83%	7.04	
			{"question": "গণঅভ্যুত্থানের ফলে পাকিস্তানের কোন সামরিক শাসক পদত্যাগ করতে বাধ্য হন?","correctAns": "আইয়ুব খান"," distractor1": "জেনারেল ইয়াহিয়া খান"," distractor2": "হোসেন শহীদ সোহরাওয়ার্দী"," distractor3": "আসাদুজ্জামান আসাদ"},	8.91452	8.91234	আইয়ুব খানের পদত্যাগ — অর্থবহ distractor।	64.95%	3.37	
			{"question": "কোন উপাধিতে শেখ মুজিবকে রেসকোর্স ময়দানে ভূষিত করা হয়?"," correctAns": "বঙ্গবন্ধু","distractor1": "জাতীয় নেতা","distractor2": "গণবন্ধু"," distractor3": "সেরা বাঙালি"},	8.0158	8.45678	বঙ্গবন্ধু উপাধি — মোটামুটি সহজ।	43.84%	17.38	

			{"question": "সামরিক শাসনকে দীর্ঘস্থায়ী করার লক্ষ্যে আইয়ুব খান কোন ব্যবস্থা চালু করেন?","correctAns": "মৌলিক গণতন্ত্র"," distractor1": "জাতীয় গণতান্ত্রিক ফ্রন্ট (এনডিএফ)","distractor2": "ঐতিহাসিক আগরতলা মামলা","distractor3": "গণঅভ্যুত্থান"},	**8.45687**	**8.78912**	মৌলিক গণতন্ত্র ব্যবস্থা — ভালো distractor।	**51.52%**	**7.75**	**2.78**
			{"question": "আগরতলা মামলায় শেখ মুজিবুর রহমানকে কোন ধারায় অভিযুক্ত করা হয়েছিল?","correctAns": "ধারা ১২১ ও ১৩১","distractor1": "ধারা ১২৬"," distractor2": "ধারা ১৫১","distractor3": "ধারা ১৩৫"},	**7.60357**	**8.12345**	আগরতলা মামলা ধারা — ঘনিষ্ঠ distractor।	**51.07%**	**12.05**	
			{"question": "পাকিস্তান সৃষ্টির আগে পূর্ববাংলা কোন দিক থেকে পশ্চিম পাকিস্তানের চেয়ে অগ্রসর ছিল?","correctAns": "অর্থনৈতিক, সামাজিক ও শিক্ষাসহ বিভিন্ন ক্ষেত্রে"," distractor1": "শুধুমাত্র সামরিক ক্ষেত্রে"," distractor2": "শুধুমাত্র রাজনৈতিক ক্ষেত্রে"," distractor3": "শুধুমাত্র শিল্প ক্ষেত্রে"},	**6.68988**	**8.56789**	অগ্রসরতার কারণ — বিশ্লেষণমূলক।	**65.62%**	**5.13**	
			{"question": "সামরিক শাসনের ফলে পূর্ব পাকিস্তানের কোন শাসকদের হাতে ক্ষমতা কেন্দ্রীভূত হয়?","correctAns": "পশ্চিম পাকিস্তানি শাসকদের","distractor1": "পূর্ববাংলার রাজনীতিবিদদের","distractor2": "জাতীয় গণতান্ত্রিক ফ্রন্টের (এনডি� marital lawএফ)","distractor3": "ছাত্রসমাজের"},	**3.93429**	**8.23456**	ক্ষমতা কেন্দ্রীভূত — অর্থবহ বিকল্প।	**58.10%**	**4.59**	
			{"question": "হোসেন শহীদ সোহরাওয়ার্দীকে গ্রেফতার করার বছর কোনটি?"," correctAns": "১৯৬২ সাল","distractor1": "১৯৬১ সাল","distractor2": "১৯৬৩ সাল"," distractor3": "১৯৬৪ সাল"},	**5.15692**	**8.91234**	গ্রেফতারের সাল — কিছুটা সহজ।	**38.65%**	**18.24**	
			{"question": "বিপ্লবী পরিষদের মূল পরিকল্পনা কী ছিল?","correctAns": "ক্যান্টনমেন্টে হামলা করে অস্ত্র দখল করে স্বাধীনতা ঘোষণা করা"," distractor1": "শান্তিপূর্ণভাবে স্বাধীনতা অর্জন করা","distractor2": "জাতীয় গণতান্ত্রিক ফ্রন্ট (এনডিএফ) গঠন করা","distractor3": "ঐতিহাসিক আগরতলা মামলা প্রত্যাহার করা"}	**8.93862**	**8.34567**	বিপ্লবী পরিষদের পরিকল্পনা — বিশ্লেষণমূলক distractor।	**50.93%**	**5.09**	
		"বাংলাদেশে জাতিসংঘের ভূমিকা"	{"question": "বাংলাদেশে জাতিসংঘের কতগুলো অঙ্গ সংস্থা মিশন পরিচালনা করছে?"," correctAns": "সবক'টি","distractor1": "পাঁচটি","distractor2": "নয়টি"," distractor3": "তিনটি"},	**6.38539**	**7.67891**	অঙ্গসংস্থার সংখ্যা — অর্থবহ বিকল্প।	**62.79%**	**4.07**	
			{"question": "জাতির জনক বঙ্গবন্ধু শেখ মুজিবুর রহমান কোন বছরে জাতিসংঘের সাধারণ পরিষদে বক্তৃতা প্রদান করেন?"," correctAns": "১৯৭৪ সাল","distractor1": "১৯৭০ সাল","distractor2": "১৯৭১ সাল"," distractor3": "১৯৭৫ সাল"},	**9.56961**	**8.91234**	বক্তৃতার সাল — ঘনিষ্ঠ distractor।	**62.54%**	**2.44**	
			{"question": "নিরাপত্তা পরিষদ কতজন সদস্য নিয়ে গঠিত?","correctAns": "১৫ জন"," distractor1": "২০ জন","distractor2": "২২ জন","distractor3": "২৫ জন"},	**5.16436**	**7.34567**	নিরাপত্তা পরিষদ সদস্য সংখ্যা — ভালো distractor।	**59.14%**	**23.46**	
			{"question": "সিয়েরা লিওনে বাংলা ভাষা কোন মর্যাদা লাভ করেছে?","correctAns": "দ্বিতীয় রাষ্ট্রভাষা","distractor1": "সরকারি ভাষা","distractor2": "প্রথম রাষ্ট্রভাষা"," distractor3": "কোনো মর্যাদা লাভ করেনি"},	**8.75203**	**8.45678**	সিয়েরা লিওনে ভাষার মর্যাদা — মোটামুটি।	**47.09%**	**12.23**	
			{"question": "আন্তর্জাতিক নারী নির্যাতন প্রতিরোধ দিবস পালনের ঘোষণা কোন সালে জাতিসংঘের সাধারণ পরিষদে দেওয়া হয়?"," correctAns": "১৯৯৯ সাল","distractor1": "১৯৯৭ সাল","distractor2": "১৯৯৮ সাল"," distractor3": "২০০০ সাল"},	**8.89799**	**8.78912**	নারী নির্যাতন দিবস — সংখ্যা ঘনিষ্ঠ।	**70.73%**	**6.45**	

			{"question": "বিশ্বে জাতি সংঘ গঠনের মূল উদ্দেশ্য কী ছিল?","correctAns": "মানবাধিকার রক্ষা করা","distractor1": "বাণিজ্য বৃদ্ধি করা","distractor2": "সামরিক শক্তি বাড়ানো","distractor3": "উপনিবেশ স্থাপন করা"},	**8.9738**	**7.12345**	মূল উদ্দেশ্য — মোটামুটি সহজ।	**74.04%**	**4.97**	**3.39**
			{"question": "জাতিসংঘের শান্তিরক্ষা মিশনে এখন পর্যন্ত কতজন বাংলাদেশি সৈন্য শহিদ হয়েছেন?","correctAns": "৮৮ জন"," distractor1": "৩৫ জন","distractor2": "৫৫ জন","distractor3": "১২০ জন"},	**6.8246**	**8.56789**	শহীদ সৈন্য সংখ্যা — ঘনিষ্ঠ distractor।	**63.93%**	**6.5**	
			{"question": "বাংলাদেশের অর্থনীতিকে উন্নত করতে কোন সংস্থাটি কাজ করছে?"," correctAns": "ইউএনডিপি (UNDP)"," distractor1": "ইউনিসেফ (UNICEF)"," distractor2": "ডব্লিউএইচও (WHO)"," distractor3": "ইউনিফেম (UNIFEM)"},	**6.83148**	**8.23456**	UNDP ভূমিকা — অর্থবহ বিকল্প।	**52.96%**	**3.51**	
			{"question": "বাংলাদেশের জনসংখ্যা পরিস্থিতি উন্নয়নে কোন সংস্থা কাজ করছে?"," correctAns": "ইএনএফপিএ (UNFPA)"," distractor1": "ইযএনডিপি (UNDP)"," distractor2": "ইযনিসেফ (UNICEF)"," distractor3": "ডব্লিযএইচও (WHO)"},	**7.20961**	**7.67891**	জনসংখ্যা উন্নয়ন সংস্থা — distractor একটু বেশি কাছাকাছি।	**55.43%**	**6.36**	
			{"question": "জাতিসংঘের সদর দপ্তর কোথায় অবস্থিত?","correctAns": "নিউইয়র্ক"," distractor1": "জেনেভা","distractor2": "হেগ","distractor3": "ব্রাসেলস"}	**5.75377**	**8.91234**	সদর দপ্তর — মোটামুটি সহজ।	**73.11%**	**14.9**	
		"বাংলাদেশের তথ্য অধিকার আইন"	{"question": "বাংলাদেশের সংবিধানে কোন অনুচ্ছেদে চিন্তা, বিবেক ও বাকস্বাধীনতা নাগরিকের মৌলিক অধিকার হিসেবে স্বীকৃত?"," correctAns": "৩৯","distractor1": "৩৮"," distractor2": "৪০","distractor3": "৪১"},	**9.44903**	**7.34567**	অনুচ্ছেদ সংখ্যা — ঘনিষ্ঠ distractor।	**54.92%**	**2.78**	
			{"question": "তথ্য অধিকার আইন কবে জারি করা হয়?","correctAns": "৫ এপ্রিল ২০০৯"," distractor1": "৫ জানুয়ারি ২০০৯"," distractor2": "৫ মে ২০০৯","distractor3": "৫ ডিসেম্বর ২০০৯"},	**3.42935**	**8.45678**	জারির তারিখ — সংখ্যা ঘনিষ্ঠ।	**60.55%**	**17.71**	
			{"question": "তথ্য অধিকার আইনে 'তথ্য' বলতে কী বোঝানো হয়েছে?","correctAns": "যেকোনো ধরনের রেকর্ড","distractor1": "শুধুমাত্র লিখিত দলিল","distractor2": "শুধুমাত্র দাপ্তরিক নথিপত্র","distractor3": "শুধুমাত্র অডিও-ভিডিও সামগ্রী"},	**5.39428**	**8.12345**	তথ্যের সংজ্ঞা — বিশ্লেষণমূলক।	**69.71%**	**12.81**	
			{"question": "তথ্য অধিকার আইনে কতজন তথ্য কমিশনার থাকেন?","correctAns": "একজন প্রধান তথ্য কমিশনারসহ দুইজন"," distractor1": "একজন","distractor2": "দুইজন","distractor3": "তিনজন"},	**6.80393**	**7.78912**	তথ্য কমিশনার সংখ্যা — ঘনিষ্ঠ সংখ্যা।	**62.35%**	**6.58**	
			{"question": "তথ্য অধিকার আইনে, যারা লেখাপড়া জানেন না তাদের জন্য কে সহযোগিতা প্রদান করবেন?","correctAns": "দায়িত্বপ্রাপ্ত কর্মকর্তা","distractor1": "তথ্য কমিশনার"," distractor2": "আপিল কর্তৃপক্ষ"," distractor3": "তথ্য কমিশন"},	**8.52551**	**8.67891**	সহযোগিতাকারী — অর্থবহ distractor।	**41.12%**	**6.18**	
			{"question": "তথ্য জানার জন্য কিভাবে আবেদন করতে হবে?","correctAns": "লিখিতভাবে বা ইলেকট্রনিক মাধ্যমে"," distractor1": "শুধুমাত্র মৌখিকভাবে"," distractor2": "শুধুমাত্র ফোনে","distractor3": "শুধুমাত্র ব্যক্তিগতভাবে"},	**8.22377**	**8.34567**	আবেদন পদ্ধতি — মোটামুটি সহজ।	**54.42%**	**4.03**	**3.99**

			{"question": "দায়িত্বপ্রাপ্ত কর্মকর্তা যদি নির্ধারিত সময়সীমার মধ্যে তথ্য প্রদান না করেন, তাহলে অনুরোধকারী কী করতে পারবেন?"," correctAns": "আপিল করতে পারবেন"," distractor1": "সরাসরি তথ্য কমিশন-এ অভিযোগ করতে পারবেন","distractor2": "অন্য মন্ত্রণালয়ে অভিযোগ করতে পারবেন"," distractor3": "কোনো পদক্ষেপ নিতে পারবেন না"},	**7.64724**	**7.91234**	আপিল ব্যবস্থা — বিশ্লেষণমূলক বিকল্প।	**56.76%**	**6.14**	
			{"question": "তথ্য কমিশনের প্রধান কাজ কী?","correctAns": "অভিযোগ গ্রহণ ও প্রয়োজনীয় ব্যবস্থা নেওয়া","distractor1": "তথ্য প্রদান করা","distractor2": "আইন প্রণয়ন করা","distractor3": "সরকারি কর্মকর্তাদের প্রশিক্ষণ দেওয়া"},	**8.73489**	**8.56789**	প্রধান কাজ — অর্থবহ distractor।	**52.05%**	**4.22**	
			{"question": "তথ্য অধিকার আইনের মাধ্যমে কোন ধরণের মানুষের উন্নয়ন নিশ্চিত করা সম্ভব?","correctAns": "দরিদ্র, প্রান্তিক এবং সুবিধাবঞ্চিত মানুষের","distractor1": "শুধুমাত্র সরকারি কর্মচারীদের","distractor2": "শুধুমাত্র ব্যবসায়ীদের","distractor3": "শুধুমাত্র রাজনীতিবিদদের"},	**4.36409**	**7.23456**	উন্নয়ন সুবিধাভোগী — মোটামুটি সহজ।	**73.48%**	**6.07**	
			{"question": "তথ্য অধিকার আইনের ফলে কোথায় স্বচ্ছতা ও জবাবদিহিতা বৃদ্ধি পাবে?"," correctAns": "প্রতিটি সংস্থার কার্যক্রমে"," distractor1": "শুধুমাত্র শিক্ষা প্রতিষ্ঠানে"," distractor2": "শুধুমাত্র রাজনৈতিক দলগুলোতে","distractor3": "শুধুমাত্র সরকারি অফিসে"}	**7.3908**	**8.78912**	স্বচ্ছতা বৃদ্ধি — বিশ্লেষণমূলক বিকল্প।	**70.76%**	**5.14**	