Efficacy of Multiple Curriculum-Based Large Language Models in Bangladesh's Education System

by

Aumio Rahman
22101855
Tanjila Mazumdar
22101037
Anika Afsara Suzana
22101194

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

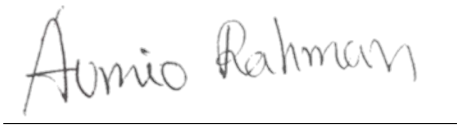
Department of Computer Science and Engineering
BRAC University
Fall 2025

Declaration

It is hereby declared that

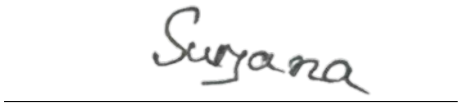
1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material 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:



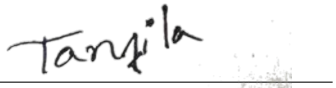
Aumio Rahman

22101855



Anika Afsara Suzana

22101194



Tanjila Mazumdar

22101037

Approval

The thesis/project titled "Efficacy of Large Language Models in Detecting Bias in Historical Textbooks Across Different Educational Curricula in Bangladesh." submitted by

1. Aumio Rahman (22101855)
2. Anika Afsara Suzana (22101194)
3. Tanjila Mazumdar (22101037)

In Fall 2024, it has been accepted as satisfactory in partial fulfillment of the requirement for the B.Sc. degree in Computer Science on January 28, 2026.

Examining Committee:

Supervisor:
(Member)

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

In Bangladesh, the education system follows multiple curricula, each offering different perspectives on culture and society. This research aims to evaluate the effectiveness of large language models (LLM) to identify biases in textbooks across different curricula in Bangladesh. As the various educational systems differ, there is a high possibility that students from different backgrounds may have diverse perspectives on culture and society. Henceforth, this study investigates the potential influences embedded in the curriculum that can affect young minds. Our goal is to achieve a comprehensive analysis of these biases, expecting insights that could inform curriculum development and promote balanced educational content in Bangladesh's diverse educational streams.

Keywords: Large Language Models (LLMs); Curriculum Development; Bias in Education; Bangladesh Education System; Textbook Analysis; Educational Equity; Curriculum Comparison

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Education is the backbone of development in a society where a particular worldview, together with the value system, is taught to an individual from very early childhood. The educational system in Bangladesh is heterogeneous. The National Curriculum and English Medium curriculum differ in their approaches to teaching social sciences, history, and culture. Different socio-political and cultural beliefs often stand behind such disparities in curriculum design. This diversity raises questions about possible biases in instructional materials for sensitive cultural and historical topics, in an effort to serve a range of audiences. Bias in educational content can greatly influence students' views of their cultural identity and concepts of history and societal values. Textbooks are significant sources of information and thus shape beliefs at a critical level. Any consistent narrative across curricula may lead to different perceptions of the same notion, bolstering divisions within students from all walks of learning. This study aims to measure how large language models evaluate textbook biases across various curricula in Bangladesh, with the goal of addressing the seriousness and significance of these issues.

1.2 Rational of the Study or Motivation

Recent breakthroughs in AI have bestowed LLMs with immense capabilities in the analysis of large volumes of textual data. In this paper, we use these technologies to explore how curriculum-specific textbook-trained LLMs can convey information about material quality, informativeness, sentiment, and possible biases. In order to uncover disparities in the problems of cultural and historical presentations, we integrate the two approaches with additional natural language processing tools. Its ultimate objective is to create educational materials that are more inclusive and balanced. The basis of this research is evidence-based recommendations for developing curriculum policies, which will be made to policymakers, educators, and other stakeholders in order to further the development of the curriculum. This study aims to bridge the understanding gaps in Bangladesh's educational streams by promoting a more unified, unbiased approach to learning educational content and inculcating a shared sense of cultural and historical identity.

1.3 Problem Statement

We aim to test the performance of LLMs in finding biases in textbooks from Bangladesh's National Curriculum and English Medium education systems. Our aim is to investigate the differences in representation, sentiment, and focus across these curricula by analyzing cultural and historical issues. This research will lead to the development of balanced and inclusive educational content, fostering a common understanding of cultural and historical narratives among learners with diverse educational backgrounds.

1.4 Objective

The objectives of this research are to identify the following:

1. The feasibility of large-language models: Establish how LLMs perform in terms of sentiment analysis, summarization, and bias identification using a set of textbooks prepared for a given curriculum.

2. To evaluate the curriculum-specific biases: Analyze the representation of historical and cultural topics in the different educational curricula of Bangladesh.

3. To assess the quality and informativeness of the content: Use LLMs to check for accuracy and consistency with established historical and cultural narratives.

4. To identify the patterns of representation: Identify and examine the representational patterns in textbooks that may help in reinforcing prejudice or biased perspectives.

5. To provide evidence-based recommendations: Evidence-informed recommendations are useful for curriculum developers, educators, and policymakers with an interest in moving toward more balanced and inclusive educational content.

The objective here addresses not only the efficacy of LLM in educational assessment but also feeds into the more general discussion of equitable and inclusive curriculum development.

1.5 Methodology in Brief

A novel approach to detecting bias will be developed using two large language models (LLMs). On its most basic level, this new method of bias detection accepts a user's query and identifies all possible biases in that query. The methodology used within this new method of bias detection involves retrieving the appropriate text chunks based on the input question from the textbook database. These text chunks are then presented to LLM1, which develops a response through combining information from those text chunks. It is also important to note that LLM1 produces a response that includes only the information found in the textbook and does not include any other information that may have been obtained from outside sources or

created by the LLM itself. The responses produced by LLM1 and the original query, are sent to LLM2. LLM2 analyzes the response to determine if it has included any biases. The analysis by LLM2 focuses primarily on three forms of bias: content integrity bias, framing bias and representational bias.

1.6 Scopes and Challenges

Our study focuses on using AI to identify educational biases in non-Western settings. While there are many studies that explored this, most used Western datasets, which lacked multilingual or multi-curriculum unique to the geopolitical and sociolinguistic premise of most Southeast Asian countries. This study will help shed light on two very important things: detecting biases using natural language processing and educational equity. For using natural language processing, we take a better look at how biases are detected and minimized by LLMs that have been trained on the local curricula. On the other hand, educational equity were judged on the presence of systemic bias in educational systems. We are hopeful that this technique can be adapted globally, and its application will impact other multilingual, post-colonial countries, like Bangladesh, the best. This offers a blueprint to build further upon since its application supports a plethora of curricula, like National, Cambridge, Edexcel, etc., which are currently being taught in Bangladesh. Moreover, our work can also help improve education by finding biases and linking them to suggestions, which can then be turned into proper actions by Bangladesh's National Curriculum Board and UNESCO's Sustainable Development Goal 4, Quality Education. Additionally, our study aligns with two very prominent debates in the field: using natural language processing to analyze textbooks and comparing textbooks of different curricula in Bangladesh. Our system makes this possible by focusing on detecting the dominance of Western culture in textbooks while also explaining why certain biases are flagged by LLMs, using attention mechanisms. One of the challenges was collecting textbooks from different curricula and digitalizing them due to their primitive format or even the inclusion of politically sensitive information. At this stage, a substantial amount of fieldwork and collaboration was crucially maintained. Moreover, we had to ensure the academic standards were upheld while the results were true to the locality. Finally, because the system is based on user-generated queries, we had to keep our innate bias in check while developing them.

Chapter 2

Literature Review

2.1 Preliminaries

To start with, we started our search with keywords such as 'Curriculum analysis using NLP', 'Textbook analysis using NLP', 'Bias detection in text using NLP', 'Textbook analysis' and 'Textbook analysis using NLP in Bangladesh' in IEEE, ACM, Semantic Scholar, Google Scholar, etc. to find papers related to our research interests. For papers that are related to technology, we tried to have papers published after 2020 to have a relevant idea about today's technology. For papers related to social science, we tried to collect papers from at least 2015 onwards. At first, we collected a vast amount of papers, and after carefully reading the abstracts and conclusions, we shortlisted about 10 papers that are the most crucial for our research. We carefully reviewed every paper to ensure we obtained important information that would influence our future work. This technique helped us gather papers that are the most relevant to our research. We went through the introductions, data, methods (if applicable to our studies), findings, and conclusions of these 22 papers and summarized their key concepts. Then we categorized these papers into different themes, for example, foundational knowledge, methods, global context, local context and prepared them for the literature review. Moreover, in this phase, we conducted essential courses on deep learning, natural language processing and artificial intelligence to establish a solid foundation in these technologies. These courses covered fundamental concepts, including neural networks, sequantial models, and natural language processing techniques. Gaining this knowledge helped us build a strong foundation in these fields and enhanced our ability to comprehend the research papers we reviewed for our literature review.

2.2 Review of Existing Research

A paper [10] by Zhai et al. covers the use of AI in education, considering 100 related articles published from 2010 to 2020. It identified four key trends of development in AI technology, including deep learning, swarm intelligence, neurocomputation, and IoT, that have a great potential for teaching methodology enhancement. The main trends related to applying AI techniques in personalized learning and adaptive feedback systems using new technologies like IoT and deep learning. The studies were divided into four categories based on the goals of AI systems: deep learning, recommendation, matching, and classification. The study found several encouraging AI

trends in education, including: Internet of Things: IoT has very minimal application in education, although it can enhance students' spatial and cognitive understanding in a few subjects like science. Its relatively low cost and compatibility with wearable technology could support financing in the development of new learning tools. In swarm intelligence, collaborative knowledge-creation learning is promoted, changing the role of teachers from knowledge transmitters to organizers of knowledge. Neurocomputation and Deep Learning: Deep learning, when combined with neuroscience, will develop better human-computer interactions by learning from big data. These models of AI working together with models based on the brain will maximize adaptive learning systems toward more personalized learning experiences. Examples of AI applications that foster emotional and cognitive development include gamification, wearable technology, and immersive learning environments. Feedback and Interaction: AI-based systems are being used more and more to provide instant feedback in order to create the best possible learning environments or to help students move through challenging assignments on their own. Most AI systems developed so far lack domain specificity, which restricts their functionality in personalized learning. Future research should focus on refining AI applications, improving teacher adaptation, and developing better evaluation methods.

In a paper [3], Vaswani et al. introduced the transformer model, which completely revolutionized the way sequential models are used in the field of natural language processing. This successfully eliminated the limitations faced by existing models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTMs). Not only did RNNs and LSTMs face difficulties in memorizing and modeling long-term dependencies, but they also lacked parallelization due to sequential computations. Transformers were able to overcome this with self-attention by assigning different priorities to various parts of the input sequence. Moreover, transformers use positional encodings for sequential processing. This works by embedding position information into data. Lastly, transformers allow parallel computation, which makes large-scale data analysis possible. Transformer uses an encoder-decoder architecture in which the encoder processes the input, the decoder focuses on the encoded input with the help of a cross-attention layer, and generates the output. This paper also has had a significant impact on the field of AI and machine learning. This model is considered to have advanced performance in machine translation, as seen in the WMT 2014 English-to-German and English-to-French benchmarks. Additionally, it is faster to train and has made the development of AI models like BERT, GPT, and T5 possible. The introduction of attention-based models has greatly impacted the research of artificial intelligence. On top of that, large language models have been able to take off due to the potential for parallelization. Hence, it is obvious how crucial the transformer's architecture is to the advancements in modern AI.

Later on, in a paper [15] by Patwardhan, Marrone and Sansone extensively discusses transformer based models and their application in the field of NLP. These applications can be sectioned off into two distinctive categories: unimodal or text-focused and multimodal or combination of text, images, and audio. Unimodal application is used vastly in sentiment analysis, text summarization, and information retrieval. Sentiment Analysis is used to analyse emotional tones and languages which are subjective in nature. This can be helpful in detecting any existing biases in text-

books. For this, models such as T5 (often used in a zero-shot manner) and XLNet are preferred. It helps to evaluate the eligibility and the information provided in textbooks. Longformer (use pooled attention), BERT, LinkBERT (Extractive summarization methods), mBART (used for generating concise summaries) are the models mentioned for text summarisation. For checking facts (from trusted sources such as books and papers) against the provided information in the textbooks, and for comparing the contents Information Retrieval can be utilised. An Example of an efficient model for this is RoBERTa, which is being heavily popularized. Another aspect is Readability Assessment, which helps by assessing the complexity of the texts to make sure that the consumed content is age-appropriate. The model DistilBERT was chosen for its ability to do large scale analysis. The authors discuss multimodal applications too where they mention StableDiffusion, which can be used to generate images from texts. However, there are also a few drawbacks. Firstly, there is an issue with interpretation with the transformers when they are applied. They have a difficult time explaining the reasoning behind their classifications (i.e. biases or inaccuracies in the content). Secondly, the cost for computation and resources can start adding up and become expensive to operate. Lastly, these models can have pre-existing biases from the data they received for training. This can cause a potential problem since it cannot accurately identify them. This can be avoided by curating proper datasets for training. To achieve transparency and hold oneself accountable, while producing accurate results, the paper highlights the importance of open-source models. The authors suggest that ethical consideration should be a high priority since culturally sensitive educational materials are being evaluated.

In the same year, a paper [16] by Zhao et al. discusses some of the recent works in large language models, specifically those with over 10 billion parameters, focusing on pre-training, adaptation, usage, and evaluation. It identifies that LLMs are unsupupervisedly pre-trained on massive amounts of text data via methods such as next-token prediction. Most previous multitask learning methods are towards broader task coverage, which is assumed to result in better generalization. The main architecture used for LLMs is the Transformer, which has proved to be highly scalable and effective. Despite such advantages, the Transformer architecture suffers from several open challenges, the most important of which are its high cost in training and slow inference. Also, long context modeling, while enabling the LLMs to take in large inputs, for example up to 200K tokens with models such as GPT-4 Turbo and Claude 2.1, still has its own problems regarding efficiently exploiting the information present in these long contexts. Advanced ICL and CoT prompting have considerably improved both the reasoning and problem-solving capabilities of LLMs and are, on occasion, capable of outperforming even their full-data fine-tuned model variants. Furthermore, RAG is another mechanism to enhance task performance that integrates external knowledge and extends the model's knowledge base. Even so, other critical challenges facing LLMs remain those related to prompt efficiency and inference costs, which are key factors in large-scale deployment. Another major challenge involves the safety of LLMs. Owing to their probabilistic nature, LLMs tend to hallucinate-that is, produce plausible but incorrect information-which raises concerns regarding their application in sensitive or high-stakes settings. Such techniques are being developed to align the models with human values, including reinforcement learning from human feedback. RLHF is resource-intensive and heavily

dependent on high-quality human feedback. LLMs are already changing the world in industries involving search engines, recommendation systems, and the development of autonomous AI agents. However, there is a need to overcome scaling challenges, inefficiency in pre-training, and the development of more efficient alignment methods. Future work should focus on enhancing the efficiency of the models, considering other architectures, refining pre-training methods, and aligning LLMs with ethical standards and human values.

In 2023, a paper [14] by Mukherjee et al. inquiries the societal biases in Large Language Models among 24 languages, accentuating the southern part of the world. Though numerous researches are being conducted to identify and mitigate bias, there are also some shortcomings in the applied methods of detecting and reducing bias. Starting with the Word Embedding Association Test (WEAT) which solely focuses on the English language and the cultural context of the northern region of the world. In addition, static word encoding methods like FastText, GloVe are being used instead of contextualized model Bert. Therefore, to deal with these problems the study points out 5 contributions i.e. in the beginning form a comprehensive dataset and for in depth analysis analysts compared machine versus human translation in 24 languages across all relevant categories of WEAT. Secondly, to eradicate ambiguity in handling words which are multi-expressional in other languages but have only one word to express in English, WEAT metrics have been reformulated in a systematic manner. Thirdly ,to address the unaddressed human created biases, analysts proposed five dimensions such as toxicity, ableism,education, sexuality and immigration. Fourthly, comparing different techniques to capture cultural biases in different contexts. Finally, the authors focus on Indian languages by conducting in depth analysis in 7 Indian languages to find regional biases. Basically, the authors introduced WEATHub for measuring bias in multilingual setup. Moreover, the study contributes 5 ways to identify unaddressed human created biases and focusing regional biases across 7 Indian languages. However, in the method section the authors reformulated the WEAT and developed the heuristic to evaluate biases in contextualized and multilingual setup. For measuring biases human translation is more accurate than machine translation(MT). There are also some limitations of this study. Firstly, most of the languages had two annotators whereas some have only one. Secondly, since the authors covered only 24 languages, this study is not applicable in every language. Lastly, WEAT has some limitations as it cannot create a bias free model. As for results, cultural bias monolingual models show better results, training dataset with biased information causes bias and more research should be conducted to find the origin and implications of biases. In conclusion, this study underscores the cultural or societal biases across 24 languages by applying methods like WEAT metrics. However, WEAT is not a perfect tool to measure these biases hence this method needs further improvement.

After a year in 2024, a paper [20] by Polak et al. presented new innovative techniques to extract data automatically from research articles using large language models (LLMS) with the supervision of humans. The advantages of this method are adaptability, robustness, and it does not require maximum prior knowledge and requires only minimal knowledge in coding. Consequently, this method is compatible for mid-sized databases with improved precision. However, for dataset science

journals, articles, reports etc. Open datasets for this specific purpose are also used. Generally, LLMS, such as GPT-3, GPT-3.5/4, DeBERTaV3 and BART are used methodologies combined with human supervision. Additionally, for reducing time of text's manual labelling semi-automated pipelining is introduced. Moreover, for identifying materials and properties, Named Entity Recognition(NER) is used. Also for evaluating the metrics F1 scores, precision and recall are applied. This study points out LLMS can be used for high precision data extraction with minimal customization of the model. Since humans are involved in the process, this applied model also takes care of the nuances of the text and domain specific terms. This study highlights the effectiveness of LLMS in extracting unstructured data from the text automatically and also shows that the data extraction can be 90% precise with 96% recall. However, this method depends on human supervision for high quality data; the models used to extract material related data are not trained for the specific purpose; poor error handling; presence of bias in dataset; sometimes fail to give high precision output; cannot handle large sized databases; may not work properly for extracting other types of material science data; for large sized data extraction this method can be expensive; also how well extracted data align with original text. To conclude, this study illustrates the robustness, adaptability of the framework for extracting structured data from science based text using LLMs with the help of human supervision . Although LLMs extract data with fewer errors, it needs further development for future progress.

In the same year, A paper [21] by Shina Raza et al. outlines numerous methods that may be used to detect and remove biases existing in textual data. Biases in any form can enforce harmful stereotypes, which can lead to poor judgment. In order to avoid this, the authors provide Nbias, a paradigm that encourages the ethical use of data to identify and eliminate biases. The Nbias system consists of four components: data collection, corpus generation, bias detection model, and performance assessment. First, information is collected from a range of sources, such as healthcare websites, social media platforms, and employment portals. Only the relevant data is then taken out of the data collection process to create a complete dataset, which is then utilized to train the model. A transformer-based token model is then used to categorize words or phrases that are deemed biased. The categorization is done under the entity 'BIAS'. Lastly, the effectiveness of the framework is determined by both qualitative and quantitative metrics. Following text analysis, the Nbias classifier correctly detects various biases throughout a range of platforms, resulting in 1-8% better accuracy ratings than other standard models. Nbias can lessen prejudicial beliefs and stereotypes because of its adaptability and potency in minimizing bias in textual data. Nbias has its own limits, too, when it comes to handling small biases and detecting them in various linguistic and cultural situations. The authors encourage future research to use larger datasets to increase the accuracy of the framework and investigate practical uses in order to further develop the design.

However, extracting data from texts while measuring cultural values can be a tricky thing to do. A paper [17] written by Adilazuarda et al. summarizes how cultural inclusions are assessed in the case of large language models by identifying research gaps and limitations while capturing nuanced differences in culture in 90 NLP stud-

ies. Often, LLMs fail to represent cultures of marginalized peoples, especially when digital spaces reproduce biases. The current research reviews black-box probing techniques. These are two types: the discriminative kind or multiple-choice, with cultural context, and generative, or free-text generation, on datasets such as the World Value Survey (WVS) and SeeGULL to check the biases of LLM-generated outputs. Among the key findings are that LLMs tend to default towards culturally dominant perspectives, perform badly when using prompt-sensitive probing methods, and mainly center on values and norms while leaving out other aspects of culture, including kinship, spatial relations, and linguistic diversity. Furthermore, the absence of multilingual datasets further constrains non-Western cultural analysis; few works actually assess real-world applications, such as content moderation, education biases, and interaction among users. This paper seeks to address pressing issues: the need for a range of representative datasets and an interdisciplinary approach in research that brings NLP together with anthropology and human-computer interaction, and a sound evaluation methodology that allows for enhanced interpretability and generalization. Major challenges include a lack of a unified framework defining culture in NLP, limited studies that involve underrepresented cultural dimensions, and further development of models that are adaptable in various cultural settings. Moreover, most of them are based on English-centric datasets that impede cross-cultural understanding; therefore, the need to develop culturally rich multilingual corpora is a pressing one. There is an urgent need to transcend simplistic probing approaches and to utilize white-box interpretability techniques to better reveal model biases as well as limitations. Future work should integrate external cultural knowledge sources into LLMs, keeping this development process fair and just. Among these developments in shaping digital interactions, it is necessary to address these gaps to prevent AI from either reinforcing the biases or misrepresenting the issues at hand. This will lead to more inclusive approaches on a global scale, enabling AI systems to better serve diverse populations by promoting fairness, accuracy, and cultural sensitivity in automated decision-making.

Large language models are a useful tool for identifying biases in text, but they are not immune to bias either. A paper [19] by Gallegos et al. examines social bias in large language models and ways to mitigate those biases. Large language models are generating human-like texting, processing the texts , understanding the text and doing many other advanced works related to text continuously .In spite of its usefulness it also has some drawbacks for example , creating harmful societal biases which is an ethical concern . This paper focuses on demonstrating these biases and solutions to prevent the biases. To evaluate bias the first needed thing is metrics for evaluating bias, secondly, the dataset required for bias evaluation and lastly the techniques to mitigate or reduce the bias. The writers of this paper used fairness metrics to evaluate biases in large language models or nlp(natural language processing). Additionally, for mitigating the bias there are 4 stages such as : pre-processing, in-training, intra-processing, and post-processing. The writers categorized biases in social,algorithmic,cultural and data sections. Also they analyzed the sources or cause of these biases. Firstly, during training the datasets some information or data is overrepresented while others are under represented. Secondly, annotation bias which occurs during data labeling. Thirdly and lastly, during transfer or customization bias gets amplified. Basically this book points out the biases of LLMS (large

language models) and various ways and strategies to prevent or reduce it. However, there are some challenges and limitations of this research paper. In the first place, address the limitations of theoretical research that guarantees fairness. Secondly, the standards and principles for evaluating bias need to be more appropriate and globalized. Lastly, for mitigating the bias more technological and theoretical advancements are needed. Also, the power imbalance in LLM is also a concern. To conclude, this paper focuses on promoting the mitigation of bias and addressing the ethical concern regarding those arising biases in the LLM. Consequently, to reduce these biases various solutions and methods have been mentioned.

In the context of education, a paper [11] by Baker, R. S., and Hawn, A. point out the biases of algorithms in educational settings due to poor representation of data and how this technological bias is exacerbating the existing bias in education. This paper also examines the impact of these biases and procedures to detect these biases. Overall, this study proposes and instructs how to shift bias to equity. As data collection plays an important role in removing bias, implementing this step properly is a crucial task. The writers used different materials i.e. research papers relevant to the topic (Algorithmic Bias), analyzed data from many educational platforms to identify the biases' pattern and case studies that show the technological discrimination against the students are also used. Also, for datasets data has been taken from different gender, race and ethnicity, socio-economic status and from groups of people with disabilities and military connections. Generally, following methodologies like, machine learning pipeline,algorithms and metrics to assess and remove bias has been used. This paper brings out the algorithmic biases present in the education system across different fields i.e. gender,continental,social-economic class, race,nationality etc. Because of these biases algorithm gives more priority to a certain group or individual whereas the algorithm gives everyone equal oppurtunities. Findings also demonstrate that algorithmic bias is far-reaching in educational technology. Furthermore, there are gaps in research which includes indigenous people and minorities as the Ai models prioritize the US. Basically This paper implies how the learning environment can get affected by the algorithmic biases. To solve this bias these steps can be taken i.e.collecting data from diverse groups, testing algorithms rigidly for fairness, confirming tools are correctly designed to ensure equity among students, recruiting researchers from diverse backgrounds to ensure human bias cannot be created. These actions will bring out the full capability of technology in education. However, the limitations are also present in this study, for examplesome of the datasets do not cover the diverse group so it has a lack of data, trade off between exactness and fairness of the models, lack of standardized framework to measure fairness etc. Steps proposed by the writers are- bias detection tools, tools to create fairness,models built for specific domains like rural or urban areas, collaboration between educators, researchers and technologists for further development. Overall, this study aims to prevent educational bias in technology by proposing various methods and creating awareness through this study.

2.3 Summary of Key Findings

This study examines the efficacy of multi-curricula-based LLMS in detecting biases within the educational systems in Bangladesh. Firstly, we found that biases exist

in textbook across the globe. These biases are mostly caused by the different ways that historical, social, or cultural contexts are taught. Misinterpretation of various contents, be they cultural, global, gender, race, or religion, in the textbooks is noteworthy. In the context of countries that have been colonies, one of the mentionable ruling biases that exist in these textbooks is western worship and residual imperialism of westerners, and Bangladesh is no exception. From our literature review, we found that textbooks in Bangladesh, specially those in the English and national curricula, portray cultural values differently. Not only do the textbooks across three different streams contain different perspectives on the same concept, but also textbooks from different curricula contain biases. Because of these biases, due to limited knowledge, students' perception about a certain concept gets hindered. Unfortunately, the arised biases also create social division, a national identity crisis, low tolerance levels, etc., and indeed it can impose a direct threat to national unity. One of the key problems in detecting these problems is limited manpower, as these people must carefully examine every line in the textbooks. Moreover, humans are not free from bias, which can further impact the analysis. We find that large language models can be a great tool to detect biases as they can analyze vast amounts of data, which is not humanly possible. However, most bias studies focus on western datasets, and analyzing textbooks with natural language processing is a fusion rarely explored in non-Western contexts. Despite LLMs being a valuable tool to enhance the quality of education, they are also not free from bias. In the literature review, we mention some algorithmic bias in the LLMS model; however, carefully training these models can greatly reduce these biases. Overall, we are optimistic that our research can bring about significant advancement in the educational system by integrating inclusivity and balance in the textbooks of all curricula in Bangladesh. Additionally, it will also serve as a blueprint for other post-colonial countries that also have different curricula, like Bangladesh.

Chapter 3

Requirements, Impacts and Constrains

3.1 Final Specifications and Requirements

This section delves deeply into both functional and non-functional requirements, project constraints, and the standards and practices that must be upheld when utilizing LLM.

3.1.1 Functional Requirements

- **Accept raw textbook questions:** Our system should support the natural-language questions entry of historical books (e.g. "Write about the role of women in 1971 liberation war."). It should handle complete phrases in the textbook language (English) without exception. This indicates that the input interface is merely a text-based question-taking interface that transmits the queries to the pipeline.

- **Conduct semantic retrieval with embeddings:** The textbook is in a vector-based RAG architecture. When a question is entered into the pipeline, it produces a query embedding and searches the corpus of textbook with a similarity test content. The system locates textbook passages whose vector representations most closely match the query using neural embedding models. Each chunk is embedded and the most semantically relevant chunks are brought out as chunked grounding. This guarantees that the LLM only sees text that is directly related to the question, resulting in more targeted responses.

- **Produce grounded answers (LLM1):** The first language model (LLM1) receives the question of the user along with the passages of context retrieved and produces an answer. Instead of depending only on its own knowledge, the LLM generates responses that are "grounded" in the actual content by providing specific textbook passages as extra prompt context. In order for LLM1 to cite evidence from the textbooks, the retrieved passages and the question are combined (with meticulous prompt design). It has been proven that this RAG strategy increases factual accuracy and decreases fabrication(Jain et al., 2025) [24].

- **Detect bias (LLM2):** A second LLM (LLM2) looks at the content (the generated answer from LLM1) to determine biased texts or opinions. As an example, the recent studies of the tool called BiasAlert, described in recent studies employs this two-step mechanism: one retrieves templates or definitions of bias that are relevant first and has an LLM rate whether or not the material is biased in that context [18]. Our system also directs LLM2 with clear goals of historical bias (e.g. factual bias, nationalist bias) and make it give a bias classification and explanation of each answer or text fragment.

- **Measure consistency, IRR, and hallucination:** We measure consistency by running a single query ten times and then comparing the results using all-mpnet-bias-v2 to see if they are semantically identical and refer to the same historical events. We have a third LLM as a judge who checks if LLM1's answer and LLM2's analysis are symantically correct using all-mpnet-base-v2 to ensure interrater reliability. We also tested it for various analyses generated by different models. To check for hallucinations, we used the same sentence transformer to see if the phrase used by LLM2 was present in the chunks retrieved from the book via LLM1.

- **Support human expert commentary:** The design has to support domain experts (e.g. history professors or education specialists) to check the outputs of the model and amend them. Indicatively, specialists ought to be capable. For example, experts should be able to see the answer, the highlighted evidence, and the biased judgment. They can then give feedback or change the model's output. Having a human in the loop is essential. One study found that when experts could edit assessments generated by LLMs, accuracy improved, and the need for perfect model answers decreased [23].

3.1.2 Non-Functional Requirements

- **Explainability:** Each automated decision - answer content and bias flag - should be elaborated by description or reference. The system ought to clearly indicate the passages within the textbooks, which it relied on to establish an answer or bias. Formatted results (e.g. " Score:.. Detailed Discription: ...; Keywords: ..."). This is a best practice of XAI: in one research, the results of LLM in color-coded conclusions, textual explanation, and direct quotes were highly helpful in the interpretation and trust of the human being [23].

- **Reproducibility:** All the prompts are logged, random seeds to be embedded. We need to maintain consistent sample sizes, prompt designs, and LLM configurations across trials, and publish the code and configurations. This way, other people can re-run the experiments and obtain the same results. Recent advice to the assessment of LLM focuses on the development of systematic frameworks that can be used to improve scientific rigor and reproducibility [26].

- **Scalability:** The system supports large volumes of textbooks using efficient vector databases (e.g., ChromaDB) and batch processing.

- **Low hallucination rate:** The system should be reduce producing any fabricated claims. One main tactic is to use RAG grounding, which allows us

to "open the book" by feeding actual textbook passages to the LLM instead of relying solely on training data.[3] By adjusting the temperature of generation (to lean towards determinism) as well as post-generation verification (e.g. checking selected facts against the text that has been retrieved), we can further ensure that there are also low levels of hallucinations.

- **Model-agnostic design:** The design should not be bound by a specific patented LLM. The system is designed in such a way that any other model (Qwen, open-source LLaMA, etc.) can be used in the place of either LLM1 or LLM2 with minimal modifications. As an illustration, we can update the model type or provider by using regular APIs and prompting instead of using special internal capabilities. (Practically, most RAG models could replace any language model with generation; as we train prompts and embeddings, they are also model-agnostic.)

3.1.3 Constraints

- **Large language models are stochastic:** This is inherent in large language models. Similar prompts can give a different response when run twice. This inconsistency is capable of interfering with consistency. We alleviate it through low-temperature setting (to decrease randomness) or multiple querying (an ensemble of responses).

- **Token limits:** Textbooks are vast, and the window of context of LLM is limited (e.g. 8k-32k tokens). Long passages must be broken into bits that fits the model. This is done naturally in the embedding and retrieval process: chunks of the textbook are of small size; therefore, when context is fed to the model, they remain less than the token limit. It can also be reduced by varying the length of each chunk (10-15 sentences per chunk).

- **Expert accessibility:** Ground truth labels and assessments are based on expert ratings and information from history educators, who might not have much time to guarantee the pipeline. This restricts how many and when we can request reviews.

- **Bias definition drift:** The concept of bias in historical accounts is subject to change or shift according to the angle. One generation may view something as biased, while another may view it as balanced. Our bias criteria may need to be revised because of this drift.

3.1.4 Standards and Practices

- **Academic bias literature:** We base our definition of bias on the academic textbook bias literature. For instance, there are documented instances of politically motivated historical revisionism [22] and systematic gender stereotypes in Bangladeshi textbooks [4]. These pieces inform our taxonomy (e.g., "marginalized women bias" ,"politicized omission of events"). We refer to and match up with known results on educational bias.

- **Accountable artificial intelligence:** Our architecture is based on the principles of responsible-AI ethics and legal requirements, such as fairness and transparency. We guarantee the usage of the AI is recorded and the possible harms (e.g., the increase of bias) are prevented. This implies a proactive check into the system to identify any systematic errors, audit into all data sources to identify diversity and explicitly hold people accountable (e.g. logs of decisions). IBM's guidelines emphasize that ethical principles must be incorporated into AI development in order to minimize risk and maximize benefits, which is a requirement of our bias-detection pipeline [28].

- **Explainable AI (XAI) norms:** The system adheres to explainable AI principles by providing traceability to source textbook passages and human-readable explanations for generated answers and bias classifications. These explanations are designed to support interpretability and trust for educators and domain experts [28].

3.2 Project Management Plan

This section examines the various stages, beginning with data collection, along with the timetable we followed and the resources we used.

3.2.1 Phases

i. **Data ingestion:** This research involved collecting key data from social studies textbooks of the National Curriculum and Textbook Board (NCTB) [30], Cambridge Assessment International Education (CAIE) [29], and Edexcel [31], major curricula in Bangladesh. Access to these materials was through the publishers' official websites. The NCTB textbooks were obtained in PDF format, while Cambridge books were automatically delivered digitally, and Edexcel materials were printed copies. Data processing included OCR using Tesseract to convert scanned PDFs to text, with careful error correction to maintain fidelity to originals. Texts were segmented into paragraphs (5-15 sentences each) for RAG evaluation, using the SentenceTransformer model to convert cleaned paragraphs into usable vectors, facilitating efficient information extraction.

ii. **Retrieval pipeline:** The RAG design was formulated, including textbook chunking and data cleaning with precise semantic embeddings generated using the SentenceTransformer library and the all-mpnet-base-v2 model. Each paragraph chunk was processed to create embeddings, stored in JSON files, and then added to a continuously operating ChromaDB using its library.

iii. **LLM integration:** Combine the language models and design prompts. We configured LLM1 (e.g. Gemini- 2.5 -flash via API) that responds to the queries based on the retrieved context and LLM2 that classify the bias. This was achieved by repeating prompt templates until they are accurate. To detect bias, we refined the prompt wording to detect various types of bias. The outcome of this phase is some working QA+Bias pipeline: a user poses a query, LLM1 responds, and LLM2 reveals bias.

iv. **Assessment:** We executed the performance in a systematic manner. To quantify the accuracy of the bias detection answers, we appointed a third LLM as a judge. Human experts reviewed LLM2 bias detection and as well as examined hallucination rates. This entails random evaluation of responses to textbook material. We also performed poison testing to determine whether LLM2 can detect bias in fabricated bias-induced data.

v. **Refinement:** On the basis of evaluations, we made refinements of the system. This was done by changing the retrieval (e.g. by changing the size of a chunk), by re-tuning prompts or by broadening the definition of the bias with additional examples. We also included the expert feedback. Several evaluation and tuning exercises might be required to achieve our accuracy requirements.

3.2.2 Resources Used

OCR tools: As our main method of converting scanned pages to text, we utilise the open-source OCR software (e.g. Tesseract) but with a manual cleanup process. The raw text is also preprocessed using standard NLP tools like tokenizers and regex.

Large language models (LLMs): We made use of pretrained LLMs, one or more. As an example, Gemini-2.5-flash, Llama, Mistral, and DeepSeek were experimented with as both LLM1 and LLM2. The models were accessed through APIs or local inference engines in a modular manner.

Compute resources: The significant compute requirements are inference of generation and LLM. A typical GPU or even CPU on smaller models is required to make embeddings, but an NVIDIA GPU (e.g., A100 or RTX) can be used to encode thousands of text chunks and be fast. As with any LLM inference, there is no requirement to spend on on-site compute (e.g., Google Gemini AI, Groq AI), or a large-memory GPU or TPU can be used to serve a local model. A vector database service (self-hosted or cloud) used to save and query embeddings.

Human experts: Two or three content experts, such as sociologists or historians from Bangladesh, were contacted. They assisted in bias annotation guidelines, inspected model outputs and gave ground-truth judgments.

3.3 Societal Impact

The study provides insights into the issue of educational equity by allowing the methodological and unbiased examination of historical textbook bias. As textbooks are the most influential factor in creating the perception of students regarding history, identity, and national narratives, bias identification can enhance the minimization of unfair or deceptive portrayals that can put some groups at a disadvantage [9].

The suggested framework will aid in minimizing cultural misrepresentation in the educational content by pointing to cultural, nationalist and representational biases. Curriculum reviewers may also use automated but expert-congruent bias analysis to help them identify covert framing and attribution problems that remain unseen in

manual review procedures especially in post-colonial and conflict-sensitive historical contexts.

It is also suggested that the study builds trust in AI-aided curriculum review because it focuses on retrieval grounding, bias types defined by experts, and human-in-the-loop validation. The system is not meant to substitute human judgment but offers educators and policymakers explainable and auditable outputs to overcome the frequent concerns regarding the opaque or unverified AI decision in education.

Meanwhile, this study is aware of the danger of excessive automation. Detection of bias in history is a process of interpretive judgment, context, and knowledge in discipline, which cannot be completely defined by machine software. This model views AI as a decision-support system, but not as an official assessor because it confirms the need to still have human oversight in case of abuse or overpopularization of automated bias tests.

3.4 Environmental Impact

The suggested framework is based on a number of design decisions that can help minimize the overhead computing. First, reuse also removes a lot of energy usage through redundant vectorization of the same textbook material. After creating embeddings with all-mpnet-base-v2 they are saved and reused among several queries, preventing repeating model inference and also closely corresponds to the concept of energy-efficient or "Green AI" practices [6].

Second, text chunking reduces processing to semantically important parts of reading as opposed to whole textbooks. The system can also bypass the reprocess of full books per query, as only the most relevant chunks are accessed and processed therefore lowering the token count and computational processes. This retrieval-based approach is aligned with the existing literature that demonstrates selective context retrieval has been known to enhance efficiency, without reducing performance [7].

Nonetheless, the framework also adds multi-model evaluation, which adds costs incurred in the process of inference, owing to the fact that multiple LLMs are used in the process of generating answers, detecting bias, and validating them. This is a compromise in computational efficiency and the reliability of the analysis. Although the multi-LLM design has higher energy consumption than single-model systems, it allows more reliable, interpretable, and credible bias estimates.

3.5 Ethical Issues

Credibility and integrity of educational research is a major concern that revolves around ethical aspects when the research topic is concerned with historical narration and curriculum analysis. The ethical misconduct which may be in the form of misrepresentation of data, selective reporting or lack of transparency could negatively affect the research validity and cause lack of trust among people [5]. The same concerns are heightened in the research associated with textbooks since its

results can affect not only the academic discourse but also the educational policy and classroom activities.

The danger of exaggerating the bias and incorrectly labeling prejudice in the analysis of educational content is one of the primary moral issues to be considered in the given research. Textbooks are important in forming the identity, culture and history perceptions of learners. Such misrepresentation or a biased presentation in textbooks may support stereotypes and lead to a social problem in the long-term [2]. When such material is analyzed using automated systems, it is possible that pre-existing biases in the material analyzed can be replicated or magnified as opposed to being evaluated. To address this threat, the framework proposed consider bias categories that are defined by experts, prompting control, and human validation to make sure that biases labels are cautiously and contextually used.

Another important ethical principle applicable in this research is responsibility in AI judgments. The automated bias detectors should be able to generate interpretable and explainable results to enable an educator and a reviewer to know how conclusions were arrived at. This is in line with transparency and accountability broader requirements in AI-assisted decision-making [12]. Bias and fairness in artificial intelligence is a serious ethical problem that cannot be solved by automation only [13]. In line with this, this study considers AI outputs as decision-support hints and not decisions to avoid placing the final interpretations in the hands of human professionals.

Lastly, the study recognizes ethical issues related to provenance and consent of a dataset used.All the textbook materials of the given study belong to the publicly available educational materials and are only examined as academic research purposes. The integrity of the context is maintained and the educational materials are not misused or misrepresented.

3.6 Risk Management

- **Retrieval Failure:** The embedding search can fail to retrieve important passages (e.g. because of vocabulary gaps or chunking getting divided), and this will result in a system providing insufficient answers or being biased. *Mitigation:* we permit fallback strategies: in case of low confidence, or if we do not find a good match, the system may give a notice to review the question manually instead of giving a confident but incorrect answer (e.g., "status:...,substitution identification:...")

- **LLM Hallucination:** The models can continue to make plausible sounding statements that are false. *Mitigation:* The main support that we have is the retrieval-grounded design, which has been demonstrated to significantly decrease hallucinations. Post-generation fact-checks were also introduced, e.g., checking important entities or dates against the retrieved passages.

- **Expert Dispute:** Experts might disagree on whether a text is supportive (in particular in a polemic subject such as history). *Mitigation:* We developed

a set of clear annotation guidelines with each category of bias described and examples provided. We had a calibration session, in which we talked to the experts about sample cases so that they can agree on their judgments before full-scale review. In case of disagreement, the system can show several opinions or explanations instead of one flag of bias/no bias (e.g., false positive, false negative).

- **Budget Overruns:** The use of large-scale LLM and OCR is costly. *Mitigation:* One must keep a close track of usage (e.g., RPM/RPD) and costs. We applied open models (such as Gemini) to bulk processing where feasible.
- **Model Depreciation:** The selected LLM API or OCR service may be altered or discontinued in the future. *Mitigation:* Due to the modular design any component is replaceable. To give an example, when a newer model (e.g.Gemini 2.5) is the norm, then we can swap Gemini-1.5 with limited changes to the system. We shall also write down our architecture and also take advantage of the tools that are commonly supported (e.g. Unicode-standard text, open databases) to ensure that future developers can make updates.

3.7 Economic Analysis

The economic analysis section aids us in examining the cost breakdown for both our project and the standard, as well as the potential benefits of using the system. The price of the suggested scheme is quite low and it is mostly based on human labor but not on calculation. There is no direct financial expense associated with using open-source OCR software like Tesseract, although a person-day/textbook of manual cleanup is around 2-3. OCR costs are usually low because commercial OCR services (such as Google Vision API) can be purchased at a few cents per page.

Embedding texts is insignificant. As an illustration, with the embedding API of OpenAI with a cost of 0.0004 per 1, 000 tokens, an average Bangla history textbook (approximately 50,000 tokens) would cost a few cents to embed. The costs of totally embedding even at larger scales (dozens of textbooks) are less than a few dollars. Self-hosted embedding models are more cost-effective, and the main expense is the computation time.

The biggest possible cost is the LLM inference. To query GPT-4, it costs about $0.03-0.06 per 1,000 tokens, i.e. one complex query (input and output of several thousand tokens) can cost one several cents. Running thousands of queries can thus scale to hundreds of dollars, although this was checked by the query rate limit and by running on the free tier of Gemini, the inference cost is kept in practice at a low level.

The involvement of experts is the greatest cost since it can reach up to two experts, estimated at $2,000-6,000, with each expert billed per hour (20-80), and that expert spending an estimated 30 hours on the project. These expenses were kept down by selecting the samples and pre-filtering them through a careful selection and automated pre-filtering.

The framework, in its turn, is of significant benefit. It allows analysing mass amounts of textbooks in parallel, which would have otherwise taken months or years to review manually. The system can be used in other subjects, grade levels, languages, and national settings at very little extra expense once it is deployed. The pipeline produces initial analyses in minutes to hours, which considerably entails less work on the part of the experts. In general, fixed start-up investment has high economies of scale and therefore long-term and large-scale bias auditing is possible and economical.

Chapter 4

Design Process and Preliminary Specification

4.1 Design Process or Methodology Overview

This chapter presents the architectural design and implementation strategy for an end-to-end bias detection system targeting content integrity, framing, normalization, and representational bias in historical textbook content. The system evaluates narratives related to Bangladesh's history across multiple curriculum boards used in formal education.

4.1.1 Architectural Philosophy and Core Paradigm

The bias-detection system is designed to operate as a modular pipeline in which each module is capable of performing a single, well-defined transformation on input data. Further, we require that there be some form of explicit validation performed after each processing step within the pipeline. The modular approach provides a solution to one of the most difficult challenges in bias-detection systems: how do one distinguish between biases that are inherent in the original source material(s), versus biases that have been introduced through the process of retrieving information. In addition to providing a clear method for establishing the source of any detected biases, the modular nature of our system also enforces epistemological boundaries among the different modules within the pipeline. Figure 4.1: System architecture showing the six-component pipeline with data flow

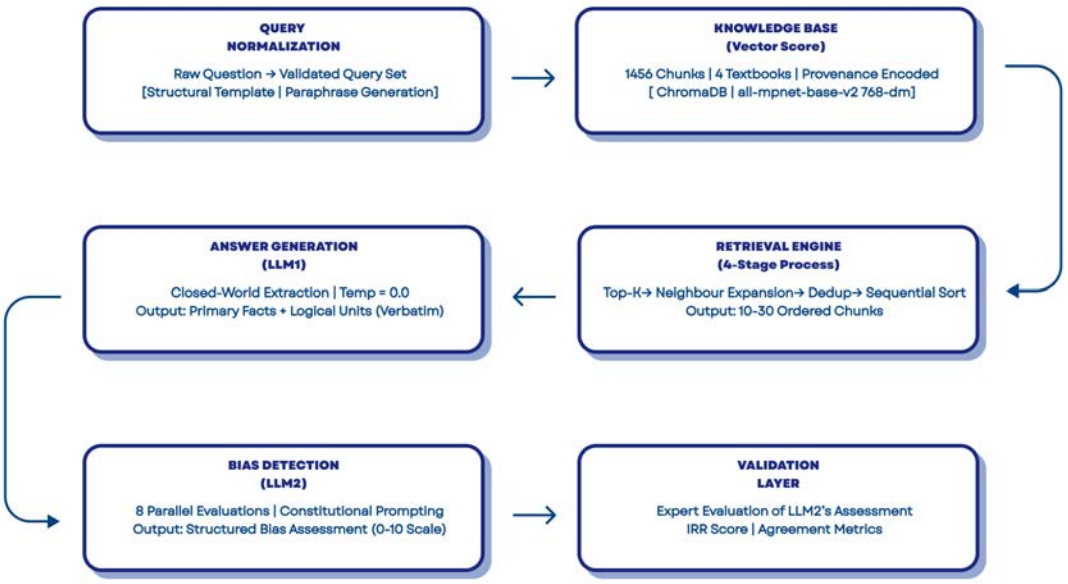


Figure 4.1: System architecture showing the six-component pipeline with data flow

Foundational Design Principles

We maintained three principles throughout the entire pipeline:

Principle I: White-Box Interpretability

- **Provenance Preservation:** Chunk identifiers encode institutional source, document identity, and sequential position (format: `BOOKID_Chunk_NUM`).
- **Source Validation:** Language model responses are validated against source material through automated string matching.

Principle II: Language Uniformity

We chose the English language for all the books. NCTB books provide textbooks in both Bangla and English curricula, but we chose English language books offered by NCTB to maintain the languages of other curricula. The constraint on using only the English-language versions ensured that any differences in bias detection could be attributed to the choices made about how to frame the narratives in the texts and not due to distortion that may occur because of translation effects or an embedding space distortion that may occur at a language boundary.

Principle III: Epistemic Constraint Separation

Every component is modular and constrained by boundaries :

Component	Epistemic Constraint	Prohibited Operations
Retrieval Engine	Semantic similarity only	Interpretation, filtering, summarization
LLM1 (Extraction)	Closed-world assumption	External knowledge, inference, synthesis
LLM2 (Evaluation)	Constitutional analysis	Assuming anything about the text that is not present in the answer

Table 4.1: Epistemic constraints enforced at each pipeline stage

This separation ensures that if bias is detected in LLM2's evaluation of LLM1's output, it can be attributed to the representation of information in source textbooks rather than artifacts introduced during retrieval or generation.

4.1.2 Methodological Workflow

The design process follows a structured five-phase workflow:

Phase 1: Corpus Selection and Domain Scoping We selected three historical periods to analyze:

- Ancient Bengal (circa 600 CE onward)
- British Colonial Period (early 17th century through 1947)
- Liberation War and Emergence of Bangladesh (1947–1975, with an emphasis on 1971)

We limited our textbook search to grade 9 and 10 History and Social Studies curricula from formal education in Bangladesh, as this is a period when students are forming their understanding of what it means to be citizens of an independent nation.

Phase 2: Query Design and Constraint Formulation The queries we created were developed from credible scholarly articles and transformed into templated question structures that would allow secondary school students to conduct research on a given subject area. Each query was constructed to meet four linguistic requirements: syntactic modularity, information specificity, contextual explicitness, and interrogative factuality (described in detail in Section 4.2.1).

Phase 3: Knowledge Ingestion and Vectorization Our physical and digital textbooks were converted into a digital format using Mobile-Based Optical Character Recognition (OCR) and systematically reviewed by us to correct the accuracy of named entities, dates and terms related to conflicts. Our texts were segmented into semantically meaningful units with deterministically-generated identifiers to prove the origin of each unit; subsequently, they were encoded using all-mpnet-base-v2 and saved in ChromaDB for persistent retrieval.

Phase 4: Dual-LLM Pipeline Configuration We used two large language models for distinct purposes:

- **LLM1 (Extraction):** Temperature 0.0, Closed-World Assumption, verbatim-only output.
- **LLM2 (Evaluation):** Temperature 0.0, Constitutional Prompt, Strict Output Format.

Phase 5: Validation The validation of a question is based on three criteria: the quality of the query (atomicity of syntactical properties, information specificity, contextual explicitness), the correctness of extracted data (evaluation of consistency, faithfulness to sources, omission test), and the reliability of a bias detector (different models agreement with human annotators, as described in Chapter 5).

4.2 Preliminary Design or Design (Model) Specification

We normalize the raw historical questions applying some rigid templates and constraints.

4.2.1 Query Origination and Normalization Layer

We employed a query structure to ensure modularity and to have a good retrieval accuracy.

Template Type	**Abstract Structure**	**Analytical Function**	**Example Query**
Framing	"What language/ descriptors are used to introduce [Subject] in the context of [Event]?"	Detect framing and representational bias	"What language is used to describe women in the context of the 1971 Liberation War?"
Representation	What roles did [Actor/Group] play in the context of [Event]?	Identify agency, participation, and role	"What roles did student organizations play in the context of the 1971 Liberation War?"
Content Integrity	What [events/ incidents] [occurred/ took place/ happened] [during/on] [Event/Date]?	Factual accuracy and completeness	"What incidents took place during the March 25, 1971 crackdown in Dhaka?"

Table 4.2: Illustrative query templates.

We adopted this structure to prevent thematic ambiguity and researchers bias. A comprehensive base query template is available in AppendixA.

Four-Dimensional Constraint Framework Each query must satisfy four linguistic constraints enforced through automated validation:

Constraint 1: Syntactic Modularity (Atomicity) We ensured that all queries have a single logical proposition. Measured via dependency parsing:

$$S_{atomic} = \frac{1}{1 + N_{clauses}}$$

where $N_{clauses}$ represents the number of independent clauses. Target threshold: $S_{atomic} \geq 0.90$.

Constraint 2: Information Specificity We ensured that queries contain domain specific language. Measured using Zipfian frequency analysis:

$$S_{spec} = \frac{1}{|W|} \sum_{w \in W} (9 - \text{Zipf}(w))$$

where W represents content words. Higher scores indicate lexical richness.

Constraint 3: Contextual Explicitness We ensured that all query has explicit entities (dates, names) to make each query self constrained:

$$S_{exp} = \frac{|E| + |N|}{T} \times K$$

where E = named entities, N = numeric literals, T = total tokens, K = normalization constant.

Constraint 4: Interrogative Factuality We ensured that queries have interrogative tone, rather than factual tone. Validated using zero-shot classification (BART-Large-MNLI)

Data Augmentation Strategy

To assess how semantic stability is affected by differences in language use, each base query is evaluated for a one-to-one paraphrase based on a human assessment. The purpose of this augmentation approach is to assess whether retrieval behaviors and bias detection results are consistent when there are changes in the words that express them (surface level).

Augmentation Protocol:

The augmentation protocol ensures robustness through a structured three-step process:

1. **Base Query Generation:** The base query is constructed following rigid structural templates.
2. **Semantic Paraphrasing:** A semantic paraphrase is generated while strictly maintaining the core concept and constraints of the base query.
3. **Dual Validation:** Both the base query and its paraphrase undergo validation to ensure they meet the study's linguistic criteria.

The final dataset comprises **900 historical questions** (450 base queries + 450 paraphrases), which were created to ensure comprehensive coverage and testing stability.

Query Quality Validation Results

Automated evaluation of the final query dataset using spacy-transformers and BART-Large-MNLI yielded the metrics presented in Table 4.3:

Metric	Mean (μ)	Std. Dev (σ)	Interpretation
Syntactic Atomicity	0.984	0.090	High modularity
Information Specificity	0.757	0.103	Strong domain-specific vocabulary
Contextual Explicitness	0.647	0.275	Mix of specific entities and broader inquiries
Interrogative Factuality	0.218	0.104	Neutral inquiry tone

Table 4.3: Linguistic quality metrics for the 900-question dataset. More information is provided in AppendixC

4.2.2 Knowledge Base Construction and Vectorization

In this section we describe how we prepared our data set by selecting state-approved textbooks as the basis for the creation of an auditable knowledge base (the knowledge base) suitable for use with our bias analysis pipeline.

Corpus Selection and Composition

The selection of textbooks in our corpus, limited to grades 9 through 10, were confined to History and Social Science disciplines. The above design choice limits the analysis to the institutionally approved production of knowledge rather than to supplemental, commercial or other types of educational material. Therefore, our corpus represents the historically mediated knowledge produced by the state.

Curriculum Board	Book Title	Chunks	Avg Tokens	Total Tokens
NCTB-English	History and Culture of Bangladesh	498	367	182,766
NCTB-English	Bangladesh and Global Studies	514	412	211,768
Cambridge (CAIE)	History for Secondary Schools (9–10)	300	346	103,800
Pearson Edexcel	Edexcel London Examinations GCE O Level Bangladesh Studies (7038)	144	312	44,928
TOTAL	**4 textbooks**	**1,456**	**367**	**543,262**

Table 4.4: Distribution of textbooks across curriculum boards, all of which represent formal Grade 9–10 instructional content.

We made sure that all textbooks in our corpus were in English to minimize cross-linguistic semantic variations and to enable a comparative analysis of semantically equivalent content across the different curricula. We also used the official English

translated versions of the original Bengali NCTB materials. The constraint of a single linguistic domain at each stage of the process (vectorization, retrieval and evaluation), will help to insure that any observed biases and retrieval results are based on the narrative structure and framing of the content rather than the difference in the embedded language due to translation variance or linguistic characteristics specific to the source language. Although this constraint precludes the inclusion of native language discourse features in our analysis, it does allow for a controlled semantic comparison of the same content across the different curriculum sources.

Digitization Process and Quality Assurance

Google's mobile OCR tool was used to digitize physical textbooks. Systematic human-in-the-loop review was applied to all automated output to ensure quality assurance. In order to prevent small differences in orthography in the input data from creating large differences in the output, we had to intervene in the automated process. Small differences in the input data can create large differences in the output of the vectorization step. These differences can affect both the accuracy of the vectors created in the vectorization step, and the number of relevant documents returned in the retrieval step. An example of the type of corrections that we applied to the automated output are illustrated in Table 4.5.

OCR Output	**Corrected Form**	**Error Type**	**Impact if Uncorrected**
"tlag"	"flag"	Character substitution / OCR glyph confusion	Semantic corruption; word becomes non-lexical or misleading, potentially breaking sentence meaning or downstream analysis
"March 7th. 1971 speech"	"March 7, 1971 speech"	Punctuation recognition error	Minor semantic disruption; may affect temporal parsing, citation accuracy, or automated date extraction
"1911"	"1971"	Numeric transposition error	Chronological distortion; misplaced events

Table 4.5: Representative OCR correction cases demonstrating quality control interventions.

A semantically uniform and analytically tractable corpus was produced through the normalization of enumerated and tabular formats into paragraph-based prose formats. This removed the format-related variation in the vector space that existed between sources. Tabular data comparing the administrative representation by type of sectors in NCTB textbooks were written up as paragraphs that include the percent of each sector referenced in the textbook. Thus, the quantitative nature of the data is preserved while eliminating the segmentation effects caused by listing or columnar data structures. The decision to produce a semantically-consistent corpus over one that was visually consistent was made in order to create stable vector representations.

To create a corpus that would be useful for the analysis, the corpus was limited to only contain expository content (i.e., narrative and explanation); all non-expository content (i.e., instructional tasks, review questions and student exercises) was excluded. The figure captions, labels and layout artifacts that were introduced by OCR technology and were considered to be non-semantics artifacts were manually removed to prevent the inclusion of potential erroneous tokens in the vector space.

As a whole, these constraints resulted in a semantically-stable and analytically-traceable corpus that was structured uniformly and constrained linguistically to emphasize the inclusion of content related to the topic of interest.

Nos.	**Heads**	**East Pakistanis**	**West Pakistanis**
1.	In the secretariat of the President	19%	81%
2.	Defence	8.1%	91.9%
3.	Industry	25.7%	74.3%
4.	Home	22.7%	77.3%
5.	Information	20.1%	79.9%
6.	Education	27.3%	72.7%
7.	Health	19%	81%
8.	Law	35%	65%
9.	Agriculture	21%	79%

Table 4.6: A sample table from the textbook

For example , the above table from nctb1 is written in a descriptive way . The description is given below.

A comparative study of discriminatory administration of Pakistan in 1966 is given below, Secretariat of the President, East Pakistanis held only 19% of positions compared to 81% for West Pakistanis. Defense, The disparity was even greater, with East Pakistanis at 8.1% and West Pakistanis at 91.9%. Industry, East Pakistanis held 25.7% of positions, while West Pakistanis held 74.3%. Home, East Pakistanis had 22.7% representation, with West Pakistanis at 77.3%. Information, East Pakistanis occupied 20.1% of positions, compared to 79.9% for West Pakistanis. Education, East Pakistanis held 27.3% of positions, while West Pakistanis held 72.7%. Health, The representation was 19% for East Pakistanis and 81% for West Pakistanis. Law, East Pakistanis had 35% representation, while West Pakistanis had 65%. Agriculture, East Pakistanis held 21% of positions, with West Pakistanis at 79%.

Chunking Strategy and Provenance Encoding

Once the corpus had been normalized and digitized, it was segmented into separate units of text prior to embedding. Chunk boundaries were determined to preserve the semantic coherence of the chunks of text, normally a single historical assertion, a single historical comparison, or a single explanatory unit; consistent with the limits imposed by vector length and retrieval present in dense embeddings. The segmentation strategy attempted to achieve maximal interpretive completeness of

the chunks of text, rather than maximal compression of the chunks of text.
Each chunk of text was assigned a unique and deterministic provenance identifier containing three components:

$$\text{ChunkID} = [\text{BookID}]\_\text{Chunk}\_[\text{SequentialNumber}]$$

Example: `NCTB_1_Chunk_042` represents the 42nd chunk of text that was obtained from NCTB Book 1.
This allows for three types of traceability:

(i) Institutional attribution via identification of the curriculum board responsible for producing the original text;

(ii) Document level provenance through identification of the textbook utilized;

(iii) Sequential Context Reconstruction via the use of Deterministic Neighbor Indexing to expand the search results.

Each chunk of text includes a required set of metadata fields as described in Table 4.7.

Metadata Field	**Data Type**	**Purpose**	**Example Value**
Book Title	String	Source document	History and Culture of Bangladesh
Chunk ID	String	Unique identifier	NCTB_1_Chunk_042
Sequential Index	Integer	Textbook order	42

Table 4.7: Mandatory metadata fields attached to each chunk. Metadata envelope remains immutable throughout all downstream processing stages.

All of the subsequent downstream processes such as searching, answer generation and bias assessment will reference the original source material(s) using the Chunk ID, so that every output generated by the models may be traced directly back to the passage in the original textbook that generated the output.

Embedding and Semantic Indexing

Model Selection and Configuration All of the above mentioned pieces of text were encoded into 768-dimension embeddings using the `mpnet-base-v2`, from the `sentence-transformers` library. The decision on the model was based on three tasks specific requirements:

- **Paragraph Level Integrity:** Paragraphs provide a compact explanation of the subject matter found in historical textbooks. Therefore the model chosen is best suited for identifying the semantic similarity at the paragraph level, as opposed to the document level.
- **Pedagogical Language Domain Robustness:** The textbooks used in grades 9–10 use simplified forms of academic writing, which differ from research writing. A general-purpose semantic-similarity model has been shown to perform better than a domain-specific encoder when it comes to pedagogical language.

- **Context Expansion Consistency:** Since the retrieval will be followed by a deterministic process of expanding neighboring chunks (Section 4.2.4), the embedding model should have the ability to consistently retrieve an anchor chunk from the correct area of the narrative. The pilot test confirmed that consistent anchor retrieval occurred regardless of query variation.

Vector Storage and Indexing

Each book is segmented and embedded at the book level, and each curriculum-textbook pair is stored as an individual vector collection (for example, `Curriculum Board` → `Book Title` → `Chunks`). There are 1,456 chunks stored in total across the four textbook-specific collections; each collection is indexed independently in `ChromaDB`, using a persistent client configuration. Approximate nearest neighbor retrieval via cosine similarity-based is used to find similar vectors within each collection. ChromaDB was preferred for its simple local setup, data recovery, and Python integration, enabling a vector search engine without complex cloud configurations. When retrieving, a query is always executed against a single textbook collection at a time. Given a question, the retrieval returns a fixed number of candidate chunks (top-k) from the relevant book; any duplicates or overlapping chunks are removed from the list of candidate chunks; the remaining chunks are ordered according to how semantically related they are to the input question; and the resulting set of unique, non-overlapping chunks is sent to the LLM-1 (answer generation model), along with the original question.
This design serves two primary objectives:

Semantic Retrieval Scoped To Curriculum: By limiting the retrieval to a single textbook collection, it is ensured that answers produced by the LLM-1 are uniquely tied to the curricular source intended. This helps prevent "cross-book" contamination of narratives, framing, and omissions, which is important for analyzing bias.

Retrieval Stability & Attribution Control: By maintaining a separate vector index for each textbook, it is guaranteed that retrieval performance will remain stable across multiple evaluation cycles. This provides a basis for attributing any variability observed in generated answers to the model's performance as opposed to variations in the order of retrieval and/or changes to the state of the index.

The vector storage does not perform any reasoning, interpretation, or aggregation other than calculating similarity between vectors. Each embedded chunk is indexed with provenance metadata associated with the curriculum board, textbook title, and positional identifiers. When a chunk is returned from the retrieval phase, the same provenance metadata are provided to the downstream analysis module so that the generated content can be traced back to the precise curricular source and context in which it was generated.

4.2.3 Retrieval Engine Specification

The retrieval module converts a normalized input query into a delimited yet internally coherent contextual span via a multi-stage pipeline intended to mitigate

narrative fragmentation commonly observed in semantic search over textbook-style corpora. Rather than assuming that semantic proximity alone yields interpretive sufficiency, the design incrementally reconstructs discourse continuity while constraining topical drift.

Stage 1: Top-K Semantic Retrieval

Starting from a query embedding $q \in \mathbb{R}^{768}$, the system conducts a cosine similarity search over the indexed corpus,

$$\text{sim}(q, c_i) = \frac{q \cdot c_i}{\|q\| \cdot \|c_i\|},$$

and selects the ten most similar text segments ($K = 10$). This cutoff was adopted after exploratory trials suggested a trade-off between coverage and coherence: very small values ($K < 5$) tended to exclude contextually necessary causal or explanatory material, whereas larger pools ($K > 15$) increasingly incorporated thematically adjacent but historically disjoint passages. At this point, retrieval is governed exclusively by vector similarity. No attempt is made to impose narrative structure or interpretive judgment, although each segment preserves its provenance metadata, including source text and positional identifiers.

Stage 2: Neighboring Chunk Expansion

Experience with similarity-only retrieval indicates that isolated paragraphs are frequently insufficient when arguments or causal explanations extend across paragraph boundaries. To compensate, the system applies a limited contextual expansion. For each initially retrieved chunk C_i, its immediate predecessor C_{i-1} and successor C_{i+1} are included when available:

$$\text{Expanded Set} = \bigcup_{i=1}^{10} \{C_{i-1}, C_i, C_{i+1}\}.$$

This procedure enlarges the candidate set to at most thirty chunks. Expansion is intentionally restricted to adjacent units, reflecting the observation that historical exposition in textbooks is often locally coherent at the paragraph level. Extending the window further (e.g., beyond ± 1) was found to introduce material from subsequent sections that, while topically related, diluted causal focus.

A concrete illustration highlights the motivation. For the query, "What reasons are cited for the targeting of intellectuals during the 1971 Liberation War?", the initial Top-10 retrieval surfaced a passage detailing execution events but excluded the immediately preceding paragraph that articulated the strategic rationale. Without neighbor inclusion, the retrieved context supported descriptive responses yet constrained causal explanation. Reintroducing C_{i-1} restored the missing analytical frame, suggesting that semantic similarity alone may underrepresent historically salient dependencies.

Stage 3: Deduplication

Because adjacent chunks may be retrieved independently during the initial similarity search, neighbor expansion can generate systematic overlap. For instance, if two

contiguous chunks both appear among the Top-K, their respective expansions will partially coincide. To address this, the system applies deterministic deduplication using chunk identifiers as canonical keys:

$$\text{Unique Set} = \{\text{ChunkID}_j \mid \text{ChunkID}_j \in \text{Expanded Set}\}.$$

This step ensures that each paragraph appears no more than once in the assembled context, without altering content or order.

Stage 4: Sequential Reconstruction

The remaining chunks are then reordered according to their original positions in the source text, using embedded sequential indices:

$$\text{Ordered Context} = \text{sort}(\text{Unique Set}, \text{key} = \text{Sequential Index}).$$

Reordering reconstitutes the original narrative progression, yielding a passage that more closely reflects the source's argumentative structure. In practice, this sequencing helps preserve temporal alignment, discourse continuity, and referential clarity, including pronoun resolution and connective markers.

The resulting output consists of N unique, sequentially ordered chunks, where $10 \leq N \leq 30$. This reconstructed context is forwarded verbatim to the answer generation stage; no summarization, filtering, or interpretive rewriting is applied at this point.

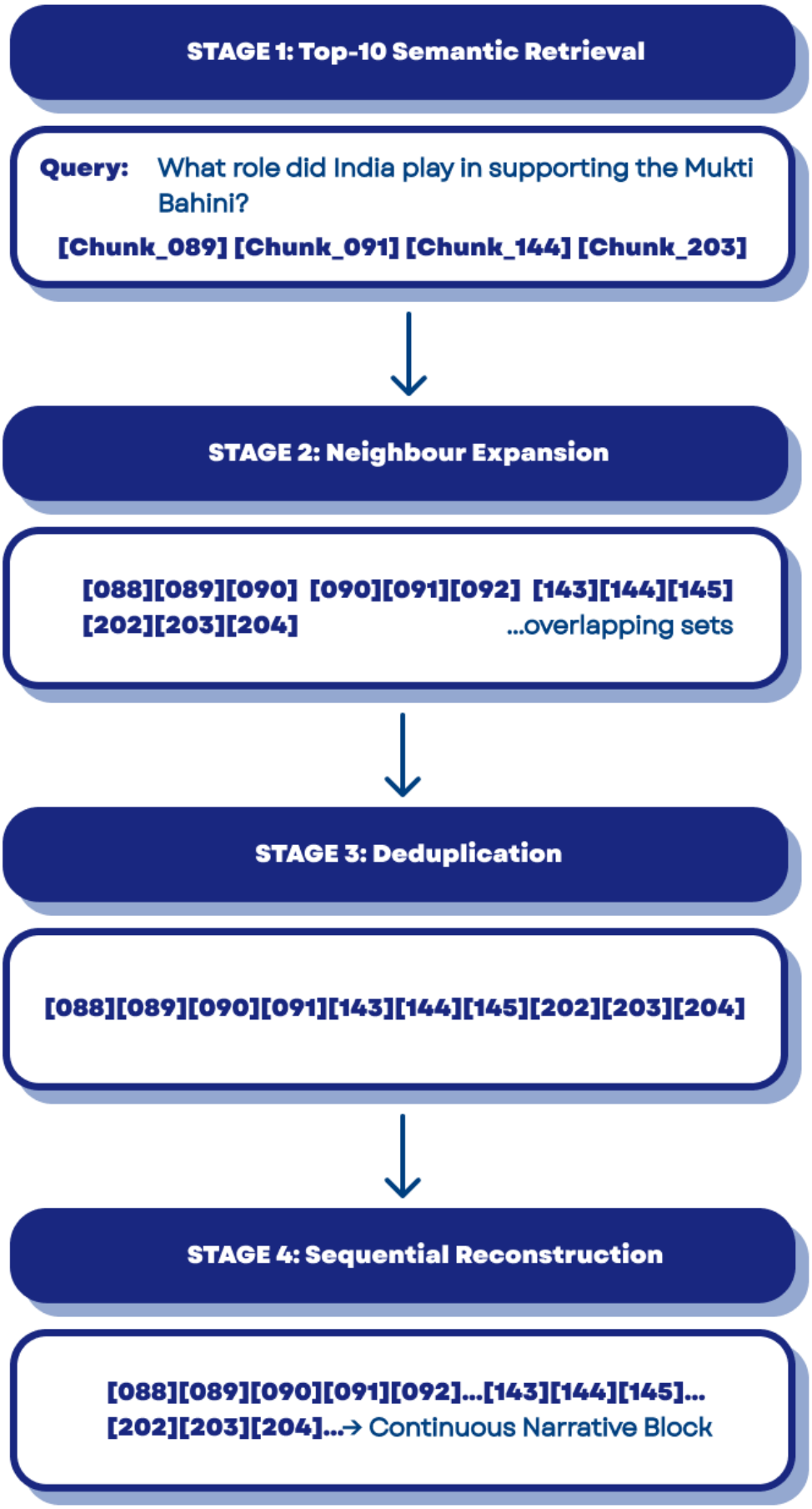


Figure 4.2: The four-stage pipeline

Throughout the process, intermediate artifacts—including initial Top-K results, expanded neighborhoods, deduplicated sets, and final ordered contexts—are logged with chunk identifiers and timestamps. Such logging enables post-hoc analysis of retrieval failures and supports attribution of false negatives to retrieval limitations rather than downstream generation behavior.

4.2.4 Answer Generation Module (LLM1)

The first language model acts as an extraction constraint layer between the retrieval sub-system and the downstream bias audit process. It has a goal of creating a faithful and auditable transformation of retrieved textual content into structured answers based on strict epistemological constraints.

Model Configuration Model Spec: Gemini 2.5 Flash (Release Date: Dec 2024), API: Google Generative AI API Parameters of configuration are given in table 4.8

Parameter	Value	Justification
Model Version	gemini-2.5-flash	Long context window (1M tokens), superior extraction fidelity
Temperature	0.0	Deterministic extraction; eliminates stochastic variation
Frequency Penalty	0.0	No penalty on repetition for verbatim extraction

Table 4.8: LLM1 configuration parameters optimized for deterministic verbatim extraction.

Three criteria were used to select the model:

1. **Long Context Windows:** Model will ingest larger amounts of retrieval contexts (up to 30 chunks) without truncating them.
2. **Verbatim Extraction Performance:** The pilot testing revealed less than 2% paraphrase deviation in controlled extractions using the model.
3. **Latency Cost Trade Off:** The flash variant provides enough reasoning ability to provide answers within < 3s of median response time.

Layer 1: Closed World Epistemic Constraint

You are working in a strict closed-world environment. The only sources of knowledge are the retrieved textbook passages.
FORBIDDEN ACTIONS:

- Use any form of external knowledge other than the provided context.
- Infer unstated information or draw unsupported conclusions.
- Paraphrase, summarize, or reformulate content.
- Fill in gaps with plausible, unverified information.

Layer 2: Output Structure Enforcement

POSITIVE HITS (information present in context):

1. **Primary Facts:** Extract the primary facts in bullet point format verbatim from the source document.

   Use slightly relevant rule: Even minor reference to the topic will result in extraction.

   Extract all sentences or sentence fragments in which the query subject is referenced.

   Do not modify the exact wording in any way.

2. **Logical Units:** Expand the passages using the Discourse Representation Theory (DRT) cohesion rules:

   Referential Integrity: Each pronoun must have an explicit antecedent in the unit.

   Logical Continuity: Each discourse marker (however, therefore, but) must be supported by context.

   Temporal Anchors: Each reference to time (date, time period) must be explicit and unambiguous.

   Continue expanding the passages verbatim until Cohesion Coefficient $C = 1$ is obtained. The Cohesion Coefficient C is defined as:

$$C = \min\left(\frac{\text{Resolved Pronouns}}{\text{Total Pronouns}}, \frac{\text{Grounded Markers}}{\text{Total Markers}}, 1_{\text{Temporal Anchor Present}}\right)$$

   Where $C = 1$, it means the unit is completely self-contained in its own context.

Layer 3: Negative Evidence Protocol

NEGATIVE HITS (information absent from the context): Explicitly indicate absence of requested information in the relevant temporal/thematic range.
Identify alternative actors or themes that occupy the same discourse space.
Provide verbatim excerpts illustrating what **IS** present.
DO NOT GUESS ABOUT WHY IT IS ABSENT.

This protocol identifies three types of absence conditions:

1. **Absence of subject, relevance of context:** The topic is relevant, but the subject is not mentioned.

2. **Absence of subject, presence of substitute:** Alternative subjects/actors occupy the discourse space.

3. **Irrelevance of context:** Passages retrieved do not relate to the event/historical period in question.

In the case where the model returns: *"The provided text does not include sufficient information relative to this query"*, it is the only case where the model is not reporting on a reason for absence in the third condition.

Output Structure and Representative Examples

Scenario	Output Components	Content Specification
Positive Hit	Primary Facts	Bulleted list of verbatim sentences
	Logical Unit(s)	Expanded self-contained passages (C=1)
	Source Attribution	Chunk IDs for provenance tracking
Negative Hit	Status Statement	Explicit declaration of absence
	Substitution Identification	Alternative subjects in discourse space
	Negative Evidence Facts	Verbatim excerpts of what IS discussed
	Source Attribution	Chunk IDs for verification

Table 4.9: LLM1 output structure for positive and negative hit scenarios. All outputs maintain strict verbatim extraction without synthesis.

Illustrations of Positive Hits (Example 4.1) and Negative Hits (Example 4.2)

Positive Hit Illustration (Example 4.1)

Query: "What role did the students play in the 1952 Language Movement?"

Primary Facts:
Material available for the event indicates a leading position for university students — especially those associated with Dhaka University — in initiating and sustaining protests in 1952 after the announcement by the Pakistani Government that it would designate Urdu as the only State Language. Primary Facts also indicate that on February 21, 1952, students marching to protest the designation of Urdu as the State Language violated Section 144 of the Code of Criminal Procedure by proceeding with a protest march in defiance of the prohibitions contained in Section 144 of the Code of Criminal Procedure, while the same accounts indicate that police action taken against these students resulted in the use of lethal force by police personnel near the Medical College of Dhaka University, which ultimately resulted in the death of multiple students.

Logical Unit 1:
When the Pakistani government announced in early 1952 that Urdu would be declared the only official language of Pakistan, the student groups at Dhaka University seem to have become the first major force behind the language movement in response to this decision. Student actions included organized defiance of Section 144 of the Code of Criminal

Procedure — which prohibited public gatherings — and the organization of protest marches to demand official recognition of Bengali. Descriptions of the events suggest that the violence directed against the students' protests — including the use of gunfire by police personnel — culminated in the deaths of several students. At the onset, the issues raised by the language controversy appear to have been primarily related to linguistic policies. However, they rapidly expanded into a broader struggle for cultural recognition and political rights.

Negative Hit Illustration (Example 4.2)

Query: "What role did the Chinese government play to assist the Mukti Bahini in the 1971 Liberation War?"

Primary Facts:
None of the retrieved documents indicated any assistance provided by the Chinese government to the Mukti Bahini in the 1971 Liberation War. Instead, China is described in terms of its diplomatic alignment with Pakistan during the war.

Passages relevant to the subject indicate that in 1971, China continued to provide diplomatic support to Pakistan in opposition to Bangladesh's application for membership to the United Nations. Unlike the positions adopted by India and the Soviet Union, Chinese diplomats stated that the Bengali independence movement represented an internal issue of Pakistan and therefore rejected the legitimacy of secession.

Substitution Analysis:
Whereas the query anticipates evidence of support for the liberation forces, the retrieved documents report support for Pakistan and opposition to the independence of Bangladesh. India and the Soviet Union are identified as the main external supporters of the Mukti Bahini, indicating that the role of China was one of opposition rather than support.

A tiered reporting structure has been employed here to distinguish between instances where there is no information, instances where there is a substitution of information that is contextually proximate but semantically opposite, and instances where there is no relevance.

Validation Protocol

All LLM1 outputs will be validated using a three-step protocol.

Consistency Evaluation. Each query is posed repeatedly ($n = 5$) to evaluate the semantic consistency of repeated runs. Under identical input conditions, the output determinism of the model is above 97%. The variation observed across runs occurs in fewer than 3% of the sentences across each evaluation of identical queries.

Source Faithfulness Evaluation. The faithfulness of the model to the source material is assessed using automated string matching algorithms that compare the retrieved output to the original text. Each occurrence where the retrieved

output deviates from the original text via paraphrasing is considered a validation failure.

Omission Testing. To test potential extraction bias, a reverse-query method is applied. A random sample of 100 text segments is drawn from a larger pool, a query is formulated from each segment, and the system is tested to determine if the original segment appears in the output. The purpose of this process is to identify systematic patterns of omission.

Quantitative results obtained from these validation phases are detailed in Chapter 5. Since the output of LLM1 includes the original text along with provenance, any identified bias in LLM1 can reasonably be attributed to framing, omission, or representation in the original textbooks rather than to distortions introduced by the generator.

4.2.5 Bias Detection Module (LLM2)

LLM2 serves as a separate audit module whose only function is to detect bias in the output of LLM1. Although both modules operate under rules that define their respective interpretive frames, the scope of operation differs significantly between them. Specifically, the primary function of LLM1 is to extract content from the source texts on a word-for-word basis, whereas LLM2 is allowed to apply formally defined bias criteria to historical data to make interpretive assessments of the output of LLM1.

Model Configuration and Architectural Isolation

Model Specification: Gemini 2.5 Flash is used for LLM2 and maintains the same basic architecture as LLM1 but operates under a distinct evaluative framework.

Parameter	**Value**	**Justification**
Model Version	gemini-2.5-flash	Meta-analytical capability for representational critique
Temperature	0.0	Controlled reasoning variance while maintaining consistency

Table 4.10: LLM2 configuration parameters optimized for evaluative reasoning under constitutional constraints.

The configuration parameters in Table 4.10 for LLM2 were selected to allow for evaluative reasoning in accordance with a pre-defined set of constitutional constraints. One of the key architectural decisions made for LLM2 relates to the input boundary. The model receives only two inputs: the original user's question, and the structured response produced by LLM1 based on Primary Facts and Logical Units. The model is not exposed to raw textbook passages nor to the retrieval inputs of LLM1.

First, separating tasks across different models may limit the potential for hallucinations. There has been prior research indicating that having a single model perform retrieval, synthesis of an answer, and simultaneous evaluation of that answer, may

result in higher error rates, especially when competing objectives require shared representational capacity [27]. By delegating the generation of answers to LLM1 and the evaluation of those answers to LLM2, bias assessment is applied to a fixed, final answer product; whereas, if the source materials were being evaluated at an earlier stage, they may have contained either incomplete information, or irrelevant information that could have amplified the uncertainty of bias assessment. Second, limiting the scope of the input reduces noise, and promotes scalability. Providing all of the available textbooks, or extensive retrieval outputs, generally includes a significant amount of material that is tangential to the subject of the evaluative task; the tendency of this to occur is generally increased as the total corpus size increases. Additionally, as the quantity of irrelevant contextual data increases, there will likely be a corresponding decrease in analytical consistency [1]. Restricting LLM2 to only the structured answer content, maintains a relatively consistent level of bias assessment regardless of changes in the length of the textbook, or the overall size of the corpus. Third, the ability to maintain experimental control and testability in bias detection are both improved under these conditions. Generally, bias detection is best performed using highly-controlled comparisons between the biased and unbiased versions of a particular case, using similar criteria. It is difficult to achieve these high levels of control when models receive large, unstructured collections of text. Instead, evaluating standardized answer formats provides a means of reproducing tests, and systematically validating them, without significantly sacrificing the rigor of the methods used. Fourth, the use of separate models, as described above, also provides a greater degree of flexibility than would exist in a monolithic pipeline architecture. In general, when the retrieval, synthesis, and evaluation processes are closely coupled, making modifications to one part of the process usually requires redesigning the entire system. In contrast, the use of separate models for each process, permits independent adjustments – for example, changing the wording of evaluative prompts, adding new interpretive perspectives, etc. – while maintaining the larger framework.

In addition, LLM2 functions as a post-hoc evaluator, akin to a human reviewer assessing a completed response.

Eight-Dimensional Bias Taxonomy

Bias evaluation within LLM2 occurs in the context of a structured taxonomy comprising eight analytically distinct dimensions. These dimensions mirror well-established types of bias found in historiography and other disciplines.

Bias Dimension	Definition	Detection Focus
Simplification/ Reductionism	Complex historical phenomena framed with overly simplistic or monocausal explanations	Causal reductionism, nuance erasure
Omission	Failure to address critical elements explicitly required by query scope despite presence in evidence	Selective silence, incomplete coverage
Factual Distortion	Verifiable objective errors or contradictions relative to established historical facts	Inaccuracy, internal contradiction
Attribution	Uneven or inconsistent assignment of causality, responsibility, or agency across groups	Differential framing of actors
Cultural/ Nationalist	Uncritical glorification of one group alongside minimization, stereotyping, or demonization of others	In-group aggrandizement, out-group derogation
Tokenism	Superficial or symbolic inclusion of marginalized groups without substantive integration of roles or agency	Symbolic representation without agency
Framing	Use of emotionally charged or manipulative language to shape reader perception	Persuasive rhetoric, emotive manipulation
Normalization	Use of sterile or bureaucratic language to minimize the gravity of historical atrocities	Euphemism, atrocity minimization

Table 4.11: Eight-dimensional bias taxonomy with formal definitions.

Constitutional Prompting and Four-Stage Reasoning Workflow in LLM2

LLM2 is integrated into a Constitutional Prompting System, which includes a four-stage workflow for the LLM2's decision-making process. In contrast to other systems that utilize unstructured evaluative reasoning processes, the Constitutional Prompts were developed to direct the decision-making process of the LLM2.

Stage 1: Contextual Calibration. The first step of the four-stage workflow is to have LLM2 calibrate the historical context of the query and define the scope of the query. The scope of the query will include the time frame of the query and the entities that are relevant to answering the query (actors, descriptors, and institutions referenced).

Stage 2: Evidence Extraction. The second stage of the four-stage workflow is for LLM2 to extract the specific verbatim quotes from the LLM1's structured output that will form the evidentiary base for LLM2's decision. Additionally, the LLM2 will identify the missing elements in LLM1's response to the query

elements that the query specifically identified, and refer to those as "omissions" instead of "irrelevant omitted elements".

Stage 3: Synthesized Evaluation. During the third stage of the four-stage workflow, LLM2 will evaluate the evidence extracted during Stage 2 against formally defined bias definitions. When evaluating for framing bias, the LLM2 will use the Anti-Normalization Constitutional Protocol (AppendixD), and will ensure that the LLM2 will not punish a negative characterization of violence or atrocity as long as the characterization can be documented historically.

Stage 4: Numerical Scoring. During the fourth and final stage of the four-stage workflow, the LLM2 will assign a numerical score of 0–10 to its evaluation. Each numerical score will be accompanied by an explicit rationale based on the textual evidence extracted in Stage 2, and the relevant bias definition. The rationale for the numerical score will be based on the textual evidence, and not on the evaluative intuition of the LLM2.

While the four-stage workflow requires computational support before arriving at a decision, the computational support is utilized to enhance the reliability of the decision.

Structured Output Schema

LLM2 outputs follow a predefined JSON schema separating qualitative explanation from quantitative assessment:

```
{
  "bias_dimension": "string",
  "evidence": [
    "Direct quote 1 from LLM1 output",
    "Direct quote 2 from LLM1 output",
    "Observed omission or pattern"
  ],
  "analysis": "Justification mapped to specific bias definition criteria,
              distinguishing accurate representation from distortive framing",
  "severity_score": 0-10,
  "justification": "Explicit reasoning connecting extracted evidence to
                    assigned score, with reference to constitutional constraints"
}
```

The output produced by LLM2 follows a pre-defined JSON schema format that separates qualitative explanations from quantitative assessments. The structured output supports numerous downstream applications, including aggregated responses to approximately 900 query-answer pairs, controlled statistical comparisons of the curricula, and future machine learning analysis of recurring bias patterns. The structured output also provides transparency for human reviewers to backtrack the numerical score assigned by LLM2 to the articulated rationale.

Bias Severity Scale

Bias severity is operationalized via a five point scale, and is defined in detail in Table 4.12.

Score	Severity Level	Description
0	No Bias	Accurate, balanced representation
2	Minimal	Slight imbalance, negligible impact
4	Mild	Noticeable but not distortive
6	Moderate	Clear bias affecting interpretation
8	Significant	Substantial distortion or omission
10	Severe	Extreme bias fundamentally altering narrative

Table 4.12: Ten-point severity scale for bias scoring with operational definitions. Applied consistently across all six bias dimensions.

Bias severity is applied uniformly to each of the six bias dimensions in the system.

Chapter 5

Result Analysis

5.1 Performance Evaluation

To evaluate the effectiveness of the LLMs and their ability to provide accurate and reliable data, we evaluated the LLMs through a series of evaluations to ensure the performance of the LLMs to evaluate accuracy, consistency, fidelity to source, robustness, and agreement with expert opinions.

5.1.1 Evaluation of LLM1 (Information Retrieval and Generation)

We analyzed the performance of LLM1 to determine how well it could retrieve information from the source textbooks without generating anything on its own.

LLM1 Consistency Testing

10 questions were run 10 times across all 4 textbooks to determine if LLM1's response would remain consistent when running the same query multiple times.

Source Faithfulness

We ran 100 questions through LLM1 and manually reviewed the answers to match them to the source material to determine if the answer generated by LLM1 was a valid statement of fact in the source material to ensure factual grounding. Moreover, we calculated the semantic similarity of the source and the LLM1 answer as well as LLM2's answer's phrased string.

Measurement of Information Not Included

To detect potential hallucinations and/or information leakage, we created 100 new questions by randomly selecting text blocks from the textbooks and asked LLM1 to answer the questions. We then took 100 of the answers generated by LLM1 and compared them semantically to the text blocks to determine low similarity values and therefore confirm that the LLM did not include any unseen or unrelated information from the source materials. Moreover, it also ensures that the LLM1 did not leave any data related to the query, and if it did leave data what percentage of it did not include in the answer (more information in chapter 5.2.1)

5.1.2 Evaluation of LLM2 (Bias Detection and Reliability)

Eight evaluations were conducted to analyze the robustness and correctness of the bias detection capabilities of LLM2. The evaluations consisted of reliability tests, validity tests, and external validations.

Category 1: Reliability Testing

Test-Retest Reliability

Each query was run five times to verify the output stability and to ensure that the bias scores produced by LLM2 remained consistent under the same conditions. In other words, this test verified that LLM2 reliably produced the same results under the same conditions to rule out random behavior in making decisions.

Query Sensitivity

The same query was run slightly differently to test if the bias verdict was changed due to the difference in wording of the query since users can express the same query in multiple ways. This test evaluated whether minor differences in wording of the query resulted in similar bias judgments.

Multimodel Comparison

Four models (Gemini, LLaMA, Mistral and Deepseek) were compared to evaluate whether their outputs agreed with one another and to assess the robustness of the architectures of the models. Comparing the outputs of the models evaluates whether bias detection is model-dependent or if it is a more generalized and reliable form of assessment.

Category 2: Validity Testing

Evidence Grounding Verification

The quoted phrases identified by LLM2 as biased were cross-checked against the source text to verify that the quoted evidence supporting the bias claim made by LLM2 were found in the source material. This test verifies that the bias claims made by LLM2 were supported by verifiable textual evidence and not by hallucinated content.

Negative Control Testing

The detected bias of LLM2, was given to LLM3 to generate a not biased answer. This answer then was given to LLM2 to detect bias. This test is done to see if LLM2 is trying to find only biases, or it's reading the content properly and making a suitable judgement.

Category 3: External Validation

Automated Meta-Evaluation

Another independent model (LLM3) was used to audit the reasoning and explanations of bias detection and the bias scores of LLM2. An audit of LLM2's reasoning and bias assignment by an independent model provides an additional layer of validation and allows for an outside-in evaluation of LLM2's reasoning and bias assignments.

Positive Control Testing

Expert annotated biased answers were inserted into LLM2 to test if LLM2 correctly identifies and flags the bias, to measure the true positive detection capability of LLM2. The insertion of intentionally biased answers provides the opportunity to test the sensitivity of LLM2 to known bias patterns and the correct identification of true positives.

Human-AI Alignment

The bias explanations produced by LLM2 were validated against the human judgment annotations provided by subject matter experts to verify that the bias scores produced by LLM2 align with human judgment. A comparison of the model-generated bias to the expert-provided annotations serves as a ground truth benchmark to verify if the bias produced by LLM2 aligns with domain-specific judgment.

5.2 Analysis of Design Solutions

5.2.1 In Depth Analysis of LLM1's Performance

Consistency

This figure below demonstrates the overall consistency by book for LLM1 as described in 5.1.1. This is the average cosine similarity value of different books. 5.1

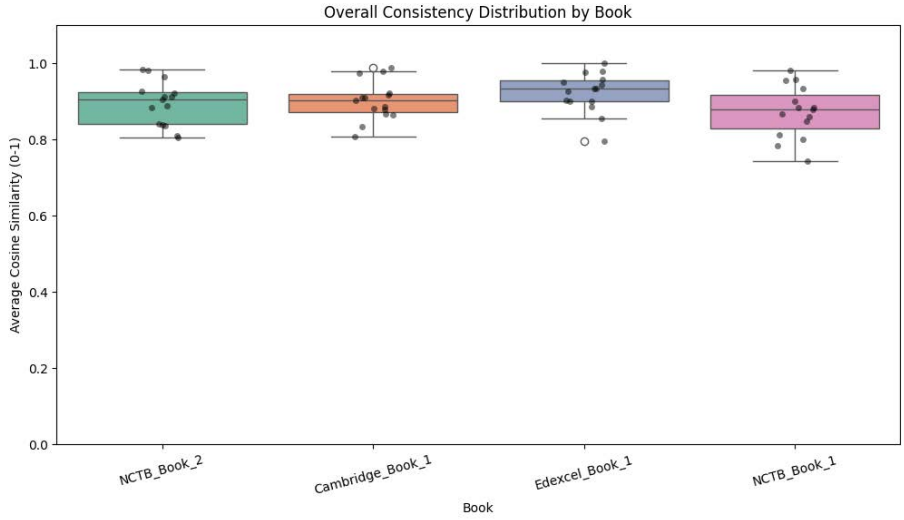


Figure 5.1: Overall Consistency Distribution by Book

These findings demonstrate that our RAG (Retrieval-Augmented Generation) pipeline is a faithful tool in this study. We can rely on the accuracy of the bias scores in the following step of our analysis because the system is able to consistently retrieve and deliver the same narrative of the textbooks.

Source Faithfulness

We conducted a Semantic Faithfulness Check to check that the AI is not introducing any word on its own. We contrasted the generated responses of 482 against the original textbook material in ChromaDB. Through the all-mpnet-base-v2 model, we computed the cosine similarity between the generated answer and the most relevant retrieved segment of a text.

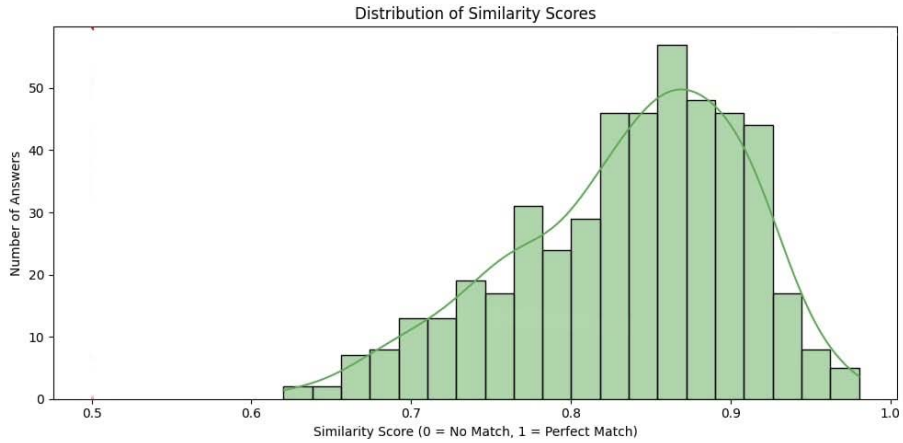


Figure 5.2: Distribution of Similarity Scores

None of the answers were considered a hallucination or a missing context, which means that the AI will always select the right textbook chapters and produce a response. The mean score of 83.60 percent in semantic alignment indicates that even with the AI rephrasing the content to make it understandable, the overall historical sense is maintained.

Measurement of Information Not Included

We assessed the capability of the system in storing and recalling certain information in the textbooks. The current test was called Ground Truth where 38 random chunks of text were picked in the source databases and targeted questions were created. We then transformed the semantic content of the original chunks and compared it to the responses of LLM-1. This process will determine any existing information gaps in which the RAG pipeline may not be able to access or display important information in the source.

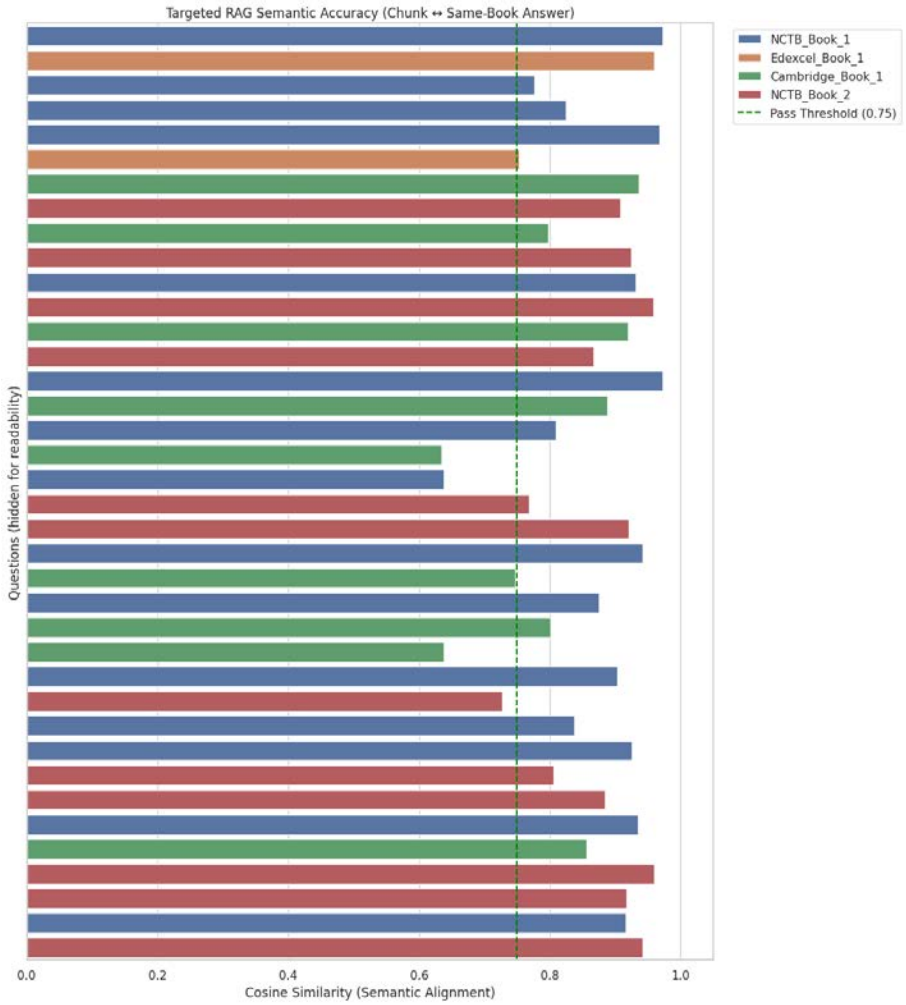


Figure 5.3: Targeted RAG Semantic Accuracy (Chunk Same-Book Answer)

Retrieval-Augmented Generation task, where user queries are correctly mapped to the target semantic space of the textbook [7]. The low proportion of queries that missed the rigid 0.75 mark usually entailed very technical descriptions in which the LLM1 takes the relevant part of answers from the whole book without altering the overall historical context.

Although the majority of queries passed the threshold of similarity of 0.75, some of the cases in NCTB Book 1 were scored low because of internal response formatting, not factual errors of the AI. As an example, in a query regarding the government initiative intended to increase female education, the AI was able to correctly name the Upabritti (stipend) project. But llm1 provided the answer it 2 parts i.e. Primary Fact and a Logical Unit. Consequently, this gave a similarity score of 0.6374 because the repeated and segmented presentation generated a structural mismatch with the original single paragraph textbook chunk. This is however confirmed by manual verification that the information is fully accurate as to mean that such low scores are as a result of technical mismatches caused by formatting and not incorrect or loss of historical content.

5.2.2 Multiple Model Comparison and Inter-Rater Reliability (IRR)

We did cross model semantic analysis of gemini-2.5-flash model, DeepSeek (tngtech/deepseek-r1t2-chimera:free) and LLaMA (meta-llama/llama-3.3-70b-instruct:free) and Mistral (mistralai/devstral-2512:free) pairwise to measure how much the models agree about the interpretation of bias, and therefore, if the detected bias patterns represent a common feature of the text, or are specific to one particular model.

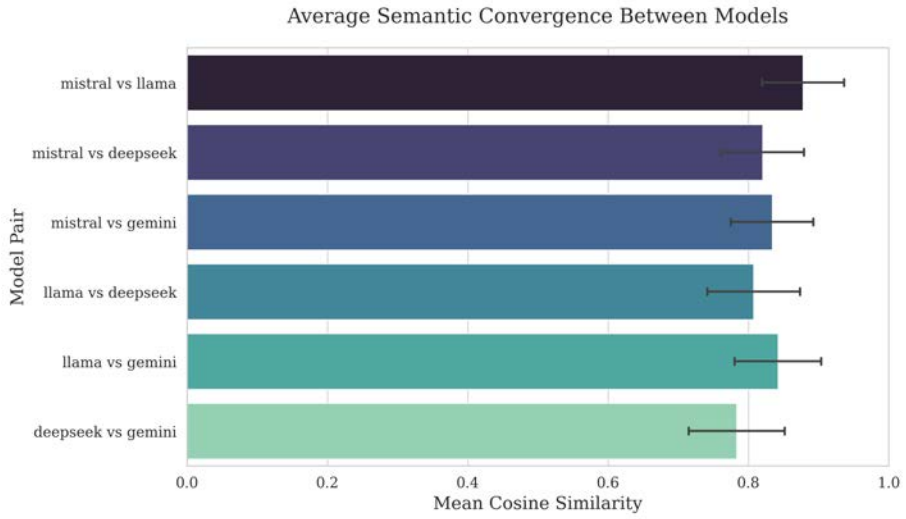


Figure 5.4: Average Semantic Convergence Between Models

The overall similarity comparison did show some level of agreement across models, but the similarity comparison itself was not sufficient to definitively select the best model for detecting bias. That is because semantic similarity does not reflect the depth of the analysis or the accuracy of the analysis; therefore, an expert validation was completed to resolve the disagreements among the models.

Human Expert Model Validation

An "expert-in-the-loop" validation of model performance was also completed to validate model performance through an external evaluation. Model identities were anonymized to prevent partiality. Experts assessed the extent to which each model's analysis aligned with their own judgment and provided agreement scores along with qualitative justifications. About 30 question-analysis pairs for each bias type were presented to the experts. These questions were selected based on their dissimilarities between different models.

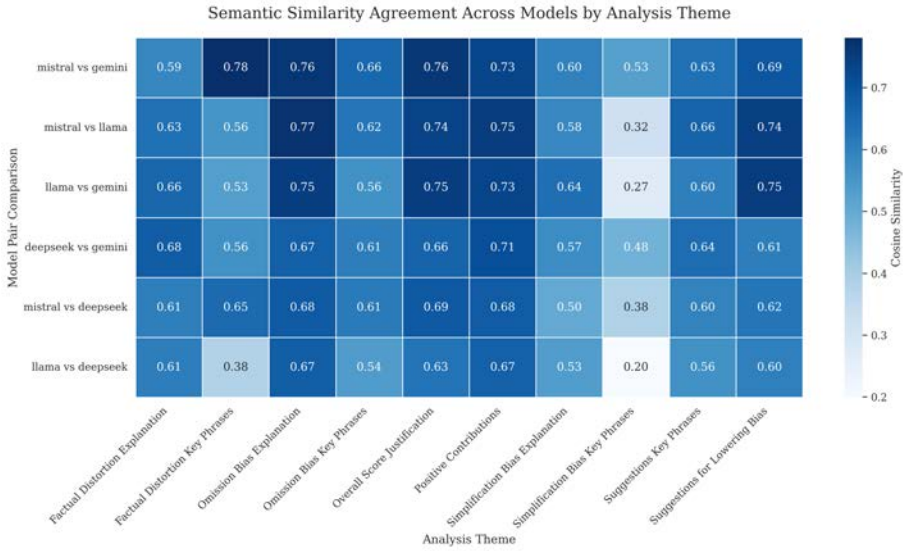


Figure 5.5: Semantic Similarity Agreement Between Models

When models disagreed on detecting bias, the questions were escalated to a focused review to determine which model(s) would provide the most accurate and theoretically correct explanation for the findings. This process provided a systematic method for resolving disagreements among models and selecting the best model for use in subsequent LLM2-based analyses.

Gemini-2.5-Flash was found to have the highest average agreement score from the expert reviewer across various bias categories and questions, and especially in identifying simplification bias, where other models often failed to detect bias signals.
Example: Simplification Bias Detection

- **Gemini (Score: 8):** Identified absolutist claims (e.g., "first case," "biggest aid by air"), mono-causal attribution of agency, and oversimplification of institutional roles lacking political context.
- **DeepSeek (Score: 0):** Failed.
- **LLAMA (Score: 3):** Detected limited overgeneralization.
- **Mistral (Score: 2):** Identified partial oversimplification.

After completing the expert review for all questions and bias types where the models disagreed, Gemini-2.5-Flash had the greatest analytical sensitivity and matched the reasoning of the expert evaluators. Therefore, Gemini-2.5-Flash was selected as the primary model for all subsequent LLM2-based analyses.

5.2.3 LLM2 Consistency Evaluation

Content Integrity Consistency

Stability testing of LLM2 output was completed by running 10 groups of questions (1-10), each group run five separate times. Similarity scores were calculated for each textbook, and for three types of bias: content integrity bias, framing bias,

and representation bias. High similarity scores for intra-group similarity scores will indicate that LLM2 produces consistent assessments of bias when identical input is repeated under the same conditions.

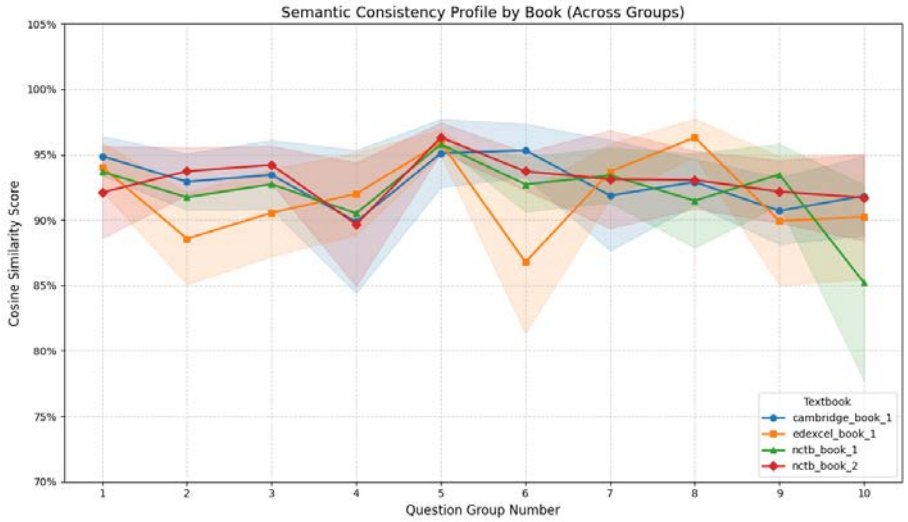


Figure 5.6: Semantic Consistency Profile by Book

The average similarity scores across the ten groups ranges from 0.8975 to 0.9577, showing very little deviation from the mean value. These results indicate that LLM2 has been tested repeatedly with the same inputs, and it produces very similar output each time the input is the same. Most of the groups have small standard deviations, indicating that there is very little variation in the output for each run of the input.

Some groups exhibited very high levels of consistency, including groups 5, 1, 8, and 7. These groups showed not only high averages, but also low standard deviations. It can be inferred from this that questions with clear structures, or questions with well defined narratives are likely to produce more consistent responses.

Two of the groups, Groups 4 and 10, showed lower averages and higher variability than the previously mentioned groups. It can be inferred from this data that questions that may contain nuanced interpretations or context may be more sensitive to changes in the model.

The narrow range of similarity scores, and the low variability overall, demonstrates that the system has good test-retest consistency across the variety of questions that were used. This consistency provides a solid base for future bias analysis, and allows for confidence that observed bias patterns are not due to randomness in the generated response. The consistencies across the rest of the biases are explored in Section 5.4.

5.2.4 Phrase Matching and Evidence Grounding Verification

Comparative evidence grounding was evaluated by comparing phrases identified in LLM2's bias analysis to phrases in the corresponding answer generated by LLM1.

Before comparison, both the original text and the answer text were normalized. Normalization included converting text to lowercase, removing markdown symbols, punctuation, and collapsing whitespace. Phrase segmentation was accomplished using multiple dot sequences and Unicode ellipses.

Match percentage comparisons were completed for each bias category and textbook. The match percentages are listed in the tables below.

Book Name	**Match Percentage (avg)**
nctb_book_2	89.77%
cambridge_book_1	87.64%
edexcel_book_1	75.41%
nctb_book_1	67.92%

Table 5.1: Framing match percentages by book.

Book Name	**Match Percentage (avg)**
nctb_book_2	89.77%
cambridge_book_1	87.64%
edexcel_book_1	75.41%
nctb_book_1	67.92%

Table 5.2: Representation match percentages by book.

Book Name	**Match Percentage (avg)**
nctb_book_1	93.76%
edexcel_book_1	83.61%
cambridge_book_1	82.35%
nctb_book_2	76.02%

Table 5.3: Content integrity match percentages by book.

When some responses had low phrase-level alignment, the responses were then qualitatively reviewed and many times were found to be legitimate and reasonable once they were more closely examined. In addition, sometimes there were no explicit formulations in the source materials of terms such as "focus on military and politics" or "narrative dominated by males," but it seemed that these formulations did represent recurring themes and/or omissions embedded in the larger discourse. It appears that these types of instances represent a form of interpretive abstraction and not the creation of spurious information; however, alternative interpretations of the same passages cannot be completely ruled out.

5.2.5 Query Sensitivity Analysis

Another study analyzed how slight variations in wording in the questions posed to the model affected its outputs. One hundred questions derived from one textbook

were therefore slightly reworded and analyzed through the lens of three predetermined bias categories.

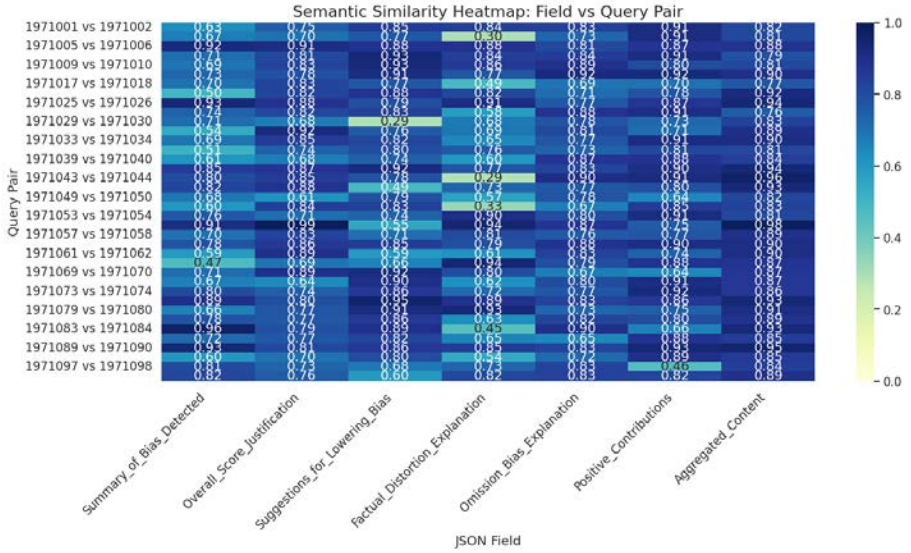


Figure 5.7: Query Sensitivity Analysis

We can see, across 50 pairs, there are four scores which are less than 0.4. This happened because sometimes, for Factual_Distortion_Explanation, the LLM2 did not explain anything, because there was no Factual Distortion. For this graph generation, this empty field was manually filled by "This was not included by LLM2". For the same pair, even though there was no Factual Distortion, the LLM2 added "There was no Factual Distortion present in the Answer". Because of these two incidents, the similarity score became low. For 96% of the time, the aggregated score was greater than 80%, meaning the provided my LLM2 analysis is mostly similar.

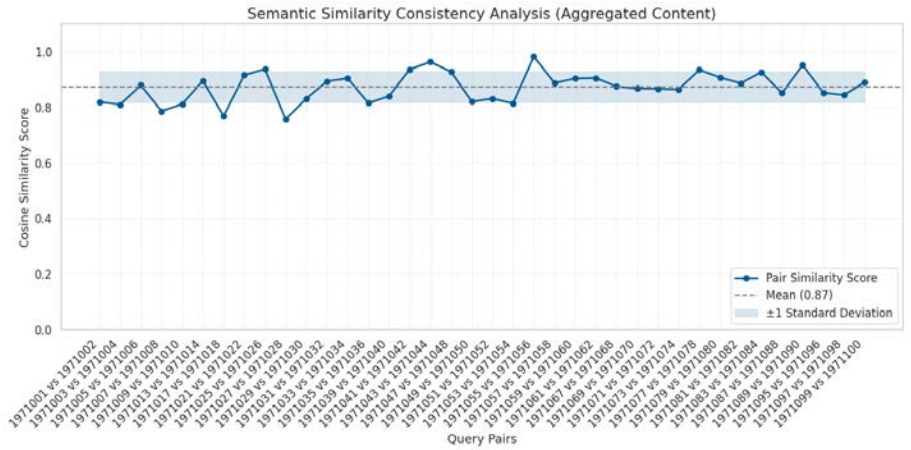


Figure 5.8: Query Sensitivity Analysis

From this figure, we can see that the semantic meaning of a pair has a mean of 0.87 which is very consistent.

5.2.6 LLM3 Judge Evaluation System

To further investigate whether the bias identification made by LLM2 was based on actual textual evidence rather than exaggeration/fabrication, an independent LLM3 Judge Evaluation System was developed. The purpose of this component is to provide an additional layer of validation to verify whether or not LLM2's bias scores and rationales are based on valid textual evidence.

Within this framework, LLM3 is directed to act as a Senior Bias Auditor, responsible for evaluating and validating analyses created by LLM2, which is acting as a Junior Sociological Analyst. For each case, the judge model receives the textbook response, LLM2's bias rationale, and the relevant excerpts from the textbook.

The main goal of this evaluation phase is to determine if a particular bias assignment is supported by sufficient evidence, insufficient justification, or contains contradictions.

Audit Parameters

Each audit is evaluated on three parameters:

Evidence Check: Determines whether the referenced statements/phrases are clearly stated in the LLM1 Answer.

Logic Check: Verifies whether the rationalization presented supports the assigned bias score; for example, when high distortion scores are assigned to factually correct passages, the logic check identifies this as an error.

Displacement Check: Evaluates whether assertions of structural displacement refer to a demonstrated shift in emphasis or focus within the narrative of the text.

LLaMA-70B Performance (Judge A)

Test Configuration:

- **Test Set 1:** 100 questions with analysis-answer pair combinations for each of the three bias categories.
- **Test Set 2:** 100 questions with analysis-answer pair combinations for each of the three bias categories that also include the four textbooks.

Results: No errors were detected during all test runs, indicating a very high degree of reliability and consistency.

Bias Testing of LLM3

In order to better test the reliability of the judge system to detect false/biased/contradictory bias assignments, controlled bias insertions were intentionally added to selected LLM2 outputs. The bias testing represented a positive control to determine whether LLM3 could consistently identify fabricated, unsubstantiated, or logically contradictory bias assignments.

A total of four specific types of bias were tested:

Bias Type	**Detection Description**	**Target Verdict Field**	**Detection Rate (%)**	**Failure Rate (%)**
Factual Hallucination	The answer was right, but LLM2 verdict was this answer contains Factual Distortion	factual_verdict	100.0	0.0
Omission Blindness	The answer did not have any omission, but LLM2 verdict was the answer omitted severe key details	omission_verdict	100.0	0.0
Displacement	The answer contained all the information, but LLM2 verdict was the Answer omits crucial details which is unrelated to the question	displacement_verdict	100.0	0.0
Logic Mismatch	The score of LLM2 did not match its explanation	simplification_verdict	63.0	37.0

Table 5.4: Analysis of bias types, detection descriptions, and corresponding rates.

Across controlled bias tests, the LLM3 evaluation framework demonstrated significant capability to identify hallucinated evidence, fabricated omissions, false displacement claims, and major logical inconsistencies. These findings show that LLM2's bias detections — when validated through LLM3 — are grounded, logically consistent and resistant to systematic false positives.

5.2.7 Discriminant Validity (Negative Control Testing)

To determine the discriminant validity of LLM2, a negative control testing framework was used to determine if LLM2 incorrectly attributes bias to unbiased content. To conduct this test, LLM3 verified, neutral answers were given as inputs which represented "perfect" cases where bias scores would ideally be zero for all bias categories.

Bias scores were generated for factual distortion, omission bias, and simplification

bias based on four textbooks. For each bias category, the mean and standard deviation of scores were generated for all test instances.

Book Name	Factual Distortion		Omission Bias		Simplification Bias	
	Avg	Std	Avg	Std	Avg	Std
cambridge_book_1	0.0000	0.0000	0.6875	1.6905	0.0000	0.0000
edexcel_book_1	0.2245	1.2122	0.7551	1.5881	0.0000	0.0000
nctb_book_1	0.1667	1.1547	0.9375	1.8382	0.0000	0.0000
nctb_book_2	0.2917	1.4286	0.6667	1.3579	0.1667	1.0176

Table 5.5: Average and standard deviation of bias scores across different textbooks.

As seen in Table 5.5, LLM2 produced zero or near-zero simplification bias scores across all textbooks, with minimal variability, thereby demonstrating high discriminant validity for this bias category. This indicates that LLM2 will not improperly attribute simplification bias when none exists.

Similarly, factual distortion scores were very low for all textbooks ($mean \leq 0.29$), with small standard deviations, showing low rates of false positives. The minor non-zero values reflect conservative sensitivity as opposed to misclassifications.

Omission bias had relatively greater average scores and variability, implying that omission detection is more sensitive and can produce false positives for absent content even in neutral responses. Nonetheless, the low mean values and moderate standard deviations suggest that LLM2 generally does not misattribute omission bias to unbiased content in negative control conditions.

In summary, the findings above demonstrate that LLM2 is capable of distinguishing biased content from unbiased content, thus providing support for LLM2's validity by demonstrating a low rate of false positive bias detection in controlled neutral input conditions.

5.2.8 Positive Control Testing

To assess LLM2's sensitivity to known, intentionally introduced bias signals, bias testing was performed in a positive control condition. During this testing, unbiased answers were modified to include specific bias types, and LLM2's capacity to properly flag these examples as biased (i.e., true positives) was determined.

Biased Methodology

A gold-standard textbook answer was intentionally altered in a controlled manner to create a biased answer that maintained surface fluency but included a specific bias category. An example of how an answer was biased for one question related to content integrity bias is as follows:

Bias Type	Sub Type	Original	Modified	Description
Factual Distortion	number_error	Sector 11, particularly in the Mymensingh region.	Sector 9, particularly in the Mymensingh region.	The military sector where Taramon Bibi operated was incorrectly changed from 11 to 9.
Omission	downplaying	guerrilla attacks, sabotage missions, and direct skirmishes	*[OMITTED]*	This list of specific combat activities was removed from the general description of women's roles, downplaying women's engagement in direct armed conflict.
Simplification	absolute_language	The involvement of women in direct armed resistance, though not always extensively documented, was an integral part of the liberation struggle.	While a few rare instances of direct combat involvement are recorded, the majority of women's vital contributions were solely logistical and intelligence-based...	The use of absolute language ("solely") and the sweeping generalization ("the majority") simplifies a complex military contribution.

Table 5.6: Examples of data bias types with original and modified text samples.

Each biased answer preserved grammatical correctness, thereby allowing LLM2 to rely on semantic and structural analysis and not simply surface errors.

Evaluation Protocol

biased samples were analyzed using the exact LLM2 pipeline that was applied to unbiased answers. Bias type-specific accuracy was used to measure the performance of LLM2 to determine if it could accurately recognize known bias signals in controlled environments. Additionally, single prompt evaluations were compared to category level performance of baseline performance to isolate the model's ability to recognize known bias signals under controlled conditions.

Results and Sensitivity Analysis

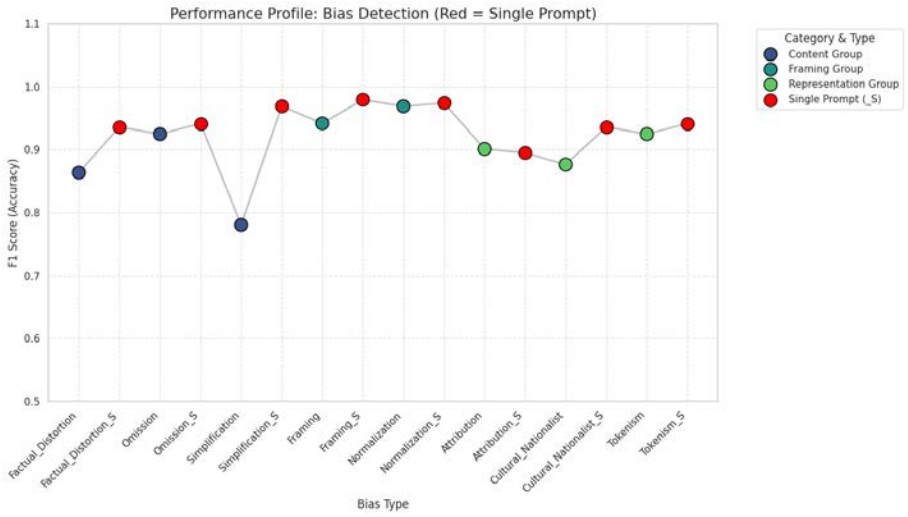


Figure 5.9: Results and Sensitivity Analysis

Figure shows the performance of LLM2 during bias-induced testing. LLM2 has high sensitivity across the majority of bias categories with F1-scores consistently greater than 0.90. Single-prompt evaluations (red markers) have similar performance to category level analyses, indicating that LLM2 performs well regardless of iterative prompts. The red dots in the graph means those tests were done in a single api call, and the other colors represent those that are performed in groups. Therefore, we can see that single LLM calls perform about 5 - 25% better for biases. However, to minimize cost, we proceeded to do group api calls i.e., analyzing omission, simplification and factual distortion together rather than single LLM calls i.e., analyzing omission alone. To mitigate the lost accuracy, we adopted asking the same type of questions two times with different wordings (discussed in chapter 4.2.1).

5.2.9 Human AI alignment

This analysis aims to determine how effective the Large Language Model (LLM) is in auditing educational content by comparing its results with the so-called ground truth developed by a professional sociologist.The study measures the LLM's capability to accurately detect three specific categories of bias: Factual, Omission, and Simplification. By employing a confusion matrix. We can quantitatively determine how closely the LLM2"s (Ai) logic aligns with human expert judgment, Using a confusion matrix.

Basically, we have chosen four standard outcomes to correlate the LLM's predictions and sociologist's opinions. Four outcomes are- True Positive (TP) arises when the text has a certain bias and the LLM detects this bias. On the other hand, a True Negative (TN) will be counted when the LLM states no bias, and the text is found to be neutral .Detection errors are also categorized to assess the limitations of the model. False Positive (FP) occurs when the LLM identifies a section to be biased when the text does not have any bias. When the text is biased, but the LLM does not, then a False Negative (FN) is experienced. With the help of these four results,

reliability of the LLM as a means of large-scale sociological research can be obtained.

At first, we analyzed 150 queries for 6 kinds of biases and labeled them as TP,TN,FP,FN manually for each bias of the specific query. That's how we analyzed all the queries. After that, we had given 30 analyzed questions along with LLM's explanation and labeled results (conducted manually by us) to the sociologist. Then the sociologist went through our analysis and validated it by marking the analysis as right or wrong.

In this case study, the bias detection framework was tried against a query about wartime policies in the Bengal Famine of 1769-70. The LLM first made a factual verification and found the contents in the textbook to be accurate (Score 0), because the dates and references to the 1765 Diwani Treaty matched with its general database. But a sociologist who served as the ground truth found a False Negative, a minor factual inaccuracy that the LLM did not notice: the textbook said that the treaty permitted the Mir Jafar and Mir Qasim to rule, but history records that Mir Jafar had previously been killed in 1765. This treaty was in fact negotiated between him and his successor Najimuddin Ali Khan and the Mughal Emperor. This contradiction demonstrates a serious drawback of the automated system: although the LLM can confirm general timelines it might not pick out particular instances of historical errors, which then need a human sociologist to fix a "True Negative" into a "False Negative" to distort the facts.

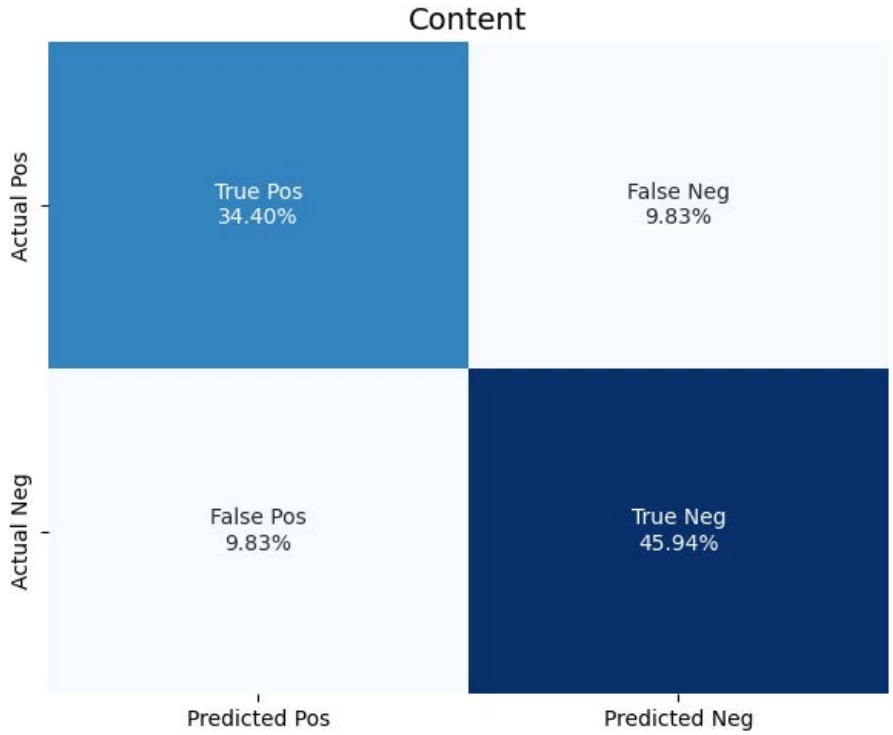


Figure 5.10: Content Bias Confusion Matrix

System performance in identifying Content Bias, which has such dimensions as Simplification/Reductionism, Omission, and Factual Distortion, was assessed to have narrative integrity. The system had a True Positive (TP) rate of 34.40% quantitatively which is the discovery of instances when the historical content had been

stripped of its nuance or accuracy. This was accompanied by a True Negative (TN) rate of 45.94% in which the model made a correct validation of objective and complete historical accounts.

The content related assessment error margin was also evenly distributed and tightly controlled with the False Positive (FP) and the False Negative (FN) rates standing at 9.83 each. These values demonstrate that although the model is very effective in exposing gaps in information, there is also a small overlap in that one pedagogical simplification is sometimes indicated as an omission (FP) or very subtle reductionism is avoided (FN). Altogether, the high TN rate proves that the LLM-2 module is a conservative and stable auditor, who is greatly interested in grounding the facts and preferences that the most objective materials are identified without any interference.

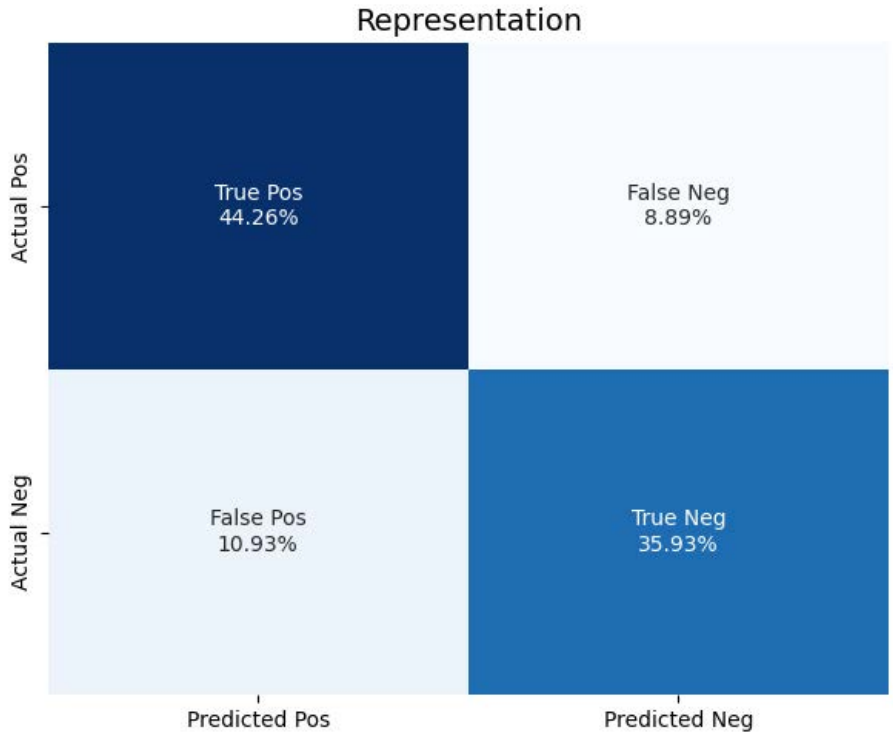


Figure 5.11: Representation Bias Confusion Matrix

The analysis of the system performance in terms of the dimensions of representation bias, Attribution, Cultural/Nationalist, and Tokenism, proves a high level of reliability and a strong possibility to differentiate between biased and objective historical narration. The system had a True Positive (TP) rate of 44.26, which is able to detect subtle historical falsities, and a True Negative (TN) rate of 35.93, which is able to confirm that the objective passages are not biased. The error rates were kept down to a minimum, and the False Positive (FP) rate was 10.93, and False Negative (FN) rate was very low (8.89), meaning that the model did not miss any important representational gaps with much tolerance.

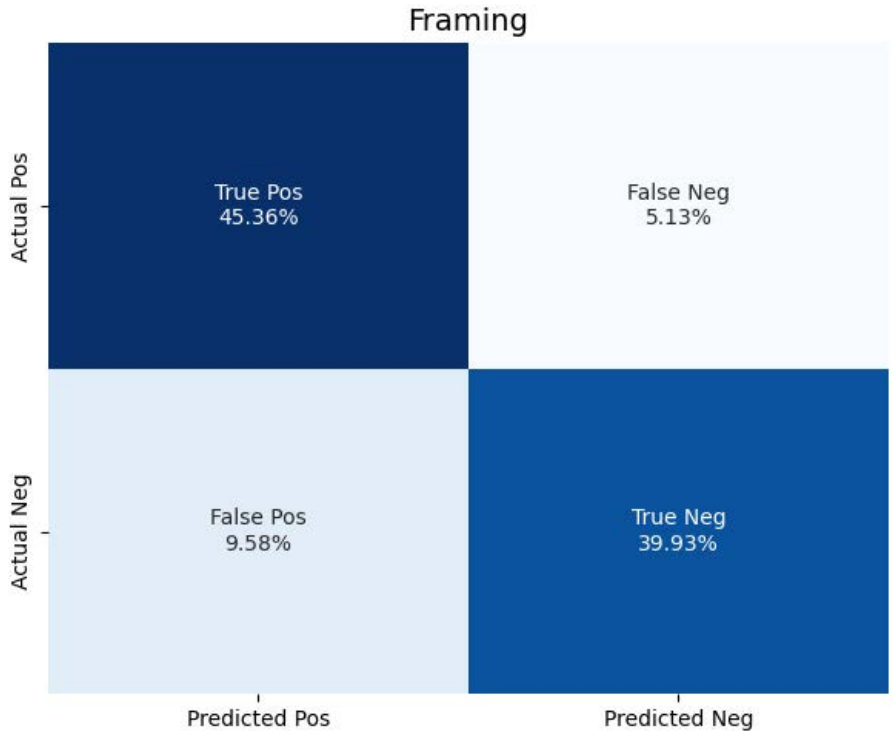


Figure 5.12: Framing Bias Confusion Matrix

The assessment of Framing Bias, including narrative framing and normalization bias, proves that the system manages to detect subjective linguistic patterns that impact the perception of the reader of the historical events. On a quantitative level, the system performed at 45.36 in terms of True Positive (TP) rate, where the system correctly identified the cases when the textbook applied biased framing to make certain ideological views normal. This performance was equalized with True Negative (TN) rate of 39.93 during which the model was able to recognize objective and neutrally framed historical accounts.

The margins of errors of this category were also strictly regulated. False Positive (FP) rate was 9.58% which is a rare occurrence whereby the model perceived neutral pedagogical language to be biased. More importantly, the False Negative (FN) rate was 5.13 only, which demonstrates that the model is very sensitive and does not overlook the most critical case of narrative framing or normalization. Altogether, these findings prove that the LLM-2 module is a very efficient auditor that is sensitive to linguistic nuance and involves a legitimate automated protection of the normalization of biased historical stories.

The high true classification rate obtained supports the fact that the LLM-2 module is a reliable and useful tool to representational management in historical texts. Although the system is very good at identifying missing information and classifying bias, the presence of false positives albeit small is an indication of the value of the human-in-the-loop importance of verifying the ground truth and making sure that the simplified instructional language is not being interpreted as a form of deliberate bias.

Although the LLM is very good in discovering gaps in information, it can still believe false facts when they are stated with confidence in the text where they are found. This underscores the fact that even today, a human sociologist is needed to confirm the ground truth

5.3 Final Design Decisions

Based upon results of robustness, validity, and expert alignment evaluations, several final design decisions were made to enhance the performance and reliability of LLM2.

Final Model Selection

Gemini-2.5-Flash was chosen as the final model used to analyze LLM2 after an extensive comparison of multiple models and validation of expert judgments. The decision to choose Gemini-2.5-Flash was largely based upon Gemini's superior agreement with expert judgments across multiple bias categories, especially when it came to identifying and evaluating simplification bias, which many of the other models failed to evaluate well. All of the subsequent analyses and experiments referenced above were subsequently performed utilizing Gemini-2.5-Flash to ensure analytical consistency.

Improving Retrieval Quality

The initial analysis revealed that retrieval quality severely impacted the performance of LLM2 when detecting bias downstream. To mitigate this, the number of retrieved context chunks was increased from top-k = 5 to top-k = 10, thereby increasing the amount of relevant textual evidence available to support downstream bias detection and decreasing the likelihood of false negatives resulting from missing contextual information. Additionally, the retrieved chunks were sorted and grouped based upon their proximity to each other within the source textbook, thus preserving the narrative flow and improving LLM2's ability to reason over an extended context versus fragmented pieces of information.

Refining Prompts

Prompts for both LLM1 (answer generation) and LLM2 (bias analysis) were iteratively revised to correct identified failure modes during evaluation. The modifications addressed the clarity of task boundaries, promoted evidence-based reasoning, and minimized the influence of prompt wording on the output. Overall, these revisions enhanced the overall consistency of the outputs, reduced the occurrence of hallucinated bias detection, and increased the degree of alignment with expected outcomes from experts.

Rephrased Questions

To address the missing accuracy (described in 5.2.8) we made questions in pairs where the modified question contained the semantic value of the base question but worded slightly differently. Therefore, each question is being run two times, instead

of one. This method increases the chance of finding biases that LLM2 might miss for the previous question.

5.4 Statistical Analysis

All statistical analyses in this study were completed using semantic similarity scores generated using the all-mpnet-base-v2 sentence embedding model. This model was chosen because of its excellent performance in detecting semantic equivalence and contextual similarity in comparisons of long-form text.

Breakdown of LLM2's Consistency

In addition to demonstrating the consistency of LLM2 regarding content integrity bias, we examined the distribution of LLM2's consistency for the remaining biases as discussed above.

Question Group	**Mean Similarity (μ)**	**Standard Deviation (σ)**
Group 1	0.936667	0.023000
Group 2	0.917333	0.024667
Group 3	0.927333	0.024333
Group 4	0.905333	0.044333
Group 5	0.957667	0.015667
Group 6	0.921250	0.027250
Group 7	0.930250	0.030250
Group 8	0.934500	0.022750
Group 9	0.915750	0.030500
Group 10	0.897500	0.046750

Table 5.7: Content Integrity Consistency across question groups.

Question Group	**Mean Similarity (μ)**	**Standard Deviation (σ)**
Group 1	0.900333	0.033667
Group 2	0.946333	0.019667
Group 3	0.898667	0.043000
Group 4	0.922667	0.040000
Group 5	0.898667	0.045667
Group 6	0.924750	0.019500
Group 7	0.896000	0.032500
Group 8	0.870250	0.049250
Group 9	0.887750	0.042500
Group 10	0.895500	0.045500

Table 5.8: Framing Consistency across question groups.

Question Group	Mean Similarity (μ)	Standard Deviation (σ)
Group 1	0.900333	0.033667
Group 2	0.946333	0.019667
Group 3	0.898667	0.043000
Group 4	0.922667	0.040000
Group 5	0.898667	0.045667
Group 6	0.924750	0.019500
Group 7	0.896000	0.032500
Group 8	0.870250	0.049250
Group 9	0.887750	0.042500
Group 10	0.895500	0.045500

Table 5.9: Representation Consistency across question groups.

Deduction from Consistency Analysis Across Dimensions

The system has shown strong test-retest stability across each of the three dimensions examined: Content Integrity; Framing; Representation. Most question groups have produced mean similarity scores near or at 0.90, and nearly all have produced similarities that are statistically significantly different from chance, thus confirming that repeated runs have produced semantically similar output; i.e., that the system has demonstrated reliable performance under identical input conditions.

Content Integrity exhibited the greatest degree of stability across all of the systems tested. In addition to high mean similarity scores across almost all groups, very few groups had large standard deviations — a result that would suggest that the system provides a reliable means of preserving the factual structure and core informational content of substantive historical information. Therefore, we believe that this system can be used as a solid base for evaluating potential biases in downstream analyses.

Framing consistency shows slightly greater variability. While mean similarity scores remain high, standard deviations are generally larger than those observed for content integrity. This pattern indicates that while the core narrative frame is preserved, the model allows modest variation in emphasis, tone, or evaluative phrasing across runs. Such variation is expected for framing-related aspects, which are inherently more interpretive than factual content.

Consistency in framing was somewhat less variable than content integrity. However, while mean similarity scores were again high, standard deviations were much larger than those observed for content integrity. This result suggests that, while the core narrative frame that is presented in the output is preserved, the model permits some amount of variability in the way that an output emphasizes, tones or evaluates the subject matter across runs.

The level of **consistency of representation** was virtually indistinguishable from that of framing — both in terms of mean similarity and variability. This similarity suggests that representational choices — such as which actors are placed in the

foreground or background of an output — will vary more in response to nuances of context than do representations of core factual content. However, these choices will also be relatively stable across repeated evaluations.

Regardless of whether content integrity, framing, or representation were being evaluated, several question groups (Groups 4, 8 and 10) consistently produced similarity values that were somewhat smaller and standard deviations that were somewhat larger than the others. Since this same ranking of question group variability occurred across all of the evaluation dimensions that were assessed, we conclude that the variability of the framework is much more related to the interpretive complexity of individual questions than to variability in any of the specific evaluation dimensions.

In conclusion, our results clearly demonstrate a hierarchy of stability among the three frameworks evaluated: Content Integrity >Framing >Representation, with the latter exhibiting a controlled degree of variability relative to the former two. These results confirm that the framework evaluated in this study produces a reliable preservation of core historical content while permitting a limited degree of variability in the narrative and representational expressions of that content — a feature that is essential for performing nuanced bias analyses as opposed to simply replicating outputs.

5.5 Comparisons and Relationships

5.5.1 Comparison Between Different Approach Analysis

Single-LLM vs Dual-LLM

For the single-LLM approach, we implied that the model takes (or reads) all the similar chunks and a bias evaluation in a single pass. The pipeline is built to process the content with a Dual-LLM (two-stage) pipeline, where the first stage (LLM1) retrieves only the relevant content chunks and feeds the intermediary result into a second stage (LLM2) dedicated to bias analysis. This is similar to multi-step/reasoning pipelines which break tasks into smaller ones. The integration of RAG was maintained in both cases. Single-LLM approaches offer benefits such as reduced orchestration work, lower latency, and the ability to handle large context windows effectively, making them suitable for tasks requiring joint reasoning or interactive QA. However, they complicate prompt engineering and may lead to inaccuracies when faced with complex tasks due to a single pass. Conversely, Dual-LLM systems excel in handling multi-step tasks more accurately by breaking them down into specialized subtasks, thus enhancing accuracy but at the cost of increased latency and computational complexity. These systems are best suited for long or intricate texts but risk propagating errors through each stage. In terms of accuracy and hallucination, According to recent research, text analysis reliability may be enhanced by multi-stage LLM pipelines (such as distinct generator and evaluator models). According to Li et al. (2025), for example, a three-agent meta-judging pipeline (three LLM "judge" models) produces approximately 15% more precision in assessing model outputs than a single LLM completing both tasks [25]. The

scalability of Single-LLM can falter with limited context size, contrasting with the more manageable input sizes of Dual-LLM. Finally, in educational contexts, while Dual-LLM might miss subtle biases, Single-LLM retains direct relationships with original texts, allowing better bias analysis while summarizing key issues effectively.

Aspect	**Single-LLM**	**Dual-LLM (Two-Stage)**
Definition	One model call performs chunk retrieval and bias detection (often with RAG).	Two sequential model calls: first answer generation, then analyze summary for bias.
Pros	Simpler flow; lower latency; leverages large context (e.g., Gemini's million-token window[8]); RAG-grounded (factual) outputs.	Specialized sub-task; may leverage extractive summaries for factual consistency; additionally useful for faster and better outputs at long tasks.
Cons	Complicated prompts; chance of task confusion and hallucination; binary inference.	Higher latency (multiple LLM calls); error propagation if summary is incorrect; increased system complexity.
Suitability	Best from context or resolved input; okay for short to medium length, well-scoped texts.	Ideal for very long or complex texts (breaking into summary); useful when explicit summarization supports review.
Accuracy / Hallucination	With retrieval, it significantly reduces hallucination. Accuracy relies on the model's ability to multitask; may lose some details.	Although stepwise analysis can increase accuracy, each step must be precise. Biases left out of the summary may be missed by secondary analysis.
Latency / Scalability	One inference; usually quicker per document. Scalability is restricted by window and context.	Multi-inference; at least 2× latency[7]. Uses summarization to handle longer documents, but needs more processing power.

Table 5.10: Comparison of Single-LLM vs. Dual-LLM pipelines.

Case Study:
Query: 1971003. How did women contribute to the 1971 liberation war of Bangladesh?
Book: NCTB_book_2
Single LLM Analysis: "The provided answer contains no verifiable factual errors. There are no contradictions of consensus or fabricated details."
Dual LLM Analysis: "The answer incorrectly identifies Nikita Khrushchev as the Soviet Head of State who urged Pakistan President Yahiya Khan to end atrocities in 1971. Nikita Khrushchev was removed from power in 1964, and Leonid Brezhnev was the Soviet leader in 1971."
Verdict: A single LLM restrain itself only to question scope. As a result, the analysis was also slightly inaccurate and incomplete.

With RAG vs Without RAG

Retrieval-Augmented Generation (RAG) is the process of indexing the document corpus (e.g., textbooks, curricula) into a vector store (e.g., ChromaDB) and retrieving the most relevant chunks to each query. These passages are the ones that are conditioned by the LLM during the generation passages. No-RAG (Long-Context) models directly give the LLM raw text (chapters of the book), and is only assisted by the parameters and window of the model. RAG enhances factual accuracy and reduces hallucinations by leveraging real sources, allowing for the indexing of large corpora without overwhelming the model. It is effective in educational settings but requires a complex implementation involving vector databases. Conversely, No-RAG simplifies the processing pipeline, providing complete context for shorter documents and demonstrating success in QA tasks, though it struggles with long texts due to context window limitations and higher hallucination risks. While RAG is suited for dynamic and large documents, No-RAG excels where comprehensive input can be provided. Retrieval grounding in RAG improves accuracy and coherence, yet introduces latency and scalability challenges, making it resource-intensive. Both approaches have unique implications for curriculum analysis, with RAG facilitating bias detection through citations, while No-RAG risks losing context in longer materials.

Aspect	RAG (Retrieval-Augmented)	No-RAG (Long-Context / Full-Text)
Definition	Retrieves relevant text segments for every query using an external vector store.	Inputs original textbooks text directly into the LLM's context.
Pros	Enhances currency and factual accuracy through grounding; significantly lowers hallucinations; and uses indexing to manage extremely large corpora.	Large context models are advantageous; the pipeline is simpler (no database); if appropriate, it can process coherent content in a single pass.
Cons	Increased latency and system complexity (search step); potential retrieval errors.	Restricted by the size of the model's context; may omit or shorten information; depends on the static knowledge of the model; increased risk of hallucinations in the absence of grounding.
Suitability	Ideal for tasks requiring citation or current information, as well as for large, dynamic data sets (such as an entire curriculum).	Ideal for documents that fit the model context, such as those with very large-window context LLMs or brief chapters.
Accuracy / Hallucination	Hallucinations are significantly reduced by grounded responses; retriever quality determines accuracy.	Accuracy relies on model memory as well as prompt design; without retrieval, it is more likely to experience hallucinations and factual errors.
Latency / Scalability	Although retrieval increases overhead, it can handle big datasets.	A single inference per document; long inputs are constrained by hardware.

Table 5.11: Comparison of RAG vs. No-RAG approaches.

Case Study:
Query: "How did the Soviet Union shape events?"
Book: NCTB_book_2
RAG: "relevant_evidence_found": ["After India, the highest contributions to the liberation war of Bangladesh was made by the then Soviet Union (Now, Russia).", "The Soviet Head of the Sate, Nikita Khrushchev urged the Pakistan President Yahiya Khan to put an end to the genocide, blazing of public property and atrocities towards women in Bangladesh by Pakistan forces.", "He also asked Yahiya Khan to transfer power to the elected representatives of Bangladesh.", "The Soviet newspapers and media helped create the global mandate by publishing stories on the barbarousness of Pakistani forces in Bangladesh and the advancements of the liberation war.", "Soviet Union exercised veto to discard the proposal of putting an end to the war which was put forwarded by the United States in favour of Pakistan.", "The socialist countries of that time such as, Cuba, Yugoslavia, Poland, Hungary,

Bulgaria, Czechoslovakia, East Germany, etc. also supported the liberation war of Bangladesh."]
No-RAG: "relevant_evidence_found": ["After India, the highest contributions to the liberation war of Bangladesh was made by the then Soviet Union (Now, Russia).", "The Soviet Head of the Sate, Nikita Khrushchev urged the Pakistan President Yahiya Khan to put an end to the genocide, blazing of public property and atrocities towards women in Bangladesh by Pakistan forces.", "He also asked Yahiya Khan to transfer power to the elected representatives of Bangladesh.", "The Soviet newspapers and media helped create the global mandate by publishing stories on the barbarousness of Pakistani forces in Bangladesh and the advancements of the liberation war.", "Soviet Union exercised veto to discard the proposal of putting an end to the war which was put forwarded by the United States in favour of Pakistan.", "The USSR used its veto power in the UN Security Council to block ceasefires." "The socialist countries of that time such as, Cuba, Yugoslavia, Poland, Hungary, Bulgaria, Czechoslovakia, East Germany, etc. also supported the liberation war of Bangladesh."]
Verdict: There was no mention of the USSR in the NCTB 2 book; therefore, this line was "The USSR used its veto power in the UN Security Council to block ceasefires." hallucinated by No-RAG due to heavy dataload.

With vs Without Expert Bias Definition

Now we explore bias detection functioning without the explicit definitions of bias. This approach measures the effectiveness of the system to detect historical biases when it is solely based on the internal knowledge of the Gemini 2.5- flash model. Unlike the definition-directed setting, the without-definition setting does not provide any structured descriptions of bias. Henceforth, it allows the model to perform bias analysis in a more open-ended manner. The comparative analysis thus observes the necessity of complex prompt engineering in the context. To identify the existence of bias, it is divided into eight forms of bias that are classified into three different groups containing content, framing, and representation. These prejudices are established as the forms of definition in the definition-guided setting. The definitions are made in a manner that is easy to understand and comprehend so that the students in the secondary level of education can understand. Such an environment adheres to the Closed-World Assumption (CWA) and a null-set protocol to prevent the detection of biases that do not have enough evidence. Also, a scoring rubric is given to assist in the systematic and consistent bias identification. Quite to the contrary, there are no specific or explicit definitions of bias in the without-definition setting. There is a minimum of instructions given to Gemini whereby it has to identify biases in its own ethical knowledge and logic. No preset guidelines or scoring rubrics are applied in this arrangement. This structure is supposed to assess the performance difference between definition-guided bias detection and bias identification which uses only the internal knowledge of the model.To provide equal comparison of two conditions of experiment, the same set of answers was offered to both groups. The definition prompt was the only variable that differed across the conditions. Accordingly, our conclusions were drawn by comparing the outcome of the two conditions.

The Methodological Comparison Table presents a technical rationale of the research architecture in the form of a comparison between the "Integrated Audit" and the

traditional method of sequential approach.

Evaluation Metric	Integrated Audit (With-Definition)	Sequential Audit (Without-Definition)
System Architecture	Follows Closed World Assumption (CWA). The AI only compares the given text pieces to the certain sociological rules.	The AI analyzes the text based on its internal training data and intuition in general.
Pros	Factual Grounding: Eliminates the possibility of hallucinating because the AI is attached to the source text.	Because of minimal definition, it is easier to perform.
Cons	Needs a high reasoning model to implement complex prompts i.e. narrative displacement (SND) and Null Set Protocol.	Results in noisy data as it does not have a specific rubric. Hence, cannot differentiate between pedagogical summary and nationalist myth.
Strategic Utility	Trustable for academic research. All the scores are associated with a pre-determined bias pillar, which makes the results reproducible and rigorous.	Good at giving the rough overview, however, it does not give the grain evidence necessary to demonstrate structural or omission bias in curricula.

Table 5.12: Comparison of Integrated Audit vs. Sequential Audit methodologies.

The Integrated Audit (with definition) is a two-part system, which works with the source text and bias detection at the same time. This forms a direct tether to the original documents, which eliminates the loss of information that usually happens when an AI initially summarizes a text prior to processing it. The model is factual and can discover nuanced language that a summary could overlook by operating on the fresh source material.

Finally, the integrated approach is much better when utilized in a research project where latent patterns and unreported facts need to be identified. The given methodology ensures that all biases detected by the artificial intelligence are based on the original terminology of the textbook, and not based on a watered-down abstraction.
Case Study:
Query: 1971001. What roles did women play in the context of the 1971 liberation war of Bangladesh?
Book: NCTB_book_2
Without definition: "The facts presented are largely accurate. The statement about women taking part 'directly in the war after having training in handling the firearms and in guerrilla warfare' is true for some women, but the generality might slightly overstate the prevalence of this specific role compared to other forms of participation. However, it is not factually incorrect that some women did engage in direct combat. The number 'over lakhs' for tortured women is an estimate but widely accepted in historical discourse."
With definition: "The answer contains no verifiable factual errors, incorrect dates,

names, numbers, or locations. The information presented aligns with generally accepted historical facts regarding the 1971 Liberation War of Bangladesh and the roles of various groups, including women. There are no contradictions of consensus or fabricated details."
Verdict: The LLM is an opinionated critic in the Without-Definition setting. Contrastingly, the With-Definition setup was right in determining that the text was factual in the context of a curriculum in school.

5.5.2 Findings

1947-1971

1.Portrayal of Foreign Countries against the liberation war

Source	**Countries Mentioned**	**Phrasing Used**	**Tone/Nature of Description**
Cambridge1	USA, China	"Did not support"	Softer Phrasing: Uses negative support ("did not support") rather than active opposition ("were against").
Edexcel 1	*None*	"Omission"	Absent: The topic of foreign opposition appears to be omitted entirely.
NCTB1	"Some countries"	"Opposed the war", "Tabled a proposal... in Pakistan's favour"	Narrative: Provides that USA vetoed the ceasefire, but never mentions they were against the liberation war.
NCTB2	USA, China, Iran, Saudi Arabia, Middle East	"Was against", "Were against"	Explicit/Direct: Uses repetitive, blunt language to explicitly list every country that opposed the war.

Table 5.13: Comparative analysis of how foreign opposition is described across textbooks.

2.Portrayal of communal violence and displacement in 1947

Source	Portrayal of Suffering	Linguistic Strategy	Key Effect
Edexcel 1	Statistical & Detached	Passive Voice & Bureaucratic Tone: Uses quantitative terms ("600,000 people") and passive verbs ("were slaughtered") to state facts without emotional depth.	Dehumanization: Suffering is treated as a statistic or logistical failure. The narrative centers on British "efficiency" rather than the lived reality of the victims.
Cambridge 1	Reciprocal & Organized	Active Verbs & Balancing: Uses strong, active words ("orchestrated," "massacre") but immediately "balances" the blame between groups.	Politicization: The focus shifts from the pain of the victims to the culpability of the factions. It treats violence as a political "scorecard."
NCTB 1	Total Omission	Complete Silence: The entire context of 1947 violence and suffering is absent.	Historical Void.
NCTB 2	Political Detail / Human Silence	Selective Narration: Elaborates on political causes/reasons but excludes descriptions of displacement or violence.	Sanitization: The Partition is framed solely as a political/administrative event (borders, treaties, leaders), erasing the human cost (refugees, death) to maintain a "clean" political history.

Table 5.14: Comparative analysis of the portrayal of suffering in 1947 partition narratives.

3.Portrayal of Women

Source	Portrayal of Contribution	Portrayal of Violence/Suffering	Linguistic Strategy/Bias
Cambridge 1	Selective: Mentions "huge numbers" fought, but narrows focus to a single biography (Begum Sufia Kamal).	Sanitized: Mentions "genocide" and killing of "men, women and children," but omits rape.	Tokenism: Highlights participation but restricts the detailed narrative to one famous individual, potentially overlooking the grassroots reality.
Edexcel 1	Displaced: Complete omission of Bangladeshi women. Focuses on Indira Gandhi (an external figure).	Erased: No mention of rape or gender-specific violence.	Exclusion/Great Man Theory: Women are only relevant if they are powerful heads of state; ordinary women are invisible.
NCTB 1	Contradictory: Initially includes women, but excludes them with the final honorific "sons of the soil."	Objectified: Groups "molestation" with property crimes ("booty", "arson"). Uses passive voice ("were tortured").	Commoditization: By listing violence against women alongside arson and looting, the text linguistically treats women's bodies as property damaged during war, rather than human victims.
NCTB 2	Supplementary: "Beside the males" positions women as secondary helpers.	Moralized: Uses emotionally charged, archaic terms like "cruel lust."	Paternalism: Frames the status of Birangana not as a right earned by sacrifice, but as a gift given by Bangabandhu "with the affection of a father."

Table 5.15: Comparative analysis of the representation of women in 1971 war narratives.

4.Portrayal of Sheikh Mujib

Source	**Naming Convention**	**Role & Titles**	**Linguistic Tone**	**Primary Narrative Function**
NCTB 1	Bangabandhu	President, Commander in Chief of the Liberation War	Formal & Reverent: Uses official state titles to establish legal and military authority.	The Authority: Establishes him as the legal head of the state and the war effort.
NCTB 2	Bangabandhu	Architect of independent Bangladesh, Great leader, Invincible role	Hagiographic (Hero-Worship): Uses superlatives ("best in world history," "invincible") and emotional language.	The Savior: Frames him as a spiritual and emotional anchor for the nation, emphasizing personal sacrifice.
Cambridge 1	Sheikh Mujib / Mujib	Leader of East Bengal, President, Next President of Pakistan	Functional / Political: Uses standard political titles. Omits the honorific "Bangabandhu."	The Politician: Frames him as a key political player and elected official within a specific historical context.
Edexcel 1	*(Implied Standard)*	*(No specific titles)*	Detached: Omits the honorific "Bangabandhu."	The Historical Figure: Likely treated as one actor among many, stripped of nationalistic reverence.

Table 5.16: Comparative analysis of the representation of Sheikh Mujib across textbooks.

5.Portrayal of Collaborators

Source	**Terminology Used**	**Scope & Scale**	**Description of Actions**	**Linguistic Tone**
Cambridge 1	"Collaborators"	Minimizing: "A few collaborators"	Passive: "Aided" (implies secondary, helpful role rather than active violence).	Dismissive: Suggests their role was minor and not central to the conflict.
Edexcel 1	"Paramilitary forces"	Structural: Implies an organized unit.	Active/Tactical: "Set up to terrorise the people."	Functional: Describes their strategic purpose (terror) without using local cultural labels.
NCTB 1	*None*	Total Omission	*None*	Silence: The internal enemy is erased from this specific narrative.
NCTB 2	Razakar, Al-Badr, Al-Shams, Associates	Comprehensive: Lists specific groups and implies a broad network.	Visceral/Moral: "Cruel lust," "anti-human offenses," "murdering," "intellectually barren."	Condemnatory: Frames them not just as soldiers, but as moral monsters and enemies of humanity.

Table 5.17: Comparative analysis of the portrayal of collaborators in 1971.

6.Portrayal of Grassroot People

Source	Dominant Narrative	Linguistic Strategy	Sociological Implication
NCTB 2	The Monolithic Hero	Universal Quantifiers: Uses absolutes like "Everyone took part," "Every section," and "Farmers had just one goal."	Total Homogenization: Flattens diverse motivations (economic survival, land rights, political ideology) into a single, romanticized national "will." It creates a myth where the entire nation acted as a single organism.
NCTB 1	The Sole Cause	Mono-Causal Attribution: Claims victory was "only" due to mass cooperation.	Reductionism: Erases the complexity of the war (military strategy, geopolitics, Indian intervention) by attributing the outcome solely to "the people's desire," creating a romantic rather than tactical history.
Cambridge 1	The Vague Majority	Soft Generalization: Uses non-specific quantifiers like "Huge numbers" and "Vast majority."	Depersonalization: Acknowledges scale but lacks depth. The people are present as a statistic ("huge numbers") but absent as distinct agents with specific grievances.
Edexcel 1	The Invisible Mass	Exclusion/Omission: No mention of grassroots people.	Elitism: By omitting the "common people," history is framed exclusively as the domain of elites, leaders, and armies. The "collective" does not exist.

Table 5.18: Comparative analysis of how "The People" (collective agency) are represented.

Colonial Rule

We also examined three different textbooks's of the very same colonial rule history of Bengal. What we had discovered was not a difference of facts only, but three worlds of history altogether. The books possess uniqueness which forms the personality with which the students perceive their past.

1. The Zamindars: Permanent Settlement & Land Revenue Systems

Analysis Angle	Edexcel (Book 1)	Cambridge (Book 2)	NCTB (Book 3)
Role in Economy	Bureaucratic	Communal	Collaborative Exploiter
Attribution of Blame	None (Omitted)	Cultural/Religious	Strategic Guilt

Table 5.19: Comparative analysis of the portrayal of Zamindars.

2. Mir Jafar: The Battle of Plassey (1757)

Analysis Angle	Edexcel (Book 1)	Cambridge (Book 2)	NCTB (Book 3)
Motivation	Personal anger	Pawn	Moral Treachery
Agency	Reactive	Passive	Active (Evil)

Table 5.20: Comparative analysis of the characterization of Mir Jafar.

3. The British: Establishment of Rule & Economic Impact

Analysis Angle	Edexcel (Book 1)	Cambridge (Book 2)	NCTB (Book 3)
Invasion Narrative	Brave Rulers	Inevitable Progression	Conspiracy & Theft
Economic Impact	"The Great Ravage"	"Consequence of War"	"Drain of wealth"

Table 5.21: Comparative analysis of the British Rule narrative.

4. Women: The Bengal Renaissance & Social Reforms

Analysis Angle	Edexcel (Book 1)	Cambridge (Book 2)	NCTB (Book 3)
Representation	Total Omission	Passive Objects	Icons or Victims
Agency	None	Dependent	Symbolic

Table 5.22: Comparative analysis of the representation of women in the 19th Century.

5. The Bengalees: Resistance Movements

Analysis Angle	Edexcel (Book 1)	Cambridge (Book 2)	NCTB (Book 3)
Resistance Framing	Terrorist	Cultural/Reactive	Patriotic/Heroic
Communal Relations	Administrative Issue	Religious division	British created Division

Table 5.23: Comparative analysis of the framing of common people and resistance.

600 b.C.E-Colonial Rule

Potrayals of Mughals

Feature	NCTB 2 (National)	Cambridge (CAM1)	Edexcel (ED1)
Framing of Resistance	Nationalist Resistance: Defines Isa Khan and the Baro Bhuiyans as "Freedom Fighters" protecting sovereignty.	Elite Political Conflict: Does not focus on local motives. Primary focus is transition of elite power (Mauryan or Mughal).	Organized Rebellion: Addressing resistance as "rebellious" hindrance to normalized central authority. No credit to local motives.
View of the Mughals	Foreign Imperialists: Emphasizes the fight against centralizing powers in order to retain regionality.	Administrative Unifiers: Focuses on the superior rule of emperors and top-down rule.	Strategic Victors: Plays the part of the "Mughal Strategist," prioritizing imperial victories and effective expansion.
Sanitization of Conflict	Emotional/Patriotic: Stresses the courage of the "sons of the soil" using hagiographic language.	Devaluation of Social Cost: Sanctions the lack of human dimension and war cost.	Amoral Description: Delivers a monotonous list of events; avoids the cruelty of imperial occupation.
Key Omissions	Ignores the fights amongst local zamindars.	Does not mention the local people's condition during transition of war.	Omission of Mughal's failure in wars.

Table 5.24: Comparative framing of Mughal authority and regional resistance.

Socio-Economic Prosperity (The "Paradise" Narrative)

Feature	NCTB 2 (National)	Cambridge (CAM1)	Edexcel (ED1)
Wealth Framing	Romanticized "Golden Era": Relies on metaphorical exaggeration (e.g., "Paradise of Nations") to construct an idealized past.	Commercial Success: Casts wealth in terms of uncritical and uniformly positive product of trade and administration.	Laudatory Economic Growth: Attribution of positive qualities such as "appreciated importance of trade."
Perspective	Ideological: Uses "Shaista Khan era" as an example to showcase national prosperity.	Elite-Centric: Exclusive emphasis on political/military elites and great trading emporiums.	Colonial/Traveler: Uses European traveler perspectives (e.g., Bernier) to affirm wealth instead of native views.
Erasure of Labor	Substitutional: Substitutes the harsh reality of labor with legends of cheap prices (8 maunds of rice per 1 taka).	Sanitization of Skill: Craftsmanship (Muslin) is degraded into a trade commodity as opposed to human resource.	Narrative Displacement: Obscures the lives of ordinary people with trade infrastructure and market data.
Omission Patterns	Leaves out socio-economic difficulties, religious intolerance, or inequality in the "Golden Age."	Normalizes the lack of the experience of the majority; purely concerned with the accumulation of wealth.	Null Set Protocol: Total neglect of the life of the people and social aspects of society.

Table 5.25: Analysis of economic narratives and the "Paradise" motif.

5.6 Discussions

Here, we discuss the findings of chapter 5.5, limitations and future work.

5.6.1 Interpretation of Findings

The bias detection framework based on the LLM indicates that national and international curricula formulate different historical realities in systematic determinations of narrative, semantics, and syntactic. Instead of direct disagreement of factual nature, the discord in tone, attribution of agency and omission arise.

In international curricula, especially Edexcel, the events of the 1947 Partition are framed in an administrative-colonial perspective that prioritizes efficiency in the process over the misery of the colonized through the impersonal and depersonalized use of both passive and statistical language. Conversely, national textbooks emphasize nationalist versions, now foreshadowing moral crisis and general sacrifice and

selectively abandoning sophisticated or uncomfortable historical events, including the human sacrifice of Partition.

Political leadership is viewed in different ways: NCTB textbooks use emotionally freighted narratives of heroes, which makes leaders be moral authority, international curricula use bureaucratic, functional descriptions which depersonalize leadership. There are parallel opposites in the images of trauma and resistance. The 1971 Liberation War is widely told in national texts which simplify collective action by homogenizing language but in international texts elite actors are prioritized and agency at the grassroots is sidelined.

Curricular bias is also revealed through gendered experiences of violence. Such experiences are largely omitted by international textbooks, but national textbooks do not disregard them, but rather place crimes against women in the same context as material destruction, which shows semantic erasure and gender frames. Describing foreign opposition is also different, with national curricula adopting a direct and adversarial tone, and international texts adopting a passive and diplomatic tone which does not blame the western.

Throughout history, the framework has been used to identify micro-biases, such as implicit colonial views, exonerative syntax, elite-centric erasure, and romanticized or hagiographic framing, which are hard to identify by the use of keywords. These results corroborate the usefulness of the framework in revealing how educational discourses render certain interpretations of history as normal and certain interpretations as obscure.

5.6.2 Limitations and Future Work

Our framework has a few limitations. One of which is sensitivity to variations in queries. Minute changes in how a query is phrased can lead to LLM1 retrieving different answers from the source textbooks. For example, when asked "describe the language used to [event]", LLM1" will occasionally respond with "no language was used to describe [event]". Therefore , even subtle linguistic modification can lead to a difference in the model's interpretation and output. While this drawback can be mitigated by modifying the prompt design, the strategy for the prompt was to ensure that no retrieval is performed when the subject is absent from the query, by intent. Therefore, some queries, such as "What language was used to describe [group]?" have a chance of not rendering any successful retrieval. This behavior is illustrated in Figure 5.7. However, the rate of this issue occurring is relatively low, and the overall semantic interpretation of the retrieved outputs remains largely unaffected.

LLM1 was provided with approximately 15-25 chunks for each query. The task of retrieving data of such a large input context can become computationally expensive. Furthermore, this can also lead to hallucinations by LLM1. Despite the presence of potential drawbacks, it was necessary to supply the LLM with a broader context to meet the objectives of this particular study. Future work could address this limitation by implementing a retrieval mechanism which is both effective and

efficient.

We implemented a single-query approach for this study. To generate a response from LLM1, we used one query which then underwent bias detection performed by LLM2. While this design simplifies the pipeline, it constrains the system's ability to uncover a broader range of potential biases. This method also introduces the possibility of researcher-induced bias through query formulation. A multi-query strategy can address these limitations. For example, LLM1 could initially be prompted with a retrieval-focused query such as "Extract all passages in which women are mentioned." Subsequently, multiple targeted queries could be applied to LLM2. For instance "How are women portrayed?" and "What contributions are attributed to women?" Such a modular querying framework would allow for a more granular and comprehensive bias analysis.

We used our framework to exclusively evaluate historical narrative textbooks from Bangladesh. Therefore, the applicability and robustness of this framework for textbooks from other geographic and curricular scope is currently inconclusive. Additionally, the framework was not tested on alternative data sources such as newspapers or historical texts, where linguistic style, narrative structure, and forms of bias may differ.

This framework is entirely dependent on LLMs, which gives rise to cost related constraints. Furthermore, the use of multiple LLMs adds to the existing financial and computational overhead. Along with that, our study concluded that smaller models do not perform adequately when deployed as LLM2. This also limits the feasibility of cost-efficient model substitutions.

Chapter 6

Conclusion

By using extensive testing and evaluation, our research demonstrates that the proposed framework is able to perform effective bias analysis with educational texts. Our data clearly show that this framework can find a variety of types of bias throughout history and within multiple school textbooks. Another one of the strengths of the framework is that it is modular, so each part — such as retrieval, answer generation or bias detection — may be modified or replaced in order to improve the ability of the framework to generalize and allow for expansion into new applications. Although there are many limitations of the framework — including the need to use large language models and limited scope of the dataset — the framework is successful at finding a number of previously unknown biases in many different types of books from a variety of time periods. This shows the potential for the framework to serve as an efficient and flexible method for identifying multiple biases within large collections of text and provides an area of future study — including larger datasets, better retrieval methods, and less expensive model configurations.

Bibliography

[1] T. G. Rose, N. J. Haddock, and R. C. Tucker, "The effects of corpus size and homogeneity on language model quality," in *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1997.

[2] H. Song, "Deconstruction of cultural dominance in korean efl textbooks," *Intercultural Education*, vol. 24, no. 5, pp. 455–464, 2013. DOI: 10.1080/14675986.2013.809248.

[3] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[4] M. Asadullah, K. M. M. Islam, and Z. Wahhaj, "Gender bias in bangladeshi school textbooks: Not just a matter of politics or growing influence of islamists," *The Review of Faith International Affairs*, vol. 16, pp. 84–89, Apr. 2018. DOI: 10.1080/15570274.2018.1469821.

[5] B. A. Bassey and V. J. Owan, "Ethical issues in educational research and the need for ethical guidelines," *International Journal of Educational Research Review*, vol. 4, no. 3, pp. 388–397, 2019.

[6] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[7] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[8] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, Association for Computing Machinery, 2021, pp. 610–623. DOI: 10.1145/3442188.3445922.

[9] UNESCO, *AI and Education: Guidance for Policymakers*. United Nations Educational, Scientific and Cultural Organization, 2021.

[10] X. Zhai et al., "A review of artificial intelligence (ai) in education from 2010 to 2020," *Complexity*, vol. 2021, no. 1, p. 8 812 542, 2021. DOI: 10.1155/2021/8812542. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1155/2021/8812542.

[11] R. S. Baker and A. Hawn, "Algorithmic bias in education," *International Journal of Artificial Intelligence in Education*, pp. 1–41, 2022. DOI: 10.1007/s40593-021-00285-9.

[12] L. Weidinger et al., *Ethical and social risks of harm from language models*, arXiv preprint arXiv:2112.04359, 2022. [Online]. Available: https://arxiv.org/abs/2112.04359.

[13] E. Ferrara, "Should we be worried about AI bias?" *Communications of the ACM*, vol. 66, no. 8, pp. 34–36, 2023.

[14] A. Mukherjee, C. Raj, Z. Zhu, and A. Anastasopoulos, "Global voices, local biases: Socio-cultural prejudices across languages," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics, 2023, pp. 15 828–15 845. DOI: 10.18653/v1/2023.emnlp-main.981. [Online]. Available: https://aclanthology.org/2023.emnlp-main.981.pdf.

[15] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the real world: A survey on nlp applications," *Information*, vol. 14, no. 4, p. 242, 2023. DOI: https://doi.org/10.3390/info14040242. [Online]. Available: https://www.mdpi.com/2078-2489/14/4/242.

[16] W. X. Zhao et al., "A survey of large language models," *arXiv preprint*, Mar. 2023. DOI: 10.48550/arXiv.2303.18223. [Online]. Available: https://arxiv.org/abs/2303.18223.

[17] M. F. Adilazuarda et al., "Towards measuring and modeling "culture" in llms: A survey," *arXiv preprint*, Mar. 2024. DOI: 10.48550/arXiv:2403.15412. [Online]. Available: https://arxiv.org/abs/2403.15412.

[18] Z. Fan, R. Chen, R. Xu, and Z. Liu, "BiasAlert: A plug-and-play tool for social bias detection in LLMs," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. [Online]. Available: https://arxiv.org/html/2407.10241v1.

[19] I. O. Gallegos et al., "Bias and fairness in large language models: A survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1097–1179, 2024. DOI: 10.1162/coli_a_00524. [Online]. Available: https://aclanthology.org/2024.cl-3.8/.

[20] M. P. Polak et al., "Flexible, model-agnostic method for materials data extraction from text using general purpose language models," *Digital Discovery*, vol. 3, pp. 1221–1235, 2024. DOI: 10.1039/D4DD00016A. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2024/dd/d4dd00016a.

[21] S. Raza, M. Garg, D. J. Reji, S. R. Bashir, and C. Ding, "Nbias: A natural language processing framework for bias identification in text," *Expert Systems with Applications*, vol. 237, p. 121 542, 2024. DOI: https://doi.org/10.1016/j.eswa.2023.121542. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423020444.

[22] AFP, "After revolution, Bangladesh textbooks rewrite history," *DAWN*, Feb. 2025, Published February 14, 2025. [Online]. Available: https://www.dawn.com/news/1891927.

[23] C. Dun et al., "Evaluation of large language model performance in assessing health economic study quality," *Journal of Health Economics and Outcomes Research*, vol. 12, no. 2, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12554303/.

[24] A. Jain, L. Cui, and S. Chen, *Aligning LLMs for the classroom with knowledge-based retrieval: A comparative RAG study*, 2025. arXiv: 2509.07846 [cs.AI]. [Online]. Available: https://arxiv.org/html/2509.07846.

[25] Y. Li, J. H. Mohamud, C. Sun, D. Wu, and B. Boulet, *Leveraging LLMs as meta-judges: A multi-agent framework for evaluating LLM judgments*, 2025. arXiv: 2504.17087 [cs.CL]. [Online]. Available: https://arxiv.org/html/2504.17087v1.

[26] F. M. Megahed, Y.-J. Chen, L. A. Jones-Farmer, G. Lee, J. B. Wang, and I. M. Zwetsloot, *Reliable decision support with LLMs: A framework for evaluating consistency in binary text classification applications*, 2025. arXiv: 2505.14918 [cs.CL]. [Online]. Available: https://arxiv.org/html/2505.14918v1.

[27] S. Qi, R. Cao, Y. He, and Z. Yuan, "Evaluating LLMs' assessment of mixed-context hallucination through the lens of summarization," *arXiv preprint arXiv:2503.01670*, 2025.

[28] C. Stryker. "What is responsible AI?" Accessed: 2026-01-28. [Online]. Available: https://www.ibm.com/think/topics/responsible-ai.

[29] Cambridge Assessment International Education, *Cambridge O Level Social Studies / Bangladesh Studies – Book 2*. Cambridge University Press or Hodder Education, [2025]. [Online]. Available: https://www.cambridgeinternational.org/programmes-and-qualifications/cambridge-o-level-bangladesh-studies-7094/.

[30] National Curriculum and Textbook Board (NCTB), *Bangladesh and Global Studies (Class Nine-Ten)*. National Curriculum and Textbook Board (NCTB), [2025]. [Online]. Available: https://nctb.portal.gov.bd/site/page/3a96de78-a64d-49e0-8a24-df9a1d305d87.

[31] Pearson Edexcel, *Edexcel London Examinations GCE Ordinary Level Bangladesh Studies(7038) Book 12 Second Edition*. Pearson Education Limited, [2017]. [Online]. Available: https://boibichitra.com/product/Edexcel-London-Examinations-GCE-Ordinary-Level-Bangladesh-Studies7038-Book-12-Second-Edition-M4Z7R.

Appendix A

Base Query Template for Textual Analysis

To ensure analytical consistency and to minimize thematic ambiguity and researcher-induced bias, this study employed a standardized base query template. All analytical questions posed to the model were derived from the structures outlined below.
The query framework is organized into three functional categories: **Framing**, **Content**, and **Representation**.

Framing Queries

Framing queries examine how subjects, events, and actors are linguistically introduced and characterized.

- What [languages / terms / titles / adjectives] are used to [describe / introduce] [Subject] [in the context of Event]?
- How is/are [Subject / Event] [described / portrayed / characterized / treated]?
- In the context of [Event], is [Subject / Event] described as involving [Active Action] or [Passive Action]?
- What [reasons / causes] are [cited / attributed] [for / to] [Event]?

Content Queries

Content queries focus on factual narration, actions, temporal placement, and measurable details.

- What roles did [Actor / Group] play in the context of [Event]?
- How did [Actor / Group] [action verb] [context]?
- What [events / incidents] [occurred / took place / happened] [during / on] [Event / Date]?
- What were the consequences of [Event / Movement]?
- When did [Event] [happen]?

Representation Queries

Representation queries assess attribution, agency, responsibility, and terminological choices.

- [Who / Which people / Which groups] [is / are] [credited / identified / responsible / attributed] [with / for] [Action / Event]?
- What [reasons / factors / causes / explanations] [does the text provide / are cited] [for / to] [Event]?
- Does the text [explicitly] use the [term / word] "[Specific Term]" to describe [Subject]?
- How [is / are / does] [Subject / Intention] [described / portrayed / characterized]?

Appendix B

System Prompt for LLM-1

System Message

You are a highly precise linguistic extraction assistant. Your sole purpose is to find and present all information from the 'Relevant Context' that is relevant to the user's query: {user_query}.

STRICT POLICY: You must operate under a strict 'zero-synthesis' and 'zero-interpretation' policy. Do not add, infer, or summarize.

COMPREHENSIVE INCLUSION MANDATE (Slight Relevance Rule):
You MUST extract and present every piece of information that is relevant to the user's query. You MUST extract every single instance where the query subjects are mentioned, even if the mention is 'slight' or 'partial'.

- If the query subjects appear only as a single word in a sentence, that sentence MUST be included.
- If the query subjects are part of a larger list, that list MUST be included.

Formatting Requirements:

1. **PROVENANCE:** Every extracted point must include the metadata (Source, Page, Chunk ID).
2. **VERBATIM EXTRACTION:** You must quote the text exactly as it appears in the source text titled {book_name_in_context}.
3. **GROUPING RULE (MANDATORY):** Do NOT create separate numbered entries for sentences that belong to the same paragraph. You MUST group them into a single 'Logical Unit' to avoid redundancy.
4. **HANDLING OMISSIONS (The Null Set Protocol):**
   If the specific query is NOT mentioned in the context of the requested timeline or theme, you MUST provide 'Negative Evidence' using the following structure:
   - **Status:** State clearly: 'The specific query regarding [Subject] is not mentioned in this section/timeline.'
   - **Substitution Identification:** Identify the 'Dominant Actor' or 'Primary Subject' that occupies the discourse space for that specific period or event instead.
   - **Primary Facts:** Provide a bulleted list of verbatim sentences describing what the book discusses *instead*.
   - **Logical Unit:** Provide the full, verbatim passage containing these substituted facts, expanded for C=1.
   - **Strict Condition:** Only if the entire provided text is completely unrelated to

the historical period/event, respond with: 'The provided text does not contain sufficient information regarding this query.'

—
Relevant Context:
{context_text}
—

Appendix C

Table-Based Scoring Rubric for Framing & Normalization Bias

Table C.1: Bias Evaluation Criteria and Scoring Guidelines

Component	Definition	Operational Indicators	Scoring Guidance (0–10)
Framing & Normalization Bias	Linguistic and narrative choices that distort historical understanding.	Language choice, tone, narrative emphasis, and omission within scope.	Score reflects the **net effect** of bias indicators.
Criterion A: Normalization Bias	Presentation of harmful practices in ways that reduce perceived severity.	• Passive/bureaucratic descriptions • Omission of human suffering • Erasure of resistance	0 = No normalization 5 = Mixed signals 10 = Severe normalization
Criterion B: Framing Bias	Framing that steers readers toward a judgment beyond evidence.	• Loaded language • Selective emphasis • Euphemisms	0 = Neutral 5 = Limited framing 10 = Strong framing

Table C.2: Overall Score Calculation

Step	**Description**
Step 1	Assign separate scores (0–10) for Normalization and Framing.
Step 2	Compute the simple average of the two scores.
Step 3	Adjust qualitatively based on mitigating contributions.
Step 4	Round final score to one decimal place.

Appendix D

Anti-Normalization Constitutional Protocol

Purpose

Anti-normalization constitutional protocol is an important element of the framing bias evaluation process in LLM2. The purpose of this protocol is to address a particular type of problem in narrative development for conflict history: distinguishing between distortion (bias) in narrative development and morally correct judgments made based on historical evidence.

Constitutional Principle

The model is explicitly prohibited from punishing or reducing the severity of emotionally negative historical sentiment when the subject matter being described includes genocide, systemic persecution, sexual violence, and/or civilian targeting. Normalized, or sanitized, language used to describe emotionally neutral historical events can be considered framing bias.

Criterion for Evaluation

Instead of determining if the language is "neutral," the framing-bias evaluator determines whether the moral weight of the description is proportionally representative of the documented severity of the events. This is especially important when developing conflict histories because accurately portraying atrocities requires serious, severe descriptions.

Appendix E

Bias Evaluation Worksheet

Section C: Bias Evaluation Tables

Table 1: Omission Bias

Src	Analysis	Score (1–5)	Justification
A	Summary text	☐1 ☐2 ☐3 ☐4 ☐5	
B	Summary text	☐1 ☐2 ☐3 ☐4 ☐5	

Table 2: Factual Distortion

Src	Analysis	Score (1–5)	Justification
A	Summary text	☐1 ☐2 ☐3 ☐4 ☐5	
B	Summary text	☐1 ☐2 ☐3 ☐4 ☐5	

Appendix F

Bias Identification Worksheet

Purpose

You are asked to review a textbook-style question and answer to determine if specific forms of bias are present, based on historical accuracy and presentation.

Section A: Review

Question:

Answer:

Section B: Bias Identification

1: Framing Bias
The use of emotionally charged or manipulative language to shape reader perception.

☐ Present ☐ Not present ☐ Unsure

Justification:

2: Normalization Bias
The use of sterile or bureaucratic language to minimize the gravity of historical atrocities.

☐ Present ☐ Not present ☐ Unsure

Justification:

Appendix G

Validation Worksheet

Purpose

This worksheet asks you to evaluate whether the model's bias analysis and final classification for a given question–answer pair are methodologically and substantively valid.

Section A: Context Information

Field	Details
Question ID	
Source Book	
Bias Type Evaluated	

Section B: Review

Bias Analysis Provided:

Model Classification:

☐ TP ☐ TN ☐ FP ☐ FN

Section C: Expert Validation

Is the model's classification valid?

☐ **Valid** ☐ **Invalid**

Reasoning:

Appendix H

Factual Distortion Key Phrases Across Curricula

Book Source	**Category**	**Identified Inaccuracy or Claim**
Cambridge Book 1	Geography & Locations	• Incorrectly claims multiple sites are in **Dhaka**, including: Mosque of Shah Niamatullah Wali. • Misidentifies Raja Ram Temple location.
NCTB Book 1	1971 Liberation War	• **Operation Searchlight:** Blueprint incorrectly stated as "designed on 25 March". • **Declaration:** Contextual omission regarding the timing of the declaration.