Thesis Report

Transformer-Based Automated Question Prediction and Generation Aligned with Bloom's Taxonomy for Educational Assessments

SUBMITTED BY

Robiul Hasan Arman

Roll No: ASH2011049M

Session: 2019-2020

Year 04, Term 02

Department of ICE, NSTU

THESIS SUPERVISOR

Dr. Md. Ashikur Rahman Khan

Professor

Department of ICE, NSTU

Year 04, Term 02

Department of ICE, NSTU

DEPARTMENT OF INFORMATION AND COMMUNICATION ENGINEERING

NOAKHLAI SCIENCE AND TECHNOLOGY UNIVERSITY

July 2025

A thesis Report on Transformer-Based Automated Question Prediction and Generation Aligned with Bloom's Taxonomy for Educational Assessments

Robiul Hasan Arman

Roll No: ASH2011049M

Department of Information and Communication

Engineering

Noakhali Science and Technology University

DECLARATION

This thesis is submitted to the Department of Information and Communication Engineering of Noakhali Science & Technology University in partial fulfillment of the requirements for the award of the degree of Bachelor of Science (B.Sc.) Engineering in Information and Communication Engineering (ICE). So, I hereby declare that this report has not been submitted elsewhere for the requirement of any kind of degree, diploma, or publication.

Robiul Hasan Arman
Student of Information and Communication Engineering, NSTU
Roll No: ASH2011049M
Session: 2019-2020

ACCEPTANCE

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science (B.Sc.) degree in Information & Communication Engineering.

Dr. Md. Ashikur Rahman Khan
Professor
Department of Information and Communication Engineering
Noakhali Science & Technology University
Sonapur, Noakhali.

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my advisor, Professor Dr. Md Ashikur Rahman Khan, for his immense support and guidance throughout my bachelor's degree at Noakhali Science & Technology University. Without his valuable assistance, completing this thesis would have been extremely difficult. I sincerely appreciate all of his guidance. His pragmatism and insightful advice have been a great help to me during my academic journey, and I am confident that the lessons learned will be beneficial for my future endeavors.

I would also like to extend my heartfelt thanks to all my classmates and the members of the Information and Communication Engineering Department at NSTU, whose support has been invaluable in many ways. I am grateful for their constant encouragement and assistance.

Additionally, I would like to thank the authorities of the online question banks and educational websites that allowed me to collect the data for this research. Their cooperation and access to the resources played a crucial role in the successful completion of this work.

Last but not least, I owe my deepest gratitude to my parents for their unwavering encouragement and support throughout my life. Words cannot fully express the extent of my appreciation for everything you have done for me.

ABSTRACT

Bloom's Taxonomy, a widely recognized framework for predicting cognitive skills, guides the design and evaluation of educational assessments. Traditional methods often struggle to maintain an effective balance between Lower-Order Thinking Skills (LOTS) and Higher-Order Thinking Skills (HOTS), while manually creating exam questions remains time-consuming and inefficient. To address these challenges, this study presents an automated system for predicting and generating exam questions aligned with Bloom's Taxonomy, ensuring comprehensive representation across all cognitive levels. The system leverages transformer-based models, including BERT, DistilBERT, RoBERTa, T5, BART, DistillBART, XLNet, and BERT2BERT, capitalizing on their deep contextual and semantic understanding capabilities. Furthermore, the research incorporates a hybrid prediction model that combines RoBERTa with Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Term Frequency-Inverse Document Frequency (TF-IDF) features to enhance question prediction accuracy and robustness significantly. The system was trained and evaluated on a diverse, manually annotated dataset spanning multiple academic disciplines, ensuring broad applicability. Experimental results demonstrate that the hybrid model achieved an impressive question prediction accuracy of 96.39% in predicting Bloom's cognitive levels, effectively handling both LOTS and HOTS complexities. Additionally, the T5-based question generation model attained a high accuracy of 95.36% alongside a BLEU score of 0.8628, illustrating its ability to produce contextually relevant and pedagogically diverse questions. This research contributes to the advancement of educational technology by offering a scalable, efficient, and intelligent solution for automated assessment design.

Keywords: Bloom's Taxonomy, Automated question generation, Transformer models, BERT, RoBERTa, T5, BART, BLEU score, Educational Assessments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Introduction

As education systems worldwide continue to modernize, the demand for innovative assessment methods has grown significantly. Practical assessments are no longer limited to testing surface-level knowledge; instead, they must evaluate a range of cognitive abilities while encouraging deeper learning. Traditional exam question creation processes are time-consuming, labor-intensive, and often fall short in terms of cognitive inclusiveness. In particular, assessments tend to be imbalanced—overemphasizing either Lower-Order Thinking Skills (LOTS) such as Remembering, Understanding, and Applying, or neglecting Higher-Order Thinking Skills (HOTS) like Analyzing, Evaluating, and Creating, as classified in Bloom's Taxonomy [1].

Furthermore, the conventional method of designing assessments has become even less adequate as educational demands have grown more complex and personalized learning experiences have become essential. Not only must their questions align with the goals outlined in the curriculum, but they also need to be adaptable to meet the diverse learning needs of students. However, it seems humanly impossible to generate such questions for every course, as this is not expected from educators themselves. Thus, a question generation and categorization system that would be a scalable solution is critical. It should also deliver its productivity without compromising educational quality.

All educational institutions must grapple with the challenge of creating practical assessments designed to evaluate a wide range of cognitive skills. The process of manually formulating exam questions is not only slow but also results in question papers that are unequal in terms of cognitive load, often featuring an overemphasis on either LOTS or HOTS questions .

Additionally, larger classes with a multimodal learning approach add even more strain to instructors to keep assessment practices effective and unbiased. The incorporation of technology in education has certainly enhanced content presentation; however, assessment design is still slow to catch up to these levels of automation and scalability. Conventional question banks tend to be

statically rigid, which is detrimental to the overall fostering of the learners' critical thinking and problem-solving skills.

There are numerous opportunities to automate and enhance the assessment process as Natural Language Processing (NLP) and Artificial Intelligence (AI) continue to develop. AI-driven technologies can generate a vast array of exam questions that are both educationally sound and contextually relevant, accommodating varying degrees of cognitive complexity by leveraging cutting-edge natural language processing models. [2]. Additionally, these tools can automatically categorize generated or existing questions into the appropriate categories of Bloom's Taxonomy, ensuring that assessments are well-balanced and accurately reflect the intended learning outcomes. This combined ability to create and categorize questions ensures thorough and insightful evaluations while also saving time. Through the use of sophisticated Transformer-based language models, this study aims to automate the process of creating and classifying test questions completely, thereby transforming conventional evaluation techniques.

The system proposed in the research will therefore automate the creation and prediction of exam questions. Likewise, it will generate assessments that are aligned with specific educational goals and learning standards. The system's design will complement existing Learning Management Systems (LMS) by broadening the scope of dynamic and scalable assessment delivery. This will enable educators to create, customize, and deliver balanced assessments quickly, while reducing the workload and enhancing the quality of education.

An automated assessment design process will thus capture a significant amount of information, bringing reforms to education assessment, addressing existing challenges, and, consequently, providing scalable solutions to enhance critical thinking and learning comprehensively. Ultimately, this will provide educators with a strong educational tool to create balanced, high-quality assessments that improve learning outcomes and engage students in higher-order thinking.

1.2 Motivation

Exam questions created and categorized using traditional methods require a significant amount of time and often fail to meet the diverse demands of contemporary education. It can be challenging to balance LOTS and HOTS when aligning assessments with Bloom's Taxonomy. To provide

scalable, contextually relevant, and cognitively balanced assessments, this research utilizes a Transformers-based NLP Model to automate question generation and prediction. Reducing the workload of educators is expected to enhance the quality of learning and foster critical thinking.

1.3 Problem Statement

Although significant progress has been made in the field of educational technologies, the challenges of manually generating and categorizing questions in examinations persist. Teachers often struggle to create balanced assessments that test all levels of Bloom's Taxonomy. As a result, most teachers either overlook Higher-Order Thinking Skills (HOTS) or overemphasize Lower-Order Thinking Skills (LOTS).

To my knowledge, there exists no automated system that can generate exam questions that consider both the context and the pedagogical validity at all levels of Bloom's Taxonomy. Moreover, prediction of questions by the remains of Bloom's cognitive domain is a significant obstacle.

The goal of this research is to develop a system that automatically generates and classifies questions in an exam using complex Natural Language Processing (NLP) systems. The system has been structured to operate according to Bloom's Taxonomy, utilizing transformer-based machine learning models that enable not only significant cognitive heterogeneity but also a high degree of educational applicability in the assessment programs.

1.4 Objectives of the Research

The primary objective of this research is to design and implement an automated system that generates and predicts test questions based on Bloom's Taxonomy. The precise objectives are as follows:

1. To develop an NLP-based predictor that predict questions into the appropriate Bloom cognitive levels.

2. To design a system of automatically creating exam questions that are educationally appropriate, grammatically correct, and contextually relevant in alignment with all levels of Bloom's Taxonomy.

3. To propose a model that can predict contextually coherent and relevant words at all stages of cognitive development by Bloom's Taxonomy.

4. To construct question sets, keep the proportion of Higher-Order Thinking Skills (HOTS) and Lower-Order Thinking Skills (LOTS) at the preferred ratio.

5. To contrast and Analyse the capabilities of different transformer-based models to generate and predict questions using models including BERT, DistilBERT, RoBERTa, T5, BART, DistillBART, XLNet, and BERT2BERT.

1.5. Scope of this Research

In this research, the goal is to design and implement an automated question generator and a question organizer that incorporates the mechanisms of oral and written examinations, built on the principles of Bloom's taxonomy. The system will utilize transformer-based NLP models to ensure that the generated questions are contextually accurate, grammatically correct, and free from typographical and semantic errors.

The scope of this study involves designing a question Prediction module that can accurately predict the cognitive level of questions, distinguishing between lower-order thinking skills (LOTS) and higher-order thinking skills (HOTS) to provide a balanced and unbiased evaluation. Moreover, the system is designed to easily integrate with currently established Learning Management Systems (LMS), allowing for the easy deployment and adoption of the system within the institutional framework.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Automating educational assessment with the help of AI and NLP has been somewhat successful; it focusing on generating questions and categorizing them according to Bloom's Taxonomy. This review Chapter investigates some of the works that have already been carried out in Question Prediction and Generation.

2.2 Bloom's Taxonomy

Bloom's Taxonomy helps classify educational learning objectives based on their specificity and complexity. It divides learning into three domains—cognitive, affective, and psychomotor. The cognitive domain focuses on levels of thinking skills and essential learning behaviors, supporting the design and evaluation of educational outcomes.

This Taxonomy is composed of six phases, including remembering, understanding, applying, analyzing, evaluating, and creating.

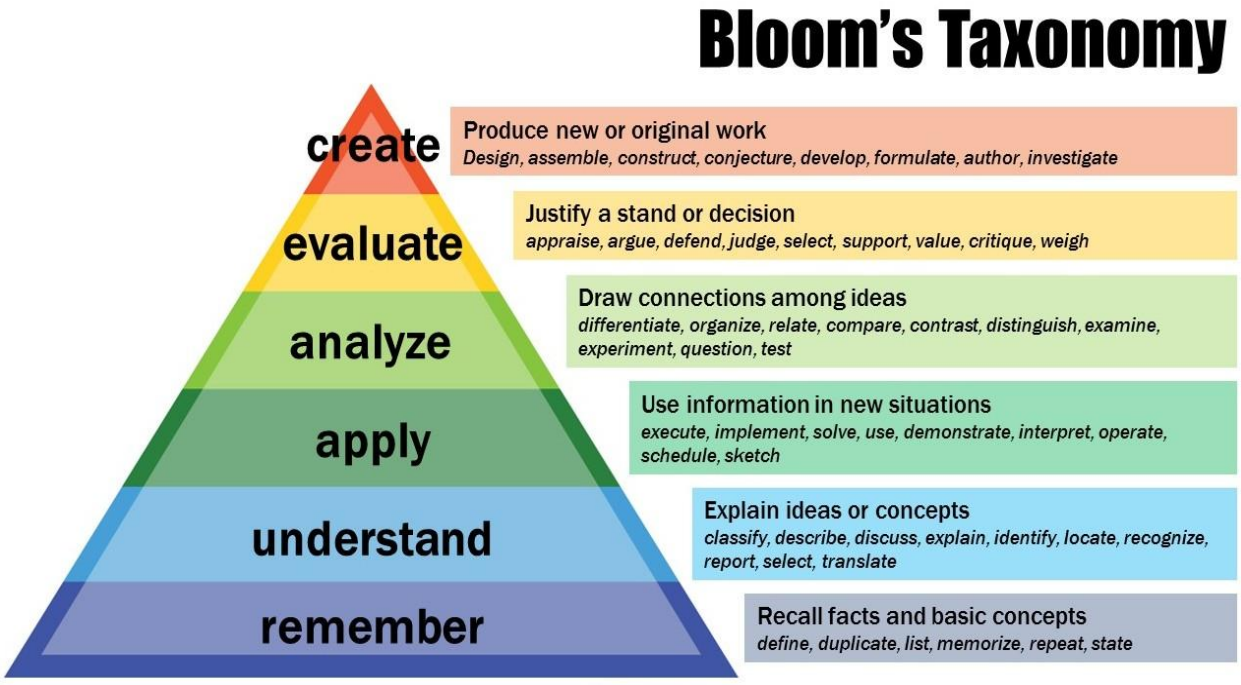

Figure 2. 1: Bloom's Taxonomy [3]

2.3 Question Prediction based on Bloom's Taxonomy

Many Studies conducted in the past have explored various methods for Question Prediction tasks, including rule-based systems, semantic analysis, and machine learning (ML) techniques. In this section we will discuss Question Prediction work that have been done under the Bloom Taxonomy.

To enhance the prediction of questions, Goh et al. applied semantic analysis and dependency parsing [4]. The method relied on syntactic parsing techniques and WordNet to comprehend semantic relationships rather than traditional machine learning training. Even though this method achieved an accuracy level of 88% its scalability was limited by the complex and resource-intensive preprocessing steps.

Kumara and Paik introduced a rule-based framework aimed at analyzing exam papers through the lens of Bloom's Taxonomy [5]. Instead of traditional training methods, the system employed predefined rules for classifying exam questions. The data was gathered from historical exam papers. The system efficiently assessed exam quality for Lower Order Thinking Skills (LOTS) but exhibited limited adaptability for Higher Order Thinking Skills (HOTS). The testing process relied on manual evaluation, resulting in an accuracy of 85%. Idris and his colleagues conducted a descriptive analysis in 2021 using surveys aimed at biology teachers to find out how well they understood Bloom's Taxonomy in assessment design [6]. This investigation did not involve the creation of any machine learning models. Data was gathered using structured surveys, which showed a significant lack of application of HOTS (44, 17 percent) but a strong emphasis on LOTS (92, 22 percent). The study received a rating of 78. 06 percent for its overall efficacy in detecting cognitive balance in assessments.

In 2015, Dhuha Abdullah Abdul-Jabbar's paper Exam Questions Prediction Based on Bloom's Taxonomy Cognitive Level Using Classifiers Combination employed an ensemble learning approach that combined k-NN, SVM, and Random Forest classifiers [7]. And an accuracy of 80.82%.A dataset of test questions from different academic disciplines was subjected to this methodology. K-fold cross-validation was used to evaluate performance, and mutual information was used to select features. The study demonstrated the effectiveness of employing classifier ensembles to increase accuracy.

Ibtihal R. Assalys' 2015 study, Using Bloom's Taxonomy to Evaluate the Cognitive Levels of Master's Students [8], looks at how exam questions are categorized based on Bloom's cognitive levels. The questions collected from master's programs were manually annotated by the researchers, who then evaluated those using statistical techniques. Their results demonstrated an unbalanced distribution of questions across cognitive levels, giving lower taxonomy levels—such as Remember and Understand—a disproportionate amount of weight.

A case study based on Bloom's Taxonomy Classifications of Exam Questions Using Natural Language Processing was carried out by Addin Usman [9]. Exam question classification using machine learning and natural language processing techniques in accordance with Bloom's Taxonomy levels was investigated in 2016. In addition to using TF-IDF and Word2Vec for feature extraction, the researchers trained on annotated question datasets and utilized algorithms like Naive Bayes and Decision Trees. In identifying higher-order cognitive levels, particularly Analyze and Evaluate, their model evaluation showed improved precision and recall. Using Bloom's Taxonomy as a guide. Shaikh et al. implemented an LSTM-based classification model, achieving 87% accuracy for Course Learning Outcomes (CLOs) and 74% for assessment items [10]. Another researcher use Ensemble approach and gain accuracy 82% [11].N. Yusof and C. J. Hui use SVM to predict question and achieve accuracy 76%. P. van Beukelen also use SVM and gain accuracy 80% [12]. Lalchhuanmawii and Zoramchhuana manually analyzed high school art exam papers. To evaluate the distribution of cognitive levels, the authors manually categorized exam questions. Their analysis revealed a dearth of HOTS questions and a notable prevalence of LOTS. This study did not make use of any computational models or automated systems. Due to its qualitative nature, it did not provide any metrics for accuracy [13].

Laddha et al. utilized Long Short-Term Memory (LSTM) models and Convolutional Neural Networks (CNN) to categorize test questions in accordance with Bloom's Taxonomy [14]. The models were trained using sizable datasets of categorized test questions that were labeled. Testing entailed validating model performance using unseen datasets. The model increased classification accuracy, but it lacked the ability to generate questions. Although the accuracy was not stated, it was noticeably better than that of conventional models. Annotated educational datasets that were in line with Bloom's Taxonomy were used by Chanaa and colleagues to develop a machine learning classification model in 2024. Although it did not address the creation of new questions, the model

achieved a classification accuracy of 90% in identifying cognitive levels when evaluated using common metrics such as precision, recall, and F1-score [15]. Yaacoub et al. assessed the alignment of AI-generated questions with Bloom's Taxonomy using classifiers such as Naive Bayes, Logistic Regression, SVC, and DistilBERT. The DistilBERT model achieved the highest accuracy of 91% [16].

S.Dhainje et al. (2023) Applied NLP and Naïve Bayes to classify programming exam questions according to Bloom's cognitive levels. The system achieved 85–90% accuracy (inferred from test descriptions, though exact value not explicitly stated). Effective for LOTS tasks, but struggled with abstract HOTS due to syntactic ambiguity, context variability, and a limited ruleset [17]. K.Jayakodi et al. WordNet and Cosine based similarity and gain 70% accuracy [18].

Chavan et al. created a question paper generation system using ML (Naïve Bayes) to classify user-uploaded questions by Bloom level and complexity [19]. Reported classification accuracy of 86% on the dataset. The system supports randomization and avoids repetition. Limitations include reliance on CSV-format input, difficulty with semantic nuances, and the lack of deep contextual validation.

Chandio et al. (2020) developed a Naive Bayes-based classification model aimed at categorizing questions into Bloom's Taxonomy levels, primarily focused on the cognitive domain [20]. The classifier was trained on a dataset of educational questions collected from computer science domains. Evaluation metrics such as accuracy, precision, recall, and F1-score were used, and the model achieved an accuracy of 81%**.**.Gani et al. proposed an extensive benchmarking framework to evaluate assessment question classification across Bloom's Taxonomy using machine learning (ML)**,** deep learning (DL**)**, and generative AI (GenAI) approaches [21]. The study introduced a domain-specific term weighting scheme for ML models, which achieved the highest performance with 87.1% accuracy and 87.2% F1-score, outperforming both DL and GenAI counterparts. While DL models like CNN and hybrid RNN-CNN showed moderate success, GenAI models (e.g., GPT-3.5-turbo, PaLM2) performed significantly worse in zero-shot settings (accuracy up to 61.8

Herrmann-Werner et al evaluated GPT-4's performance on Bloom-classified medical MCQs. It achieved 91–93% accuracy, excelling at LOTS but showing errors in recall and reasoning, especially with HOTS. The study revealed GPT-4's limitations in semantic understanding, hallucination issues, and contextual reasoning [22].Stevani & Tarigan (2022) manually analyzed English textbooks for Bloom's levels. Most questions targeted Understand (26%) and Remember (17%), with minimal HOTS. It highlighted the lack of high-level cognitive tasks and the unsalable nature of manual analysis [23].Tika Reformatika Fuadi used Descriptive method for Analyzing Exam questions and concentrated on lower levels (C1-C4); C5 and C6 were absent [24]. And accuracy was Not Mentioned. Shadi Diab and Badie Sartawi. Use Semantic similarity; NLP techniques. And find cognitive levels based on action verbs Precision: 94% F1:89% [25].

2.4 Question Generation based on Bloom's Taxonomy

Several studies have explored question generation using Bloom's Taxonomy to ensure cognitive diversity and alignment with learning objectives. These approaches range from rule-based systems to advanced machine learning models that generate questions across various cognitive levels. Here are some of the works that have been done in Question Generation using Bloom's Taxonomy

In order to create exam questions that are in line with Bloom's Taxonomy, Timakova and Bakon created an automated system based on rules in 2018 [1]. The system used rules that were manually created to generate questions that focused on Lower-Order Thinking Skills (LOTS) like comprehension and memory. Manually selected test questions made up the training data, but the system was not flexible enough to accommodate Higher-Order Thinking Skills (HOTS). Manual validation and expert reviews were the main methods used for testing. Even though the model's accuracy was 89% its efficacy in dynamic learning environments was limited because it was unable to scale and adjust to increasingly complex assessments. Contreras and his colleagues [2] created a hybrid method in 2022 that combines Support Vector Machines (SVM) and Natural Language Processing (NLP) to generate essay questions. A collection of essay questions arranged in accordance with Bloom's cognitive levels was used to train the model. The model's performance was evaluated through the use of cross-validation and other common machine learning validation techniques. The

system achieved an accuracy rate of 85. 7% demonstrating significant progress in generating pertinent essay questions. However, it struggled to create Higher Order Thinking Skills (HOTS) questions and produced content that lacked diversity.

A hybrid natural language processing model was developed by Sahu and his colleagues in 2021 to generate comprehension-based questions that are consistent with Bloom's Taxonomy [26]. Using both rule-based and statistical NLP techniques, this model was trained on academic texts and comprehension datasets. The assessment procedure comprised contrasting the quality and relevancy of the generated questions with those created by humans. Although the model improved comprehension tests with a 91 percent accuracy rate, it was not generalizable and had trouble scaling across different subjects.

S. Kadiyala Created a rule-based formula for multiple-choice questions (MCQs) based on Bloom's Taxonomy [27]. Experts in education examined this system to guarantee its quality. Medical textbooks and course materials provided the information. With a 92% accuracy rate, it effectively raised the caliber of multiple-choice questions (MCQs) but showed a lack of diversity in the generated questions. Moholkar, K., Patil, used an Ensemble Deep Neural Network (LSTM) and achieved 77.4% accuracy [28].

Tanuja et al. (2023**)** introduced Question Guru**,** a rule-based automated system designed to generate multiple-choice questions (MCQs) from input text. The system employs template-based generation and POS-tagging, extracting subject-verb-object patterns to identify key answerable concepts. Distractors are created using lexical databases such as WordNet and syntactic structure rules. While primarily targeted at lower-order Bloom's levels (Remember, Understand) The paper does not report empirical metrics like accuracy or BLEU scores but includes qualitative evaluation examples and suggests future integration [29].

Wijanarko et al. (2020) proposed a key phrase-based question generation system using context-free grammar aligned with Bloom's Taxonomy. The system produced over 92,000 essay questions with a BLEU score of 0.821 but showed limitations in handling numerical or schematic content and adaptability beyond textual domains [30]. V. Sudhir (2018) proposed an NLP-based system to automatically generate educational questions based on Bloom's Taxonomy using templates and

syntactic parsing. The system parses sentences using Stanford NLP tools, extracts key concepts, and applies Bloom-aligned question templates to generate MCQs. Although it targets mainly lower-order levels [31].

2.5 Summary

This section provides an overview of existing research efforts that have been undertaken using a variety of approaches. These works reflect the diversity of methodologies applied to tackle the problem from multiple angles, offering valuable insights and laying the groundwork for further study.

Table 2.1 Summary of the previous work

SI No.	Title	Author(s)	Methods	Observations	Accuracy	Limitations
1	Bloom's Taxonomy Based Exam Question Generation System.	Timakova et al. [1]	Rule-based	Effective in LOTS but lacked adaptability for HOTS.	89%	Unable to generate HOTS; manual validation.
2	Essay Question Generator Based on Bloom's Taxonomy	Jennifer O. Contreras el al. [2]	SVM	SVM improving question relevance.	85.7%	Weak in HOTS; low content diversity.
3	Question Classification Using Universal Dependency and WordNet	Goh Thing et al. [4]	Dependency parsing, semantic analysis	Enhanced classification accuracy with complex preprocessing.	88%	High preprocessing complexity; scalability concerns

4	Rule-Based Question Analysis for Exam Papers	Kumara & Paik et al. [5]	Rule-based analysis	Effective in measuring exam quality but lacked adaptability.	85%	Not adaptive to diverse or evolving exam styles.
5	Profile of Prospective Biology Teachers Using Bloom's Taxonomy	Idris, Ferazona & Safitri et al. [6]	Descriptive analysis with surveys	Strong in LOTS (92.22%), weak in HOTS (44.17%).	78.06%	Insufficient development of HOTS among prospective teachers.
6	Exam Questions Classification Using Classifier Combination	Dhuha Abdullah Abduljabbar et al. [7]	Naïve Bayes, SVM, k-NN	Enhanced accuracy with ensemble models	80.82%	Moderate accuracy; lacks generative capabilities.
7	Classification for Revised Bloom's Taxonomy Using Deep Learning	Manjushree D. Laddha, Varsha et al. [14]	CNN, LSTM	Improved classification but no question generation.	Not Mentioned	Focused only on classification.
8	Cognitive Level Evaluation with ML and Bloom's Taxonomy	Abdessamad Chanaa, Noureddine et al. [15]	SVM, KNN	Achieved precise classification of cognitive levels.	90%	Did not evaluate generative or adaptive learning integration

9	Assessing AI-Generated Questions Alignment with Cognitive Frameworks in Educational Assessment	A.Yaacoub, J. Da-Rugna et al. [16]	Logistic Regression, SVC, and DistilBERT	Transformer models could effectively differentiate cognitive levels	91%	Focused on alignment, not question creation; lacks depth in HOTS evaluation.
10	Analysis of Final Semester Assessment Questions in Grade XI Chemistry	Tika Reformatika et al. [24]	Descriptive method	concentrated on lower levels (C1-C4); C5 and C6 were absent	Not Mentioned	No automation; HOTS (C5, C6) questions missing.
11	Classification of Questions and Learning Outcome Statements into Bloom's Taxonomy	Shadi Diab et al. [25]	Semantic similarity; NLP techniques	Precise classification of cognitive levels based on action verbs	Precision:94% F1:89%	Focused only on classification; lacks generative functionality
12	Comprehension-Based QA Using Bloom's Taxonomy	Pritish Sahu et al. [26]	Rule Based	Improved comprehension assessment.	91%	Not suitable for large-scale applications
13	Applying Bloom's Taxonomy in Framing MCQs	Silpa Kadiyala et al. [27]	Rule-based	Enhanced question quality in medical education.	92%	Domain-specific (medical)

14	Automated Question Generation Using Keyphrase-Based Context-Free Grammar for Bloom's Taxonomy	W. Wijanarko et al. [30]	Key phrase-Based Context-Free Grammar	Generated grammatically valid questions aligned with Bloom's levels	BLEU score 0.821	Limited to grammar-based generation; lacks semantic depth
15	Taxonomy-Based Educational Question Generation Using NLP	V. Sudhir et al. [31]	Templates and syntactic parsing	Structured generation of taxonomy-aligned questions	Not Mentioned	Template-driven model restricts flexibility question diversity.

2.6 Research Gap

Recent advancements in Natural Language Processing (NLP) and Artificial Intelligence (AI) have significantly enhanced the capabilities of automated question generation and prediction. While traditional systems perform adequately for generating questions targeting Lower-Order Thinking Skills (LOTS), they often lack the scalability and adaptability required to address Higher-Order Thinking Skills (HOTS). More recent models utilizing Transformer-based architectures, such as DistilBERT, have demonstrated potential in distinguishing cognitive levels aligned with Bloom's Taxonomy, yet remain limited in their ability to produce a broad range of high-quality questions. To overcome these limitations, this study proposes a novel Transformer-based NLP framework capable of both generating and predicting questions across all cognitive levels of Bloom's Taxonomy. This approach ensures scalable, context-aware, and pedagogically aligned question generation, thereby advancing the automation of educational assessment systems.

CHAPTER 3

METHODOLOGY

3.1 Introduction

Creating an automated system that can quickly classify and produce questions in accordance with Bloom's taxonomy is the main objective of the thesis. Several machine learning techniques and transformer-based models, such as BERT AND T5, and hybrid approaches can help to achieve this goal. This chapter provides a comprehensive overview of the automated techniques employed, including the model execution process, Data collection, and model development. The goal of each stage is to guarantee that the questions generated and predicted are not only sound from an educational perspective but also suitable for the context and consistent with cognitive learning goals.

3.2 Workflow Diagram

The suggested Automated Question Generation and Prediction System workflow is designed to systematically generate and classify learning questions using Bloom's Taxonomy

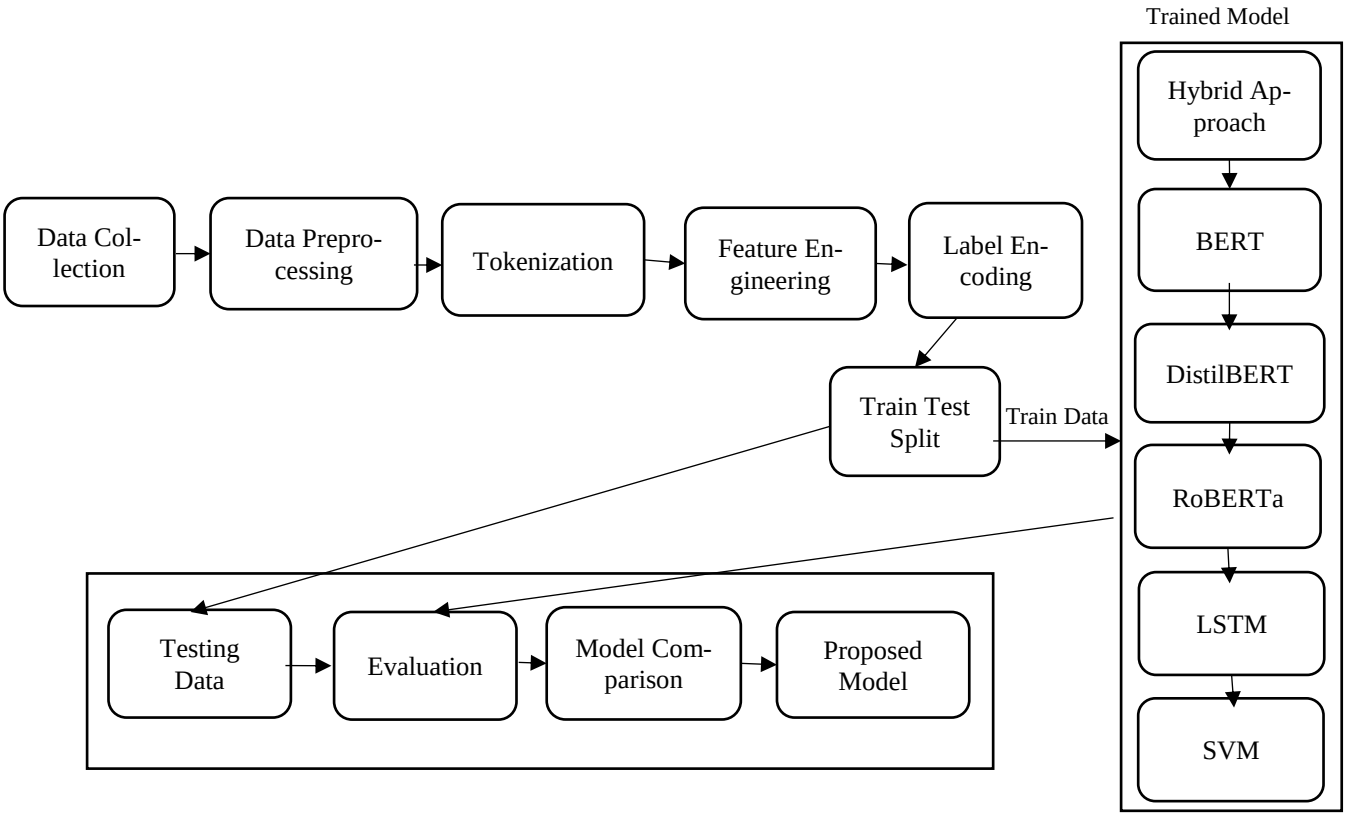

Figure 3. 1: Workflow Diagram for Question Prediction

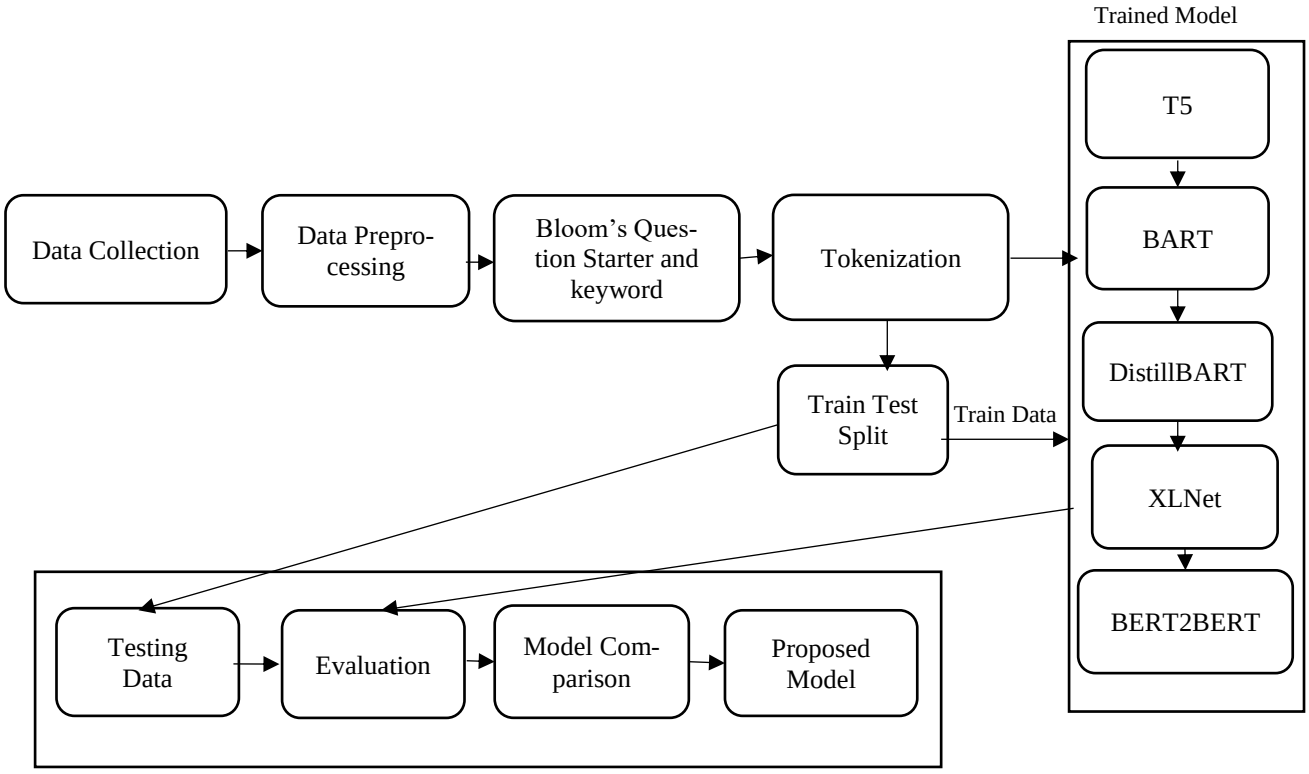

Figure 3.2: Workflow Diagram for Question Generation

3.3 Parameters to predict the cognitive level of Bloom's Taxonomy

When determining educational questions in terms of the Bloom Taxonomy by identifying and defining the specified cognitive indicators incorporated into a question, it is essential to initially determine which of these indicators are present in a question. In our research, we examine a widespread range of parameters that affect cognitive categorization. These parameters are derived from the linguistic structure, semantic content, and instructional context of each question. The database encompasses a diverse array of questions related to various levels of cognition, including remembering, understanding, applying, analyzing, evaluating, and creating. All parameters are chosen about the underlying cognitive demand, making it accurate to classify with the protection of both manual annotation and automation. The most important parameters considered in this Prediction are listed below.

Table 3.1 Keyword for Question Prediction

SI No.	Remember	Understand	Apply	Analyze	Evaluate	Create
1	Define	Arrange	Adapt	Advertise	Evaluate	Anticipate
2	Draw	Associate	Apply	Analyze	Judge	Assemble
3	Enumerate	Classify	Compute	Appraise	Justify	Collaborate
4	Find	Convert	Coordinate	Survey	Argue	Combine
5	Label	Describe	Demonstrate	Categorize	Defend	Compile
6	List	Discuss	Develop	Classify	Rate	Compose
7	Locate	Explain	Dramatize	Compare	Support	Construct
8	Match	Exemplify	Employ	Conclude	Score	Create
9	Memorize	Interpret	Establish	Connect	Select	Design
10	Name	Locate	Examine	Contrast	Mark	Develop
11	Recall	Match	Extrapolate	Correlate	Conclude	Devise
12	Recite	Paraphrase	Illustrate	Criticize	Recommend	Express
13	Record	Report	Implement	Deduce	Rank	Facilitate
14	Recognize	Research	Instruct	Devise	Estimate	Formulate
15	Select	Sort	Interview	Diagram	Prioritize	Generalize
16	State	Summarize	Manipulate	Differentiate	Validate	Hypothesize
17	Tabulate	Translate	Modify	Discriminate	Compare	Infer
18	Duplicate	Ask	Operate	Dissect	Refute	Integrate
19	Examine	Associate	Order	Distinguish	Grade	Intervene
20	Identify	Compare	Practice	Divide	Hire	Invent
21	Observe	Contrast	Predict	Estimate	Predict	Manage
22	Omit	Differentiate	Prepare	Evaluate	Verify	Modify
23	Quote	Discover	Produce	Experiment	Opinion	Negotiate
24	Read	Distinguish	Utilize	Explain	Prescribe	Originate
25	Repeat	Estimate	Act	Focus	Critique	Plan

26	Reproduce	Convert	Administer	Illustrate	Check	Prepare
27	Retell	Associate	Articulate	Infer	Determine	Produce
28	Tell	Infer	Calculate	Order	Decide	Propose
29	Visualize	Illustrate	Change	Plan	Debate	Reorganize
30	Recite	Contrast	Chart	Prioritize	Criticize	Report
31	Record	Visuals	Adapt	Select	Resolve	Revise
32	Rearrange	Distinguish	Build	Separate	Refine	Rewrite
33	Select	Detail	Compute	Subdivide	Disprove	Role-play
34	State	meaning	Prepare	Survey	Weigh	Simulate
35	Tabulate	Report	Practice	Rank	Assess value	Solve
36	Duplicate	Review	Transfer	Reason	Approve	Speculate
37	Examine	Cause	Sketch	Determine	Challenge	Make
38	Highlight	Defend	Employ	Infer motives	alternatives	Produce
39	Observe	Process	Perform	Map	Interpret rules	Rearrange
40	Omit	examples	Produce	Survey	Align crite-ria	Revise
41	Quote	Conclude	Measure	Sequence	Reflect	Animate
42	Read	relationships	Execute	Test	Reframe	Structure
43	Repeat	themes	Dramatize	sources	Rule on	Synthesize
44	Reproduce	Match	Organize	Points	Confirm	Engineer
45	Cite	Estimate	Choose tools	Identify	Gauge	Brainstorm
46	Note	Group	Interpret data	Decompose	Influence	Write
47	Index	Detail	Capture	Correlate	Perceive	Dictate
48	Pick	Elaborate	Design	Examine	Examine	Adapt
49	Express	Predict	procedures	Detect	solutions	Simulate
50	Demonstrate	Relate	Input	Reasoning	Select	Initiate
51	Spell	Express	Use tools	Prioritize	Review	Collaborate
52	Recognize	implications	Use	Solve	Release	Visualize

53	Define terms	Purpose	Execute	Interpret	Award	Code
54	Track	Rewrite	Implement	Structure	Find errors	Prototype
55	Recall details	Trace ideas	Demonstrate	logic	impact	Customize
56	Associate	Reconstruct	Calculate	Probe	Pros and cons	Rewrite
57	Locate	Comment	Solve	Sort	effectiveness	Perform
58	Identify dates	Sum results	Modify	Cluster	performance	models
59	Memorize	Review	Show	components	Check validity	Author
60	Index	Give	Illustrate	differences	Appraise quality	Launch
61	Alphabetize	Predict impact	Construct	Unpack	Argue points	Fabricate
62	Detect	effects	Change	Group s	arguments	Innovate
63	Isolate	Reframe	Operate	Trace origins	Rank responses	Direct
64	Discover	Rephrase	Experiment	Distinguish	evidence	Develop outline
65	Select	Translate ideas	Predict	inconsistencies	Weigh evidence	Sketch plan
66	Match	Modify	Relate	logic	appropriateness	Mix ideas
67	Recall names	key points	Manipulate	case studies	Scrutinize	Build
68	Number	Extrapolate	Develop	Verify facts	Measure	Shape ideas
69	Remember	procedures	Sequence	perspectives	Critique approach	Form solutions
70	Acknowledge	Units	Model	Map data	Performance	Invent
71	Recognize	Link ideas	Simulate	Distinguish relevance	Question validity	Devise campaign

72	Label	Use analogies	Adapt	Identify assumptions	Assess processes	Generate alternatives
73	Call	Provide rationale	Build	Assess bias	Compare results	Construct framework
74	Retain	Paraphrase sentences	Compute	Classify materials	Defend opinion	Compile list
75	State laws	Interpret quotes	Prepare	Correlate factors	Judge clarity	Design structure
76	Review	Visualize	Practice	Spot flaws	Rate decisions	Generate process
77	Reference	Present facts in context	Transfer	Break apart	Justify reasoning	Formulate policy
78	Write down	Interpret data	Sketch	Analyze findings	Score judgment	Build database
79	Print	Arrange	Employ	Organize arguments	Recommend actions	Create prototype
80	Decode	Describe change	Perform	Relate parts	Value contributions	Produce outline
81	Designate	Subtract	Produce	Split elements	Select priorities	Refactor
82	Store	Factor	Measure	Define structure	Validate	Handle
83	Tell facts	Synthesize meanings	Execute	Test components	Evaluate methods	Draft proposal
84	Register	Decode	Dramatize	Critique	Rate clarity	Plan
85	Copy	Depict	Organize	Spot	Debate issues	Design flow
86	Access	Visualize meaning	Choose	Identify strategy	Assess significance	Map strategy

87	Scan for data	Trace outcomes	Interpret data	Compare logic	Identify weaknesses	Illustrate concept
88	Count	Compare definitions	Initiate	Segment elements	Determine accuracy	Create story
89	Point	Describe impact	Design	Validate structure	Score effectiveness	Program
90	Recount	Relate facts	Show procedures	Outline flaws	Justify plans	Initiate project
91	Determine	Interpret conditions	Input	Isolate patterns	Review decisions	Develop workflow
92	Describe features	Sort	Use software/tools	Sift through evidence	Identify best options	Combine methods
93	Name characters	Group by function	Paint	Differentiate features	Critique findings	Develop strategy
94	Know	Translate language	Solve routine problems	Extract elements	Evaluate sources	Make
95	Show recall	State meaning	Translate theory	Draw inferences	Evaluate decisions	Produce
96	Provide	Recognize connections	Implement instructions	Unpack structure	Argue position	Rearrange
97	Check	Reconstruct meaning	Apply laws/principles	Assess variables	Judge accuracy	Revise
98	Map	Identify significance	Use a formula	Reclassify	Dispute	Animate
99	Relate	Define relationships	Apply methods	Diagnose faults	Check	Structure
100	Isolate facts	Link information	Test ideas	Trace outcomes	Determine	Synthesize

101	Match terms	Provide interpretation	Perform tasks	Relate categories	Decide	Engineer
102	Trace	Detect meaning	Use technology	Evaluate models	Debate	Brainstorm
103	List procedures	Reorder ideas	Modify plans	Map interactions	Criticize	From idea

3.4 Parameters to generate questions according to the cognitive level of Bloom's

Different Keyword and Question Starter Prompts can be used for Question Generation Aligned with Bloom's cognitive levels**.** This keyword and question Starter helps to form questions from the content. To start a question, a Question starter is used to start the question. The keyword is used to create a perfect question that addresses different cognitive levels of Bloom's taxonomy. Without considering both questions, the answer would be less accurate and perfect**.**

Keywords and questions considered for Question Generation from Different Contents are outlined below:

Table 3.2 Keyword and Question Starter for Question Generation

SI No	**Cognitive Level**	**Keywords**	**Question Starters**
1	**Level 1 - Remembering**	Choose, Define, Find, How, Label, List, Match, Name, Omit, Recall, Relate, Select, Show, Spell, Tell, What, When, Where, Which, Who, Why, Outline, Quote, Read, Review, Select, State, Study, Tabulate, Trace, Write, Cite, Enumerate, Index,	What is...? Where is...? How did ___ happen? Why did...? When did...? Can you recall...? Can you list the three...? Who were the main...? Which one…? How is…? When did ___ happen? How would you explain…? How would you describe…? Can you recall…? Can you select…? Can you list the three…? Who was…? What do you remember about _____? How do

		Indicate, Meet, Draw, Label, Point, Record,Pick,State,Reicte,Record,Reproduce,State,Note.relate,find,state, Know, MapPoint, Detect, Memorize, Isolate, Acknowledge, Associate, Label, Recognize,Locate,,Map,relate.	you define___? How do you choose____? List the _____ order in? How would you outline_____? How do you use it? What example can you find to____? How would you organize_____? What would result if____? What facts would you select to show___? What elements would you choose to change___? What other would you plan to_____?
2	**Level 2 - Understanding**	Classify, Compare, Contrast, Demonstrate, Explain, Extend, Illustrate, Infer, Interpret, Outline, Relate, Rephrase, Show, Summarize, Translate, Add, Approximate, Articulate, Associate, Characterize, Clarify, Convert, Defend, Extrapolate, Factor, Subtract, Rewrite, Translate, Picture graphically, Differentiate, Predict, Review, Describe, Distinguish, Elaborate, Estimate, Example, Explain, Express, Extend, Give, Infer, Observe, Paraphrase, Rewrite	How would you classify the type of...? How would you compare...? Can you explain what is happening...? What facts or ideas show...? What is the main idea of...? Which statements support…? Explain what is happening…? What is meant…? What can you say about…? Which is the best answer…? How would you summarize…? Which is the best answer__? How Would you describe___? How would you express____? What can infer from_____? Will you restate__? Elaborate on___. What would happen if_____? What is the main idea of____? What can you say about___? What you observe____? Distinguish between___. Graphically Present the_____.

3	**Level 3 - Applying**	Acquire, Adapt, Allocate, Alphabetize, Apply, Ascertain, Assign, Attain, Avoid, Back up, Calculate, Capture, Change, Classify, Complete, Compute, Construct, Customize, Demonstrate, Depreciate, Derive, Determine, Diminish, Discover, Draw, Employ, Examine, Exercise, Explore, Expose, Express, Factor, Figure, Graph, Handle, Illustrate, Interconvert, Investigate, Manipulate, Modify, Operate, Personalize, Plot, Practice, Predict,Prepare,Price, Process, Produce, Project, Provide, Relate, Round off, Sequence, Show, Simulate, Sketch, Solve, Subscribe, Tabulate	How would you use...? How would you solve...? What approach would you use to...? How would you show your understanding of...? Can you make use of the facts to...? What elements would you? Choose to…? What facts would you select to …? What questions would you ask in a …? What actions would you take from___? How would you develop_____to present? How do you apply____? How would you solve____? Why does____work? What examples can you find that_____? What example can you find that_____? How do you change_____? How do you present_____? Discover the method to solve_____? What would the result be if____? How would you modify_____? How you investigate it_____? What factors would you consider to solve the_____? How would you plot____? How would you Calculate_____?
4	**Level 4 - Analyzing**	Analyze, Audit, Blueprint, Breadboard, Break down, Characterize, Classify, Compare, Confirm, Contrast, Correlate, Detect, Diagnose, Diagram,	What are the parts or features of...? What inference can you make...? What is the theme? What motive is there? Can you list the parts? What inference can you make? What conclusions can you draw? Who would you classify…?

		Differentiate, Discriminate, Dissect, Distinguish, Document, Ensure, Examine, Explain, Explore, Figure out, File, Group, Identify, Illustrate, Infer, Interrupt, Justify	How would you categorize? Can you identify? What evidence can you find…? What is the relationship…? Can you distinguish between…? What is the function of…? What ideas justify…? Why do you think____?
5	**Level 5 - Evaluating**	Appraise, Assess, Compare, Conclude, Contrast, Counsel, Criticize, Critique, Defend, Determine, Discriminate, Estimate, Evaluate, Grade, Hire, Interpret, Judge, Justify, Measure, Predict, Prescribe, Rank, Rate, Recommend, Release, Select, Summarize, Support, Test, Validate, Verify	Do you agree with the actions...? Would it be better if...? How would you rate the...? What would you cite to defend the actions...? What judgment would you make about...? What information would you use to prioritize _____? What choice would you have made _____? What data was used to evaluate _____? How could you verify____? Rank the Importance of____.
6	**Level 6 - Creating**	adapt, build, change, choose, combine, compile, compose, construct, create, delete, design, develop, discuss, elaborate, estimate, formulate, improve, invent, make up, Abstract, Animate, Arrange, Assemble, Budget, Categorize, Code, Compile, Cope, Correspond, Create, Cultivate, Debug,	What changes would you make to solve...? How would you improve...? Can you propose an alternative...? How would you adapt ___ to create a different...? What changes would you make to revise...? Design a solution for__? Create a plan for? What do you invent to solve___? What alternative would you suggest for _____? Devise a way to _____. How would you compile the facts for

		Enhance, Generalize, Generate, Handle, Improve, Integrate, Interface, Join, Model, Overhaul, Portray, Reconstruct, Reorganize, Revise, Rewrite, Specify	_____? How would you portray____? How would you enhance the____? How would you change_____? How would you specify____? Devise a way to reorganize____ What changes would you make to formulate___? What changes would you make to Generate___?

3.5 Dataset and Data Collection

In order to begin training a transfer learning based model, a well-constructed dataset is necessary. A successful undertaking of the preprocessing process of the dataset not only reduces the time of computation but also brings the model to convergence faster, enhances the performance and the accuracy of the model.

Data collection is the initial phase of work that has a substantial impact on the performance of any machine learning system. A diverse, well-annotated dataset ensures the effective learning and generalization of the models. In our work, the primary goal is to develop a robust system for the automated prediction and generation of educational questions in alignment with Bloom's Taxonomy of Cognitive Domains. Accordingly, two distinct datasets were curated:

i. Question Prediction Dataset

ii. Question Generation Dataset

3.5.1 Dataset to predict questions according to cognitive level

A set of 8,768 individual educational questions was composed of a broad collection of available sources, such as: Textbooks, Online learning resources, Exam question papers, Educational websites, and free publicly accessible question banks The questions were then annotated manually according to Bloom's Taxonomy levels so they fell into Lower-Order Thinking Skills (LOTS) and

Higher-Order Thinking Skills (HOTS). In order to achieve the representative balance, the questions were taken based on various topics such as Computer Science, Physics, Biology, and Mathematics. This even spread to levels of cognitive ability and fields of study makes the Prediction model more generalized and more robust.

Table 3.3 Summary of Question Prediction Dataset

Sources Diversity	Number of Questions
Online platforms, Examination Papers, Education websites, Question Banks	8,768

Table 3.4 Data Description for Question Prediction

SI NO.	**Question**	**keyword**	**Bloom's Cognitive Level**
1	What is the chemical symbol of water?	Symbol	Remember
2	Describe the steps to solve a quadratic equation	Describe	Understand
3	Name the components of a computer	Name	Remember
4	Interpret the meaning of a metaphor	Interpret	Understand
5	Build a sustainable urban transportation model for a city with over 10 million residents.	Build	Create
6	Discuss how climate change impacts agriculture in developing countries.	Discuss	Understand
7	Examine the trend in global CO2 emissions using the provided graph.	Examine	Analyze
8	Compose an algorithm that sorts data efficiently.	Compose	Create
9	Compute the volume of a cylinder	Compute	Apply

10	Calculate for x in the following equation: 12x=36	Calculate	Apply
11	Explain How photosynthesis works in plants	Explain	Understand
12	Compare mitosis and meiosis	Compare	Understand
13	Examine the cause and effect of air pollution	Examine	Analyze
14	Develop a simple calculator using Java	Develop	Create
15	Differentiate between renewable and non-renewable energy	Differentiate	Analyze
16	Interpret the following economic growth chart	Interpret	Analyze
17	Recall the main events of French Revolution	Recall	Remember
18	List three types of triangles	List	Remember
19	Identify the capital city of japan	Identify	Remember
20	Summarize main points of a speech	Summarize	Understand
21	Contrast Hardware and Software	Contrast	Analyze
22	Defend the use of mobile phones	Defend	Evaluate
23	Propose a new law to improve road safety	Propose	Create
24	Design an energy-efficient model using the given data on renewable energy trends.	Design	Creating
25	Investigate the impact of deforestation on global biodiversity using the provided deforestation rate data for the last 20 years.	Investigate	Analyze
26	Evaluate the effectiveness of renewable energy policies in reducing carbon emissions.	Critique	Evaluate
27	Design a cost-effective water purification system for rural communities	Design	Create

. . . .	. . . .	. . . .	. . . .
8765	Invent a device that help to reduce plastic waste	Invent	Create
8766	Explain the difference between hardware and software	Explain	Understand
8767	Illustrate the process of the digestive system of human body	Illustrate	Understand
8768	Justify the effectiveness of global vaccination campaigns.	Justify	Evaluate

3.5.2 Dataset to generate questions according to cognitive level

Similarly, the model that was used in automated question generation was trained with 4,630 educational contents. These were also obtained through various sites, like textbooks, online learning tutors, and question banks that were publicly available. Here, attention was paid to teaching the model to come up with new questions with the same levels as the taxonomy developed by Bloom, covering both categories, LOTS and HOTS. Various reliable sources of education were used when selecting the content in order to cover the content widely and be instructionally relevant.

Table 3.5 Summary of Question Generation Dataset

Sources Diversity	Number of Contents
Online platforms, Examination Papers, Education websites, Question Banks	4,630

Table 3.6 Data Description for Question Generation

SL no	Context/Passage	Bloom's Level	key-word	Question Starter	Generated Question
1	The process of photosynthesis involves plants converting sunlight into energy and releasing oxygen.	Understand	describe	Can you	Can you describe how photosynthesis contributes to the oxygen supply on Earth?
2	The water cycle includes evaporation, condensation, and precipitation, which circulate water throughout the environment.	Remember	process	What is the	What is the process of the water cycle, and how does it work?
3	Newton's third law states that every action has an equal and opposite reaction.	Analyze	differences	What are the	What are the differences between Newton's third law and the law of inertia in terms of motion?
4	The concept of relativity revolutionized the understanding of space, time, and gravity.	Understand	Explain	How does you	How do you explain Einstein's theory of relativity, which challenges classical mechanics?
5	An algorithm is a step-by-step procedure for solving a problem or task, often used in computer programming.	Apply	Demonstrate	Can you	Can you demonstrate how a binary search algorithm works on a sorted list?
6	The quadratic formula is used to find the solutions of	Apply	Solve	Can you	Can you solve the quadratic equation

	a quadratic equation ax^2+bx+c = 0				2x^2+3x−5=0 using the quadratic formula?
7	The theory of evolution proposes that species change over time through natural selection.	Create	Design	What alternative model could you	What alternative model could you design to explain the changes in species over time, apart from evolution?
8	Advances in artificial intelligence are transforming industries, such as healthcare and automotive manufacturing.	Evaluate	use	In what ways Can you	In what ways can the use of AI in healthcare lead to ethical dilemmas regarding privacy and decision-making?
9	The process of cellular respiration involves converting glucose into energy in cells	Understand	describe	Can you	Can you describe how cellular respiration contributes to energy production?
10	The theory of evolution by natural selection explains how species adapt and evolve over time.	Understand	explain	Can you	Can you explain how Darwin's theory of evolution is supported by fossil records?
. . . .	. . . .	. . . .	. . . .	. . . .	. . . .
4628	Operating systems manage computer hardware and	Understand	Clarify	How would you clarify	How would you clarify the meaning of an

	software resources and provide essential services for computer programs.			the meaning of	operating system's role in resource management?
4629	The law of supply and demand states that when the supply of a good increases and demand remains constant, the price tends to fall.	Remember	Relate	How does it	What is thermodynamics, and how does it relate to energy transfer?
4630	Sustainable development involves meeting the needs of the present without compromising the ability of future generations to meet their own needs.	Analyze	Impact	How would you analyze the	How would you analyze the impact of sustainable development practices on economic growth?

3.6 Data Preprocessing

In order to support the process of classifying questions in terms of the Bloom Taxonomy as well as generation of questions, preprocessing pipeline was used to clean and normalize the data including standardization. The step is critical in the reduction of the linguistic noise and the increase in the data consistency to have the textual data of quality, concise, and semantically rich. In using this pipeline, the data is streamlined to ensure an effective prediction to the Bloom-based cognitive level and question formulation to the respective cognitive level.

3.6.1 Data Preprocessing for Question Prediction

i. Text Cleaning: All the texts of the context were lowered to lower cased and URLs, punctuations, special characters, and excessive whitespace were removed. That action creates consistency and less noise in the dataset.

iii. Tokenization: A pre-trained RoBERT tokenizer was applied to clean and stem text. This is done to transform text to sub word level but preserve both syntactic and contextual features that are important to subsequent semantic processing.

iii. Stemming: Stemming was used in converting the words into roots. This assists in reframing semantically similar words and reduces the sparsity of vocabulary, leading to better generalization of the model.

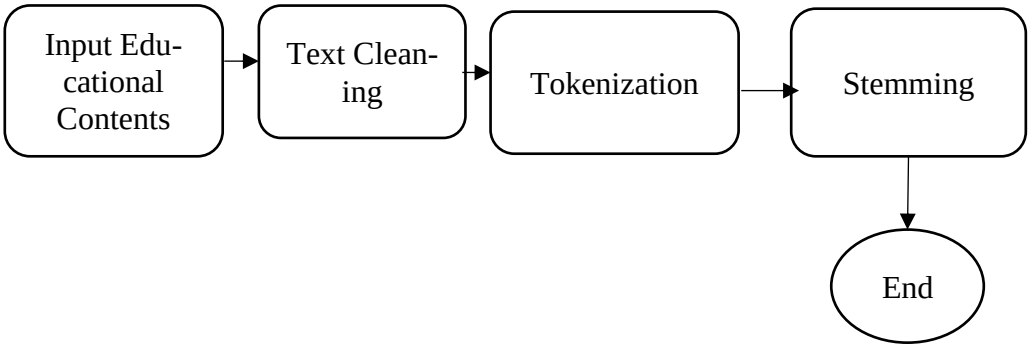

Figure 3.3 Data Preprocessing for Question Prediction

Table 3.7 Data Cleaning for Question Prediction

Question	**Text Cleaning(Stop Words Removal & Lowercasing)**	**Tokenization**	**Stemming**
Explain how Newton's Laws apply to driving a car safely.	'explain', 'newton', 'laws', 'apply', 'driving', 'car', 'safely'	'Explain', 'how', 'newton', "'s", 'laws', 'apply', 'to', 'driving', 'a', 'car', 'safely', '.'	'explain', 'newton', 'law', 'appli', 'drive', 'car', 'safe'
What steps should be taken to design a secure database system?	'steps', 'taken', 'design', 'secure', 'database', 'system'	'what', 'steps', 'should', 'be', 'taken', 'to', 'design', 'a', 'secure', 'database', 'system', '?'	'step', 'take', 'design', 'secur', 'databas', 'system'

3.6.2 Data Preprocessing for Question Generation

i. Text Cleaning: All textual contents were lowercased, and the URLs, punctuations, special characters, and APIs were removed. The reason is that this process provides consistency and removes noise in the dataset.

ii. Combination of the Content, Bloom level, and question starter: the set of the content, Bloom level, and question starter were entered into a single input string. This framework assists the model to come up with questions, according to the Bloom taxonomy levels.

iii. Tokenization: The cleaned and lemmatized text was tokenized by a pre-trained tokenizer (e.g., T5Tokenizer). This transforms the input and target texts to the subword units, so that syntactic and semantic data remain intact to generate the questions.

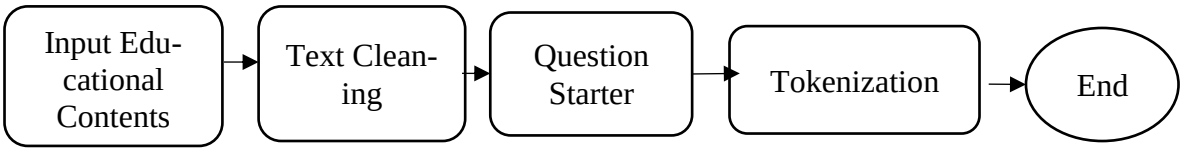

Figure 3.4: Data Preprocessing For Question Generation

Table 3.8 Data Cleaning for Question Generation

Content	**Bloom's Level**	**Question Starter**	**Text Cleaning**	**Tokenization**
Machine learning is a subset of artificial intelligence that focuses on using algorithms to identify patterns in data and make predictions.	Analyzing	"What process would you follow to"	"Machine learning is a subset of artificial intelligence that focuses on using algorithms to identify patterns in data and make predictions"	['machine', 'learn', 'is', 'a', 'subset', 'of', 'artificial', 'intelligence', 'that', 'focus', 'on', 'use', 'algorithm', 'to', 'identify', 'pattern', 'in', 'data', 'and', 'make', 'prediction', '|', 'Bloom', "'s", 'Level:', 'Analyzing', '|', 'Starter:', 'What', 'process', 'would', 'you', 'follow', 'to']

DevOps is a set of practices that combine software development and IT operations to automate and improve the process of building, testing, and deploying software.	Evaluating	"What strategy would you use for"	"devops is a set of practices that combine software development and it operations to automate and improve the process of building testing and deploying software"	[devop', 'is', 'a', 'set', 'of', 'practic', 'that', 'combin', 'softwar', 'develop', 'and', 'it', 'oper', 'to', 'autom', 'and', 'improv', 'the', 'process', 'of', 'build', 'test', 'and', 'deploy', 'softwar', '|', 'Bloom', '"s", 'Level:', 'Evaluating', '|', 'Starter:', 'What', 'strategy', 'would', 'you', 'use', 'for']

3.7 Data Augmentation

Data augmentation was employed to expand the training set and address the issue of class imbalance. Raw samples were initially gathered, totaling 8,768. To improve the robustness of the model and support the balanced representation of all six cognitive levels of Bloom's Taxonomy, data augmentation procedures were employed, resulting in a larger dataset comprising 15,492 data samples. These methods involved creating syntactic differences in the original questions, such as rephrasing or modifying sentence structures, to expose the model to greater diversity in input form. This method not only enhanced the generalization capacity of the model but also the flexibility in responding to alternative formulations of questions. Diversification of the training data made the model more capable of handling different question types in a real-life task, allowing for more objective learning and testing the model in all aspects of cognition.

3.8 Train Test Split

The preprocessed datasets were also split based on training and testing subsets in a stratified manner to assess the generalization capability of the model. In case of both question Prediction training

and question generation training, an 80:20 ratio was followed so as to have enough training data, but at the same time have an evaluation set representative in size.

When training on the Prediction task, stratification was used at the levels of Bloom's Taxonomy to avoid class imbalance. In the same way, the data used to generate questions was divided equally, maintaining the proportion of the distribution among the cognitive levels and academic topics.

Table 3.9: Summary of the dataset splitting

Task	Total Samples	Training	Testing Set
Question Prediction	15,492	12,393	3,099
Question Generation	4,630	3,704	926

3.9 Model Development to predict the cognitive level for the given question

This is the most important phase, which includes building the models that will aid in the Prediction task. We implement different models, including machine learning and transformer models, for the question Prediction task. We used BERT, DistilBERT, ROBERTa, LSTM, SVM, and a Hybrid model combination. All of these algorithms will be discussed in the following subsections.

3.9.1 Model Using BERT

Based on Bidirectional Encoder Representations of Transformers (BERT) illustrate in Figure 3.5 [31].The system identifies the questions under Bloom's cognitive categories. BERT pre-trained model is an excellent fit for working on a task that requires text Prediction since the model performs extremely well in recognizing the meaning of words in the context of a sentence. BERT model is perfected with the help of a set of test questions previously grouped based on the levels of Bloom's Taxonomy. This refinement refines the model to determine the various levels of cognitive complexity, such as determining the nature of the question, whether the question to be answered requires a complex analysis or mere recalling. To acquire semantic relations and contextual

semantics, the prediction method transforms the input question into a high-dimensional vector representation.

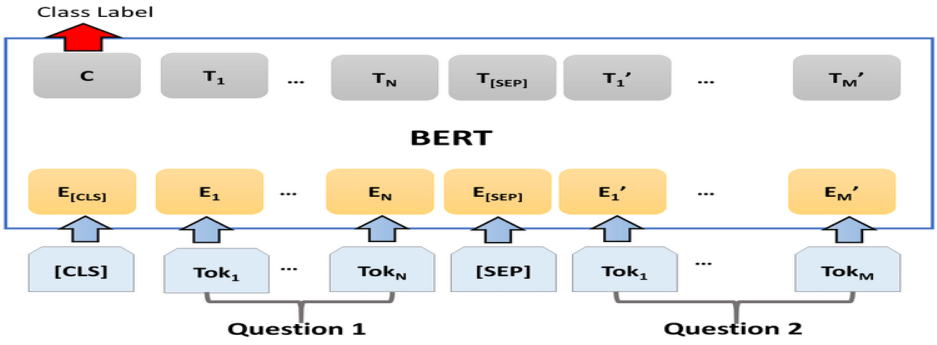

Figure 3.5: Bert Model Architecture [32]

3.9.2 Model Using DISTILLBERT

DISTILLBERT is a more knowledge-distilled version of BERT and is smaller and more efficient illustrate in Figure 3.6 [32]. DISTILLBERT will carry over the contents of the bigger version of BERT into a smaller model that has a lot of the accuracy of the original and is quicker and uses less processing capacity. Regarding the prediction activity, it can be employed to estimate the cognitive levels of Bloom, with less computational cost although DISTILLBERT.

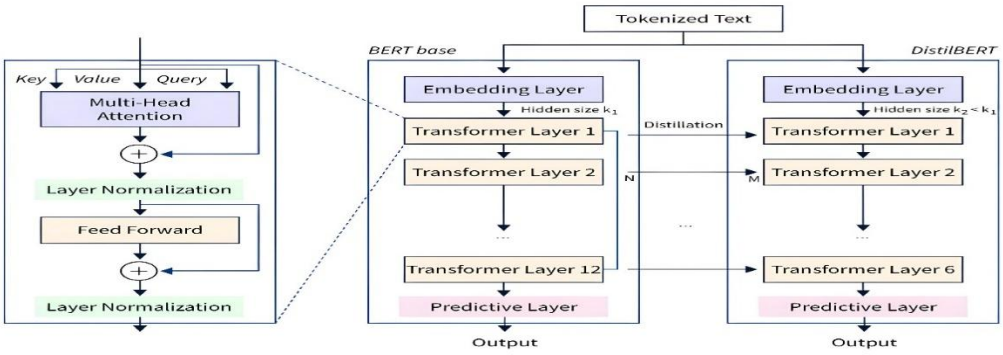

Figure 3.6 DistilBERT Architecture [33]

3.9.3 Model Using RoBERTa

RoBERTa is a fine-tuned form of BERT, which was trained upon more effective techniques such as dynamic masking and increased batch shows in Figure 3.7 [33]. It ablates the Next Sentence

Prediction (NSP) task, leading to a more robust model that does better than BERT in multiple tasks. One of the applications of ROBERTa includes Bloom level prediction, where difficult questions are properly grouped into difficult and easy questions based on the level of cognitive demands. Its ability to train on bigger data and better techniques enables it to tackle complex tasks in a better manner. The augmented question training procedure renders the training of Roberta very effective as far as comprehension and categorization of questions are concerned.

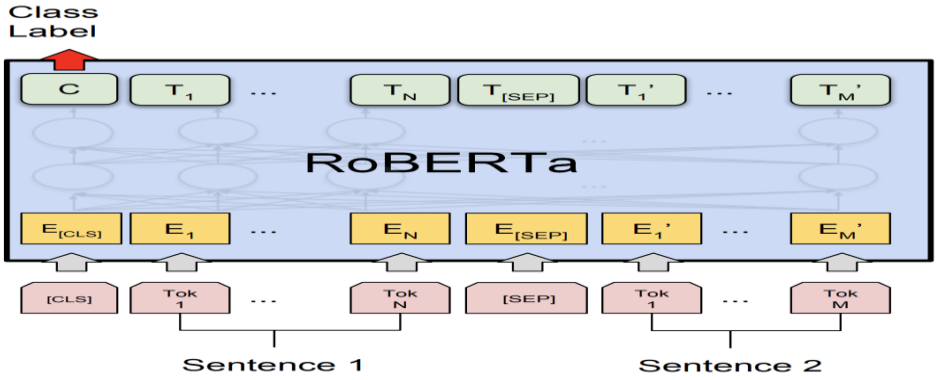

Figure 3.7 RoBERTa Model Architecture [34]

3.9.4 Model using LSTM

LSTM (Long Short-Term Memory) is a special form of RNN (Recurrent Neural Network), which is used in a sequential context and helps address the vanishing gradient problem regular RNNs have illustrate in Figure 3.8 [34]. It has input, output, and forget gates on memory cells that enable storing long-term dependencies and is thus useful in context processing and understanding of sequences. The LSTM is used in the level prediction of Bloom, where it defines questions by their cognitive complexity by capturing the contextual flow of words in an entire question. It is appropriate in categorizing educational material in which word sequence and meaning development are vital because it is capable of retaining significant information over a long series.

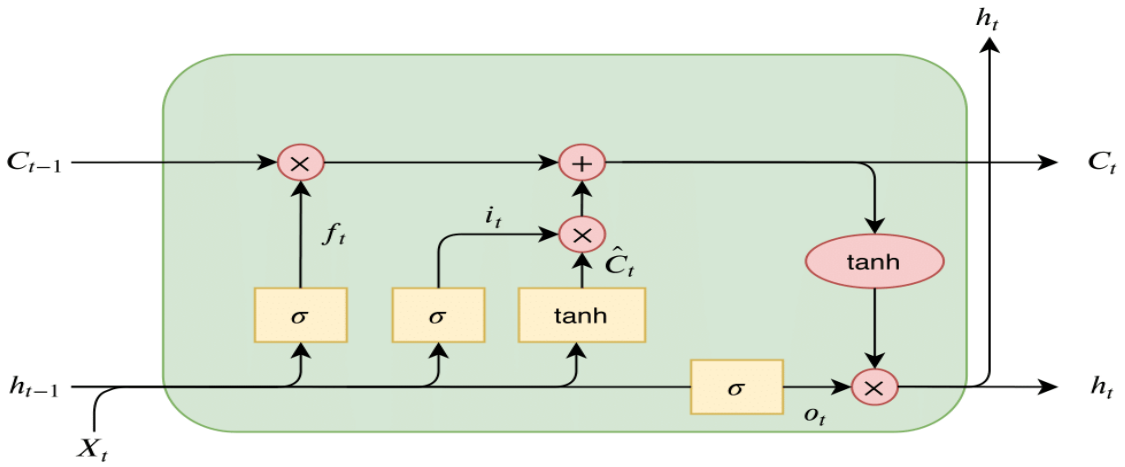

Figure 3.8 Single Unit LSTM [35]

3.9.5 Model Using SVM

SVM (Support Vector Machine) is a strong classical machine learning algorithm and is applied to perform classification shows in Figure 3.9 [35]. It operates in a manner that is optimal in finding the hyperplane that can best partition the data points into dissimilar groups. SVM plays the role of classifying the level of questions in terms of cognitively based on feature representation of questions, such as TF-IDF vectors, which are analyzed using SVM in level classification according to Bloom. SVM excels when working with highly-dimensional structured data and, provided that the dataset size is not very high, can be quite efficient. Its advantages are associated with toughness and the propensity to generalize.

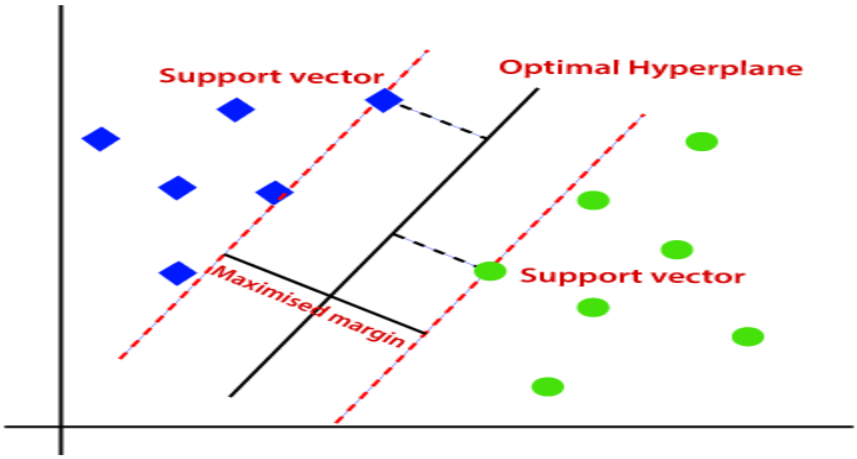

Figure 3.9 Support Vector Machine [36]

3.9.6 Hybrid Model Using RoBERTa, LSTM, CNN, and TF-IDF

The Hybrid Model combines the strengths of contextual representation, sequential representation, and feature-based representation by combining the abilities of RoBERTa, LSTM (Long Short-Term Memory), CNN (Convolution Neural Network), and TF-IDF (Term Frequency-Inverse Document Frequency). The hybrid method is especially useful in the case of question prediction because it is indispensable to capture both the deep semantic notion of the text and its flow. RoBERTa has the contextual understanding, which would be fundamental when it comes to categorizing questions into the level of Bloom's cognitive levels,

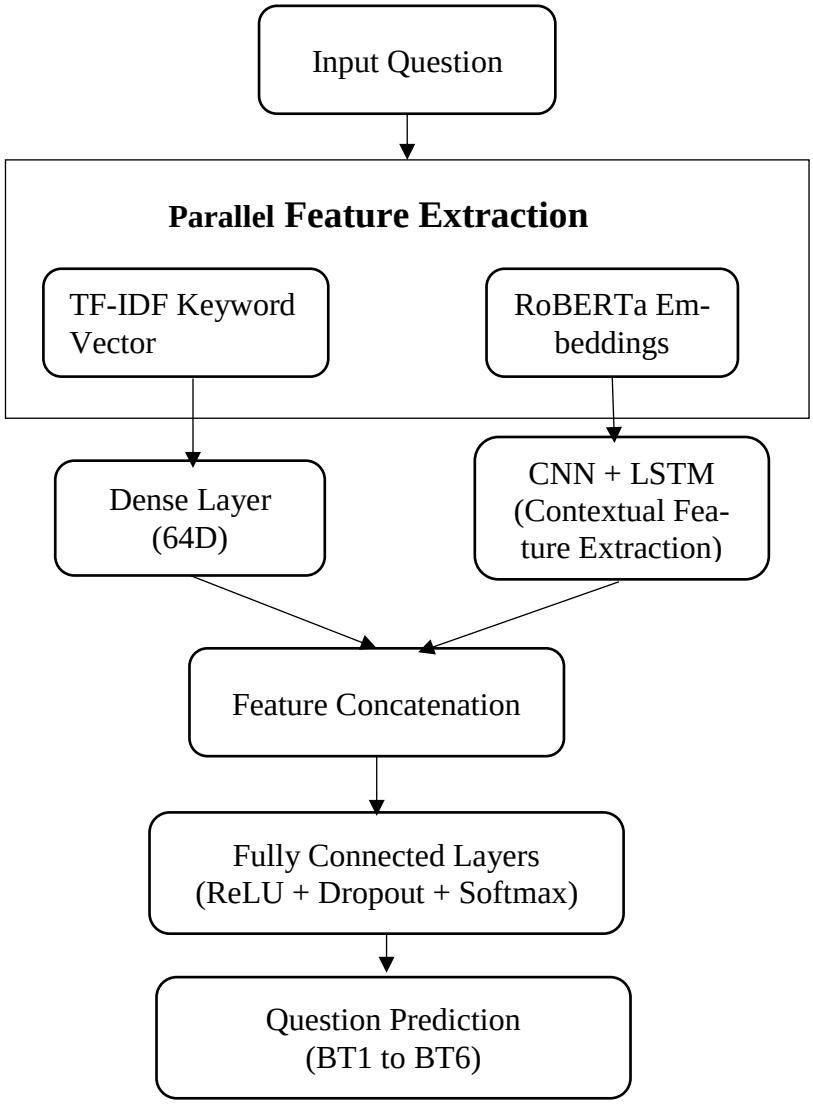

Figure 3.10 Proposed Hybrid Model for Prediction

Where the slight differences in semantics are based .LSTM is a particular form of the Recurrent Neural Network (RNN), which is meant to be very helpful when it comes to sequential data. In contrast to the classical feed-forward systems, LSTM implements long-range dependencies in the text and preserves a memory of the past tokens, which is essential to comprehend questions with the meaning spread over several words. In this hybrid model, LSTM is applied to capture the order in the text, which is how various words and phrases relate to each other in a question.

Using the memory of LSTM and contextualization in Roberta, this combination enables the model to grasp the general meaning of a question and follow the flow of ideas simultaneously. And also, CNN is used for local patterns.TF-IDF is the statistical process that can tell the significance of the words within a document compared to the whole corpus. TF-IDF also gives greater value to the use of unique or distinctive words in a particular topic or concept by taking into consideration the frequency of word occurrence on individual questions and as a whole in the dataset. This operation of the preprocessing pipeline will assist in the identification of important words that can be vital in the prediction, particularly between the levels of cognition in Bloom's Taxonomy.

3.10 Model Development to generate questions according to Bloom's taxonomy

We implement different models and transformer models for the question task. We've used T5, BART, DistillBART, XLNet, and BERT2BERT. All of these algorithms will be discussed in the following subsections.

3.10.1 Model Using with T5

Advanced transformer models are used in the question generation process, specifically T5 (Text-to-Text Transfer Transformer), and these models have been specifically honed to produce instructional materials that align with Bloom's Taxonomy's various cognitive levels shows in Figure [36]. And it's well-known for using its deep knowledge of linguistic structures to produce text that is both coherent and contextually relevant. In this structure. The T5 model is more effective at producing well-structured questions because it transforms all natural language processing tasks into a standard text-to-text format. Multiple-choice true/false and open-ended questions are among the question types that the

Model can generate thanks to its versatility. It ensures that the questions it generates are contextually accurate and educationally appropriate by incorporating domain-specific datasets during the fine-tuning phase.

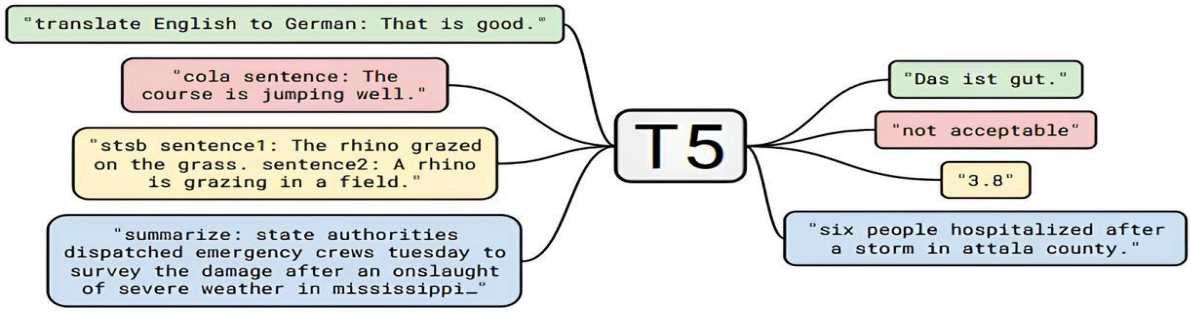

Figure 3.11: Text-to-Text Transfer Transformer, adapted from [37]

3.10.2 Model using BART

BART ((Bidirectional and Auto-Regressive Transformers)) is a diminishing auto encoder designed for sequence-to-sequence tasks. It combines the bidirectional encoder of BERT with the autoregressive decoder of GPT, making it highly effective for text generation tasks. For question generation, BART was fine-tuned to generate grammatically correct and contextually relevant questions based on the given content and Bloom's level. Its powerful encoder-decoder architecture allows it to handle both understanding and generation efficiently. That architecture shown in Figure 3.12 [37].

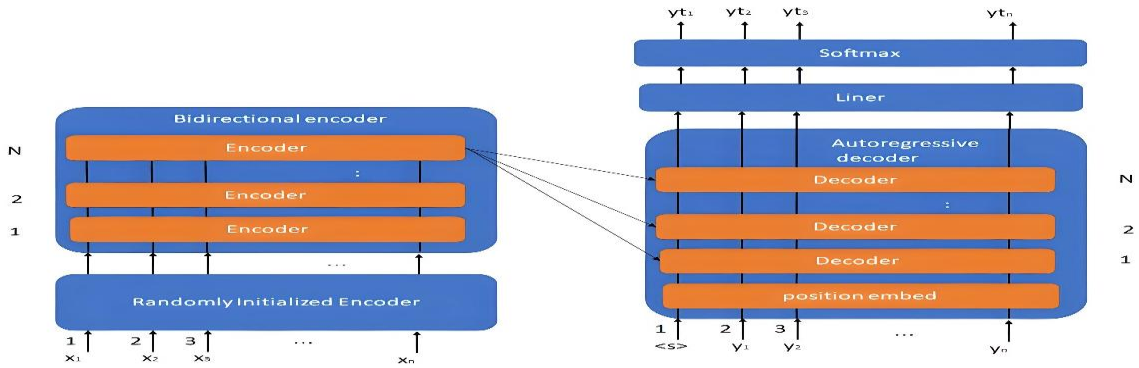

Figure 3.12 Bidirectional and Auto-Regressive Transformers [38]

3.10.3 Model using DISTILLBART

DISTILLBART is a smaller and more efficient version of BART, created through knowledge distillation. For question generation, DISTILLBART was fine-tuned to maintain high-quality question generation while reducing inference time. Its efficiency makes it suitable for deployment in resource-constrained environments without compromising too much on accuracy.

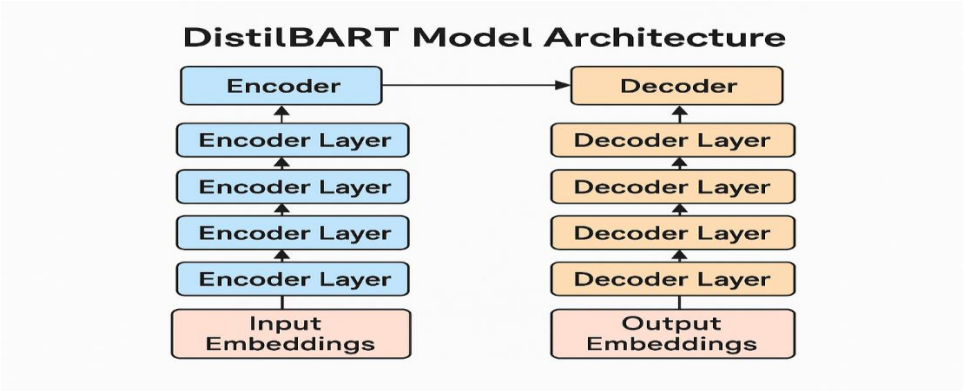

Figure 3.13 DistillBART Architecture

3.10.4 Model using XLNet

XLNet is an autoregressive transformer model that combines the advantages of BERT's bidirectional context with GPT's autoregressive generation. It models long-range dependencies more effectively than previous models. XLNet was used for question generation, leveraging its ability to capture complex dependencies and context in the input text. It architecture show in Figure 3.14 [33].

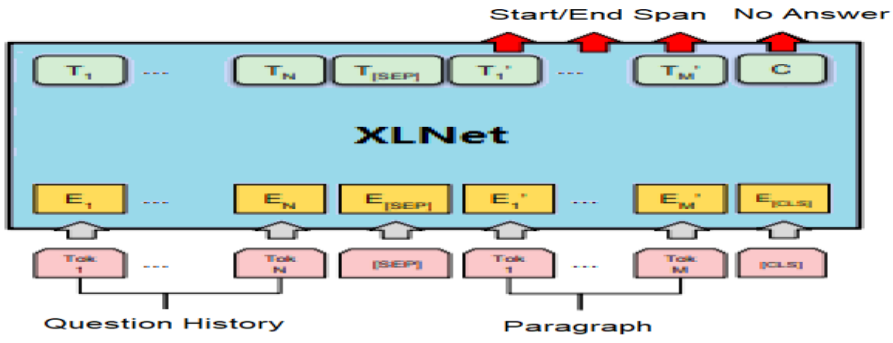

Figure 3.14 XLNet Architecture [39]

3.10.4 Model using BERT2BERT

BERT2BERT is a sequence-to-sequence model using two separate BERT models: one for encoding and one for decoding shown in Figure 3.15 [39]. For question generation, BERT2BERT was fine-tuned to generate questions from the given content and Bloom's taxonomy levels.

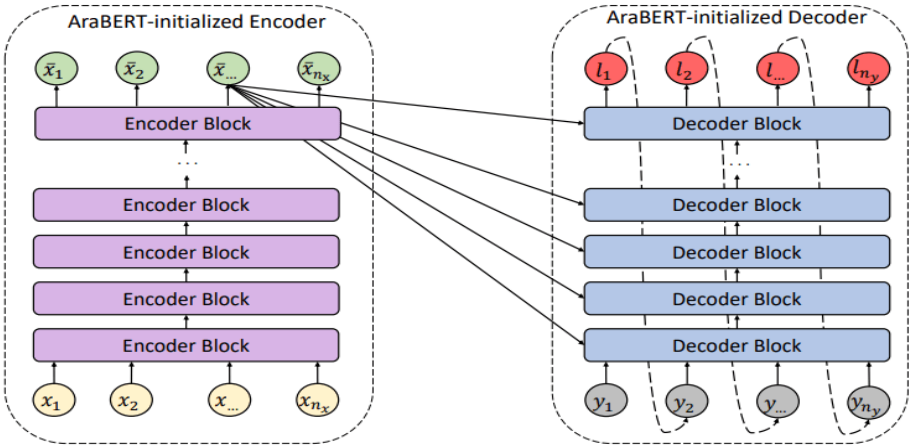

Figure 3.15 BERT2BERT Architecture [40]

Chapter 4

Implementation

4.1 Introduction

In this part, we discuss the implementation of a Hybrid Prediction Model for predicting educational questions based on Bloom's Taxonomy and a T5-based Question Generation Model for generating questions from given content aligned with Bloom's Taxonomy levels.

4.2 Hybrid Model Implementation for Bloom's Taxonomy Question Prediction

The Model categorizes educational questions into Bloom's six cognitive levels (Remember–Create). The implementation includes: This model classifies educational questions into categories (e.g., Remember, Understand) based on the content and text analysis. It combines TF-IDF vectorization, RoBERTa embeddings, and a CNN-LSTM hybrid architecture to accurately classify the questions.

1. **Feature Extraction**

The tokenized text is transformed into numerical features using TF-IDF (Term Frequency-Inverse Document Frequency), which helps quantify the importance of each word in the context of the entire corpus. Then, RoBERTa embeddings for contextual feature extraction. These embeddings provide rich representations of the text, which are further processed using CNN layers to capture local patterns (e.g., phrases or key terms) and LSTM layers to capture long-range dependencies between words.

2. Data Handling and Label Encoding

The Bloom's Taxonomy levels represent the target labels for prediction. Since machine learning models cannot work directly with categorical labels, these Bloom levels are encoded into a numerical format using LabelEncoder. It converts each category (e.g., BT1, BT2) into a unique integer, making it suitable for model training. For example, BT1 could be encoded as 0, BT2 as 1, BT3 as 2, etc. This encoding ensures that the model can compute predictions based on numerical representations of the categories.

3. Prediction

After feature extraction and data handling, the model predict educational questions into Bloom's Taxonomy levels by concatenating the outputs from CNN (capturing local patterns) and LSTM (capturing long-term dependencies). These combined features are passed through a fully connected layer with a softmax activation function, which produces a probability distribution over the possible categories.

4.3 T5 Model Implementation for Question Generation

To automate the creation of educational questions, a T5-small model was fine-tuned on a dataset containing instructional content, Bloom's levels,

1. Input Formatting with Cognitive Level Conditioning

Each training input was structured as:

"Generate question: <Content> | Bloom's Level: <Level>"

This format conditioned the model not only on the instructional content but also on the intended Bloom's level (e.g., BT1: Remember, BT6: Create), allowing it to generate questions aligned with specific cognitive objectives.

2. Tokenization and Padding

The T5Tokenize**r** was used for both input and target sequences. Input texts were tokenized with a maximum length of 512 tokens, while output questions were truncated or padded to 128 tokens. This ensured compatibility with GPU memory constraints while preserving relevant semantic information.

3. Transfer Learning and Fine-Tuning

The model was initialized with pertained T5-small weights and fine-tuned on the educational dataset. Only task-specific weights were updated, preserving the general linguistic understanding learned during retraining.

4.4 Optimization Techniques

In the model development phase, a variety of training strategies and hyperparameter optimization techniques, including the Adam optimizer, were applied to enhance the model's performance, generalization, and resilience. These methods ensure the model is robust, efficient, and capable of accurately handling both question prediction and question generation tasks. The following techniques were employed:

4.4.1Optimization Technique for Question Prediction

To ensure the model performs well on new, unseen data, several strategies were used to prevent overfitting and improve generalization:

1. Regularization: Techniques such as dropout and L2 regularization were employed to reduce the model's complexity and prevent it from memorizing the training data.

2. Learning Rate Modifications: Learning rate schedulers and adaptive learning rate optimizers were utilized to adjust the learning rate during training, allowing the model to converge smoothly and avoid overfitting. Here we use Cosine Annealing for dynamic learning rate adjustment. Focal Loss: To address Class imbalance, Focal Loss is used, ensuring better performance for harder-to-classify examples

3. Early Stopping: Training was halted when the validation performance stopped improving for a certain number of epochs. This prevents unnecessary training, saving computational resources and avoiding overfitting.

4.4.2 Optimization Technique for Question Generation

To ensure the model performs well on new, unseen data, several strategies were used to prevent overfitting and improve generalization:

AdamW Optimizer: Optimizes model weight with weight decay

Early Stopping: Training was halted when the validation performance stopped improving for a certain number of epochs. This prevents unnecessary training, saving computational resources and

avoiding overfitting.

4.5 Performance Measurement

Performance Measurement means verifying that the system is successfully generating and classifying questions, which requires evaluating the model. The system's performance will be assessed using the following metrics.

4.5.1 Question Prediction

Whether the Weather Model is working perfectly or not can be evaluated through the following matrices

Accuracy evaluates how many questions were correctly classified out of all the questions..

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions} \quad (4.1)$$

Precision: calculates the percentage of the classified questions that were right.

$$Precision = \frac{True\ Positivs}{True\ Positives + False\ Positives} \quad (4.2)$$

Recall (Sensitivity): evaluates how well the model finds every pertinent question.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (4.3)$$

F1-Score: A balance between precision and recall.

$$F - 1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)$$

4.5.2 Question Generation

BLEU (Bilingual Evaluation Understudy) Score

BLEU is a widely used metric in machine translation and natural language generation tasks. It compares the n-gram**s** (i.e., contiguous sequences of n words) in the generated question against those in one or more reference questions. The score ranges from 0 to 1, with 1 indicating perfect matches between the generated and reference questions. A higher BLEU score indicates that the generated questions are more similar to the reference questions.

$$BLUE = BP.exp \sum_{n=1}^{N} w_n . logp_n \quad (4.5)$$

Where,

BP = Brevity Penalty to penalize short sentences.

p_n is the precision for n-grams (e.g., unigram, bigram, etc.).

w_n is the weight assigned to each precision value (typically equal across all n-grams).

Diversity Score

The Diversity Score quantifies the variety in the generated questions, ensuring the model produces a wide range of unique and diverse questions, rather than repetitive ones.

$$Diversity\ Score = \frac{Unique\ n-grams}{Total\ n-grams} \quad (4.6)$$

Expert Evaluation

It involves human reviewers assessing the quality, relevance, and grammatical correctness of the generated questions, ensuring they align with Bloom's cognitive levels and educational goals.

However, it usually depends on several criteria, such as:

- Clarity: How clear and understandable the generated question is.
- Relevance: How well the question matches the intended Bloom's cognitive level.
- Grammatical Correctness: How grammatically correct the question is.

4.6 Evaluation on an Independent Test Set

To further assess the robustness and generalizability of the proposed system, evaluation was conducted on an independent dataset not used during training or initial validation. This separate testing focused on both the question prediction and question generation components to ensure comprehensive performance analysis. For the prediction task, the system's ability to accurately classify questions into Bloom's cognitive levels was evaluated on diverse, unseen academic content. For the generation task, the model's effectiveness in producing coherent, contextually relevant, and pedagogically sound questions was similarly examined. This evaluation provides a rigorous measure of the system's applicability and reliability across varied educational settings.

4.7 Implementation Tools

This study utilizes various free and open-source tools for model development, data processing, and evaluation. Each tool contributes to the efficient execution of the tasks, ensuring high-quality results throughout the study.

1. Python: Python is a high-level, general-purpose programming language that is widely used in data science and machine learning due to its simplicity and versatility. It offers extensive libraries and frameworks to implement machine learning models efficiently.

2. Virtual Environment: We will use Jupyter Notebook as the environment to conduct our experiment. Anaconda framework provides Jupyter Notebook along with other tools that provide assistance in conducting such experiments. We will create a virtual environment inside Anaconda and install all the required tools.

3. NumPy: NumPy is a powerful library for numerical computing in Python. It provides support for multidimensional arrays, along with a collection of mathematical functions to operate on these arrays, making it essential for scientific computing and data manipulation.

4. Matplotlib: Matplotlib is a popular plotting library for Python that allows you to create static, animated, and interactive visualizations. It is widely used for data visualization, offering various

chart types such as line plots, histograms, and scatter plots.

5. Scikit-learn: Scikit-learn is one of the most widely used libraries for machine learning in Python. It provides a range of simple and efficient tools for data mining and data analysis, including algorithms for prediction, regression, clustering, and more.

5. Pandas: Pandas is an open-source Python library used for data manipulation and analysis. It provides data structures like DataFrames for handling and analyzing structured data, along with powerful methods for data cleaning, merging, and transformation.

6. TensorFlow/Keras: TensorFlow is an open-source framework for machine learning, particularly deep learning, and developed by Google. Keras is a high-level API for building and training models in TensorFlow, making it easier to experiment with deep neural networks and complex architectures

Chapter 5

Results and Analysis

5.1 Introduction to Results

This section presents the evaluation results of the automated question Prediction and question generation systems developed for Bloom's Taxonomy. The primary objective is to assess the effectiveness of these systems in generating and predicting questions in alignment with the cognitive levels of Bloom's framework. The evaluation focuses on examining the ability of the models to produce contextually relevant and balanced assessments, with an emphasis on key performance metrics to validate their effectiveness and practical applicability in educational settings.

5.2 Performance Analysis of Distinct Models for Question Prediction

This section presents the performance analysis of various transformer-based models used for question prediction based on Bloom's Taxonomy. The evaluation focuses on the effectiveness of these models in predicting questions according to Bloom's cognitive levels. Performance is assessed using key metrics such as precision**,** recall, and F1-score. The results provide insights into how each model performs in predicting questions across the different cognitive levels and their ability to meet the objectives of the prediction task.

5.2.1 Question Prediction with BERT

The BERT model was fine-tuned for predicting questions based on Bloom's Taxonomy by leveraging its pre-trained embeddings. Training involved tokenizing the questions, extracting features, and fine-tuning the model over 30 epochs. Figure 5.1 presents the training and accuracy. The training accuracy also doubled throughout the 30 epochs, increasing from 0.32 to 0.9622. The validation accuracy experienced a steep rise in the first few epochs, reaching a consistent value of 0.9445. This consistent and parallel growth in training and validation performance is evidence that not only did this model learn, but it generalized well on out-of-sample data with very little overfitting.

Moreover, Figure 5.2 demonstrates the training and validation loss plot. First, the training loss began at 0.94 and continuously decreased due to the model's increasing knowledge of the data, and finally leveled off at a small value of approximately 0.1192. The validation loss also started at 0.76 and shows a downward trend, finishing at 0.1022.

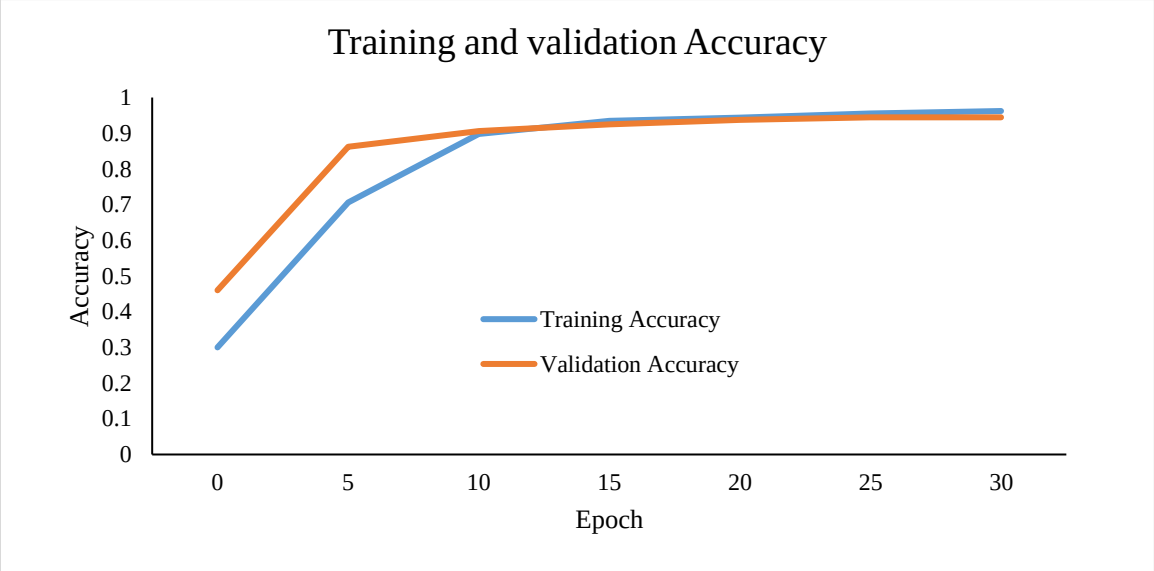

Figure 5.1 BERT MODEL Training and Validation Accuracy

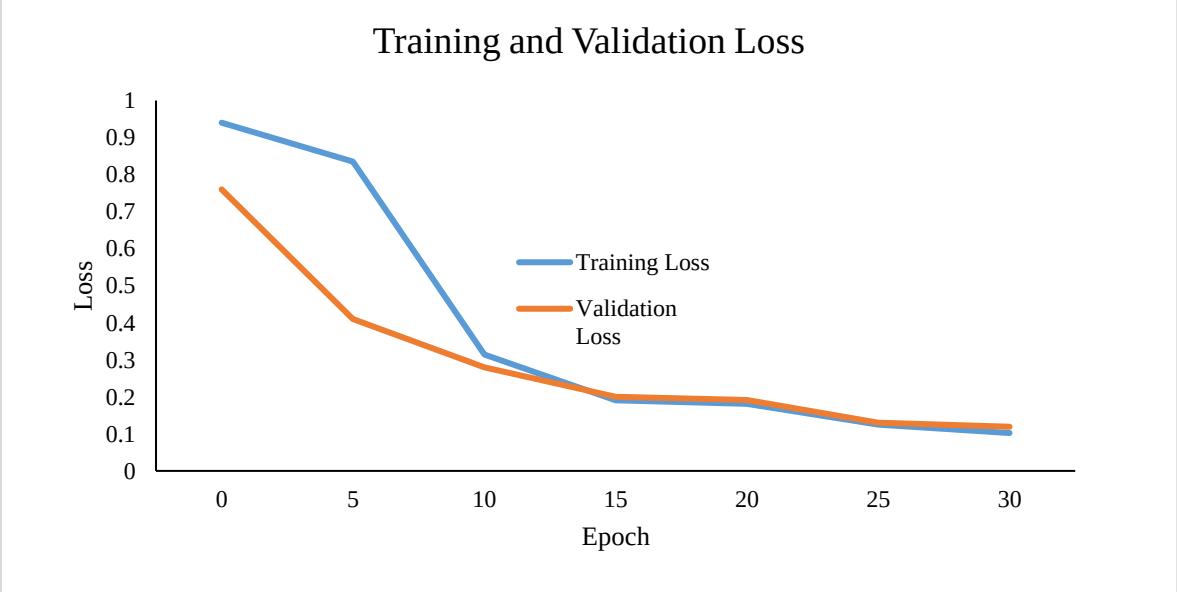

Figure 5.2 BERT MODEL Training and Validation Loss

In the Confusion Matrix Figure (5.3), it is evident that the question predictor performs well overall. Demonstrating a good number of correct predictions across all Bloom's Taxonomy levels, as reflected by the strong values along the diagonal elements in the matrix. However, there are still some mispredictions between Remember (BT1) and Understand (BT2), as well as between Remember (BT1) and Analyze (BT4), and between Remember (BT1) and Analyze (BT3). BT1, BT2, and BT3 show the maximum number of mispredictions.

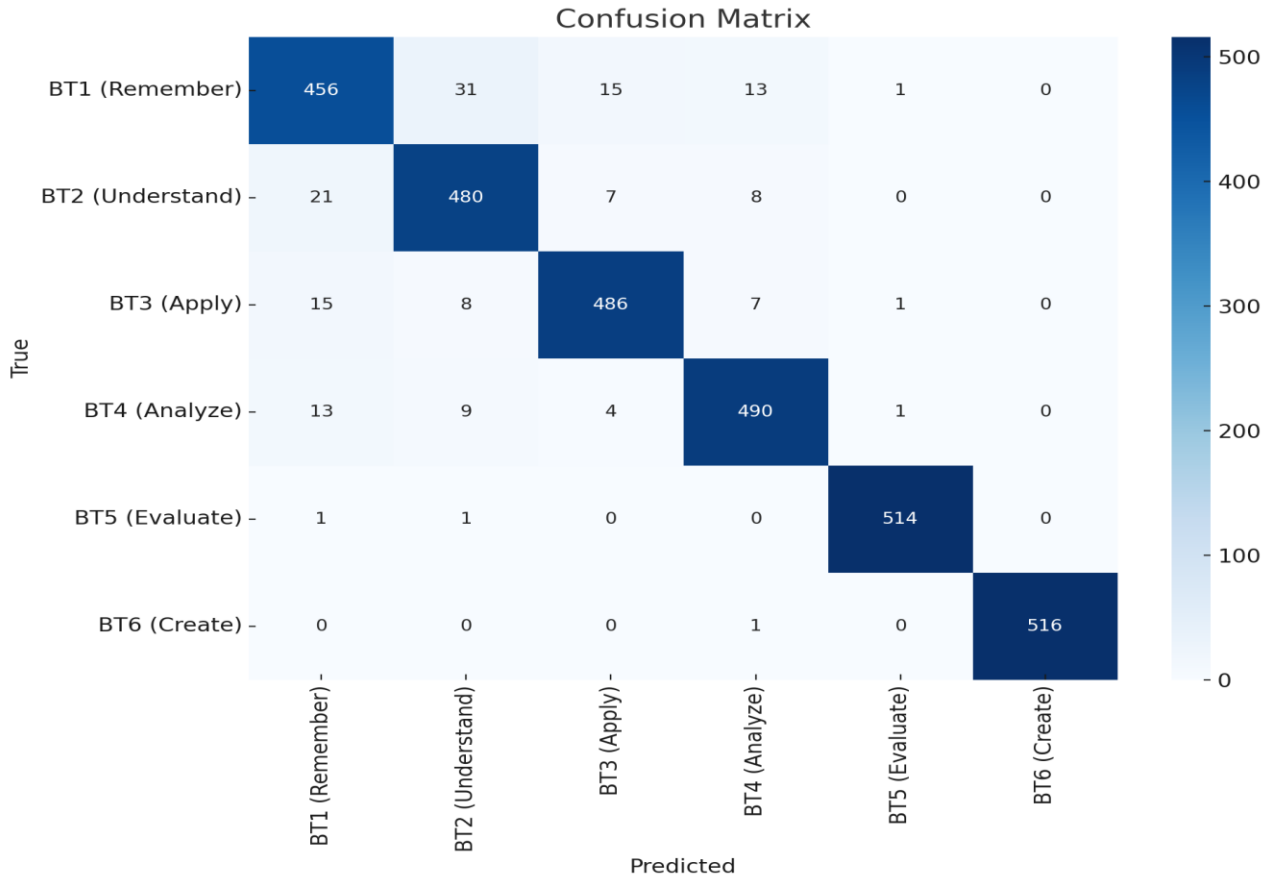

Figure 5.3 Confusion Matrix for BERT

Table 5.1 provides an overall summary of the results obtained from the BERT model, achieving an overall accuracy of 94.45%, with a precision of 94.00% and a recall of 93.67%, indicating its effective prediction of questions across Bloom's cognitive levels. Table 4.2 presents the prediction report for the BERT model at each Cognitive level. BT5 demonstrated excellent performance, with a precision of 1.00, a recall of 0.99, and an F1-score of 1.00, highlighting the model's strength in higher-order thinking skills.

Table 5.1: Performance matrices Result.

	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
BERT	96.22%	94.45%	94%	93.67%	93.83%

Table 5.2: Question Prediction Report for BERT

Cognitive Level	Precision	Recall	F1-Score
BT1 (Remember)	0.91	0.83	0.87
BT2 (Understand)	0.90	0.91	0.91
BT3 (Apply)	0.92	0.95	0.93
BT4 (Analyze)	0.91	0.95	0.93
BT5 (Evaluate)	1.00	0.99	1.00
BT6 (Create)	1.00	0.99	0.99

5.2.2 Question Prediction with Distill BERT

The Distill BERT model, a smaller and more efficient variant of BERT, was fine-tuned on a dataset of questions labeled according to Bloom's Taxonomy. In Figure 5.4, we can see that the training started at 0.32 and finished at 0.9622, achieving an Accuracy of 0.9445 after 30 epochs. However, in Figure 5.5, the validation loss increased in some cases, starting from 0.50 and reaching 0.095 in the final epoch, reflecting that the model demonstrated good performance in identifying both lower- and higher-order cognitive levels.

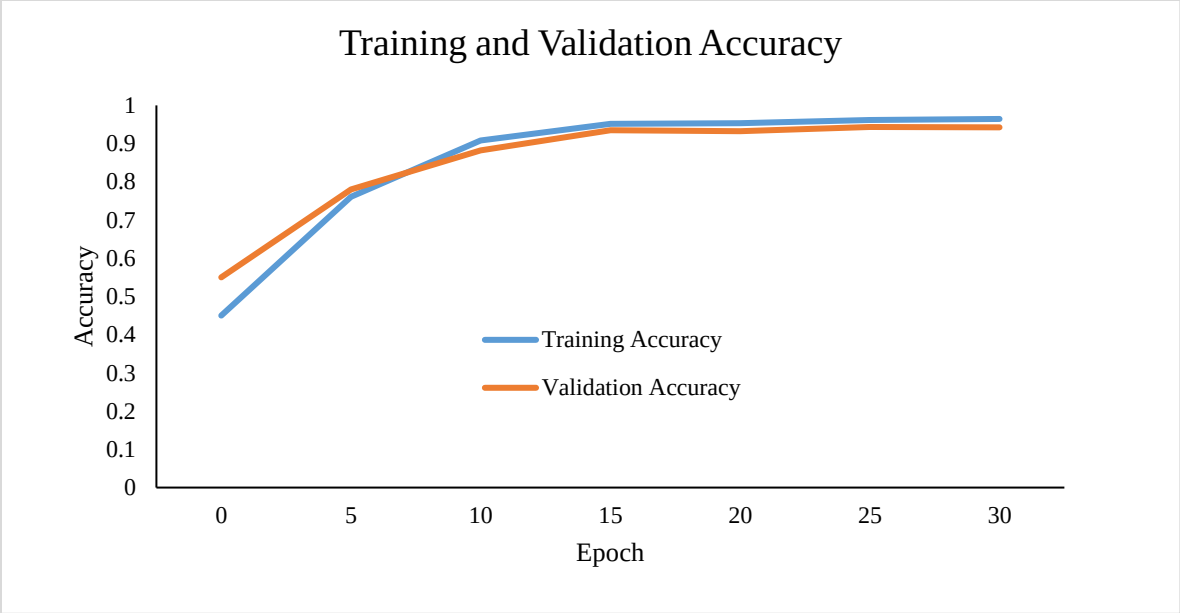

Figure 5.4 Distill BERT MODEL Training and Validation Accuracy

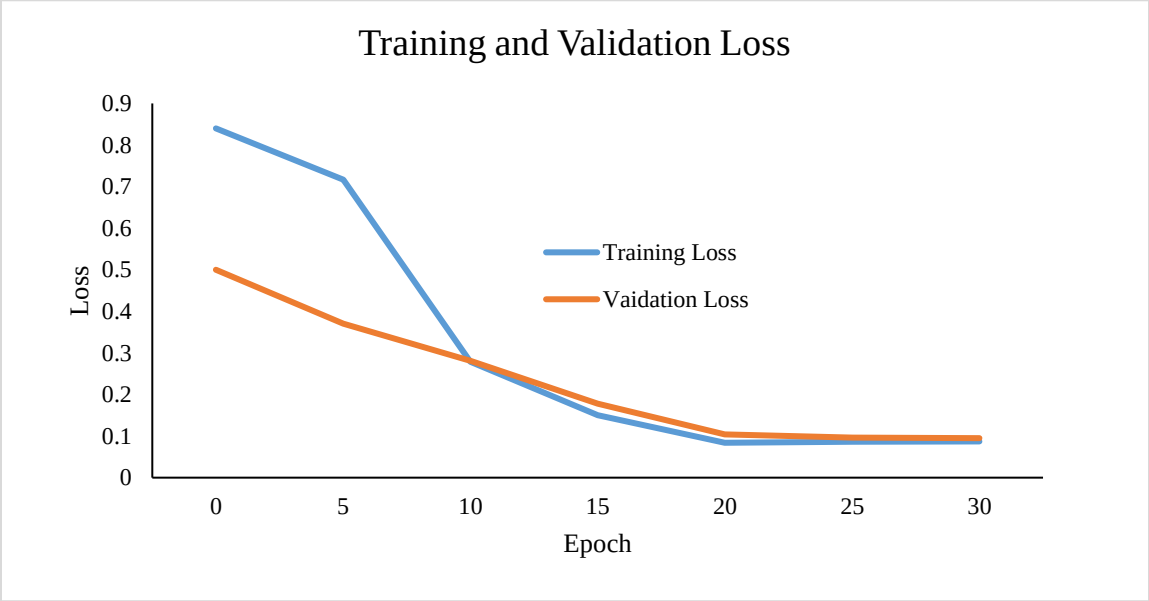

Figure 5.5 Distill BERT MODEL Training and Validation Loss

In the Confusion Matrix Figure (5.6), it is evident that the question predictor performs well overall. Demonstrating a good number of correct predictions across all Bloom's Taxonomy levels, as reflected by the strong values along the diagonal elements in the matrix. However, there are still

some misclassifications between Remember (BT1) and Understand (BT2), as well as between Remember (BT1) and Analyze (BT4), and between Remember (BT1) and Analyze (BT3). And Remember (BT1) and Understand (BT2) show higher mispredictions.

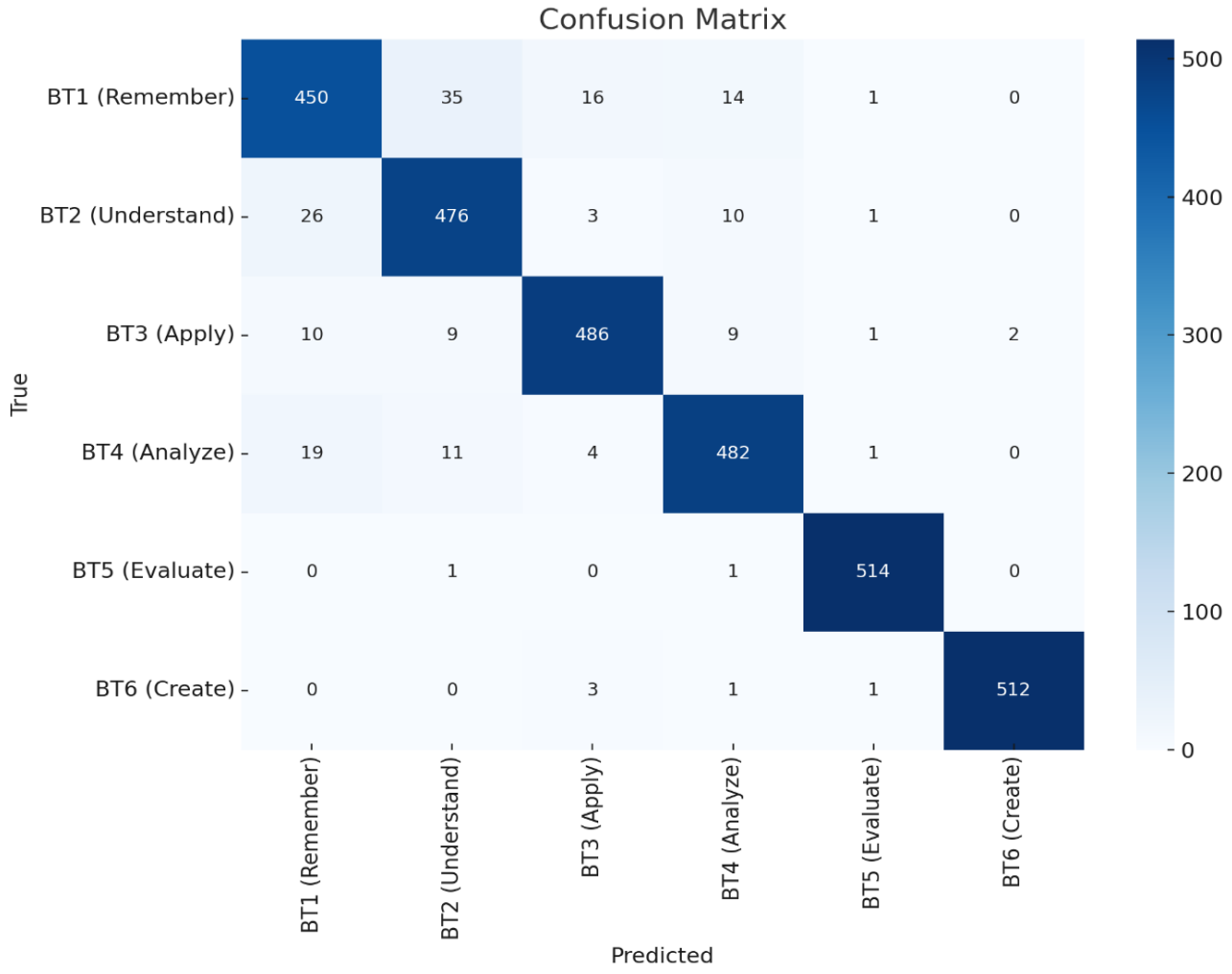

Figure 5.6 Confusion Matrix for DistilBERT

Table 5.3 gives an overall summary of the results obtained from the DistilBERT model, which achieved an accuracy of 94.21%, with macro average precision of 94% and recall of 93%, demonstrating solid performance across various Bloom levels. Table 5.4 presents the Question Prediction report for the DistilBERT model at each Cognitive level, demonstrating its performance in BT5, with a precision of 1.00, a recall of 0.99, and an F1-score of 0.99.

Table 5.3: Performance matrices Result.

	Training Accuracy	**Testing Accuracy**	**Precision**	**Recall**	**F1-Score**
DistilBERT	96.45%	94.21%	94.08%	93.43%	94.005

Table 5.4: Question Prediction Report DistilBERT

Cognitive Level	**Precision**	**Recall**	**F1-Score**
BT1 (Remember)	0.89	0.87	0.88
BT2 (Understand)	0.89	0.92	0.91
BT3 (Apply)	0.95	0.94	0.94
BT4 (Analyze)	0.93	0.93	0.93
BT5 (Evaluate)	1.00	0.99	0.99
BT6 (Create)	0.99	1.00	0.99

5.2.3 Question Prediction with ROBERTa

The RoBERTa model, a robustly optimized version of BERT, was fine-tuned for predicting questions based on Bloom's Taxonomy. In Figure 5.7, we can see that the training started at 0:33 am and finished at 0:9584, achieving an Accuracy of 0.9487 after 30 epochs. However, Figure 5.8 shows that the validation loss increased in some cases, starting from 0.62 and reaching 0.1085 in the final epoch, reflecting that the model demonstrated strong performance in identifying both lower- and higher-order cognitive levels.

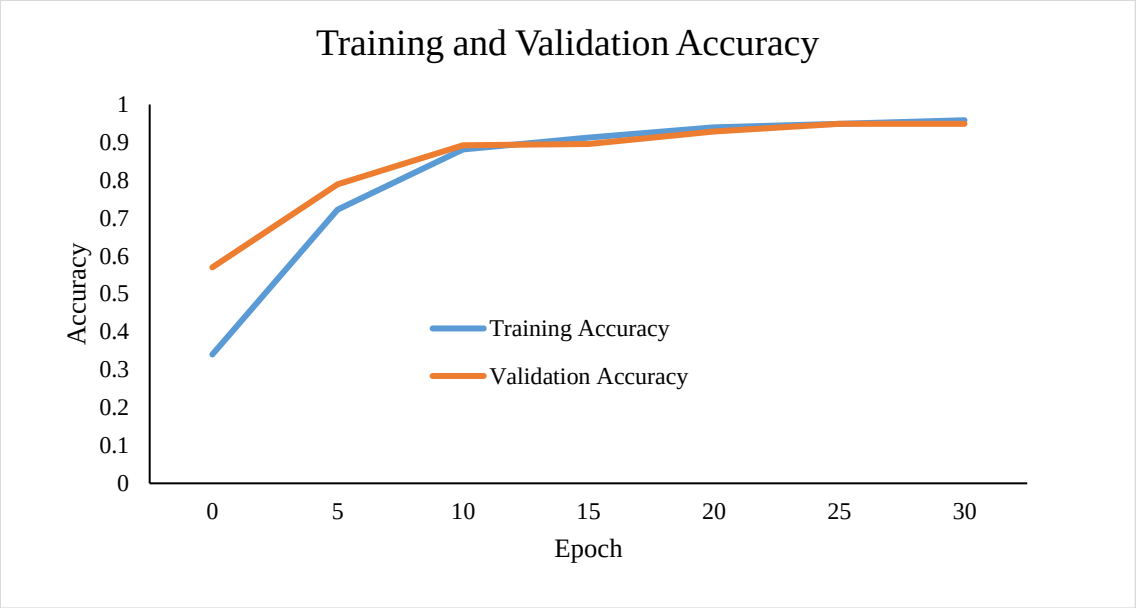

Figure 5.7 ROBERTa MODEL Training and Validation Accuracy

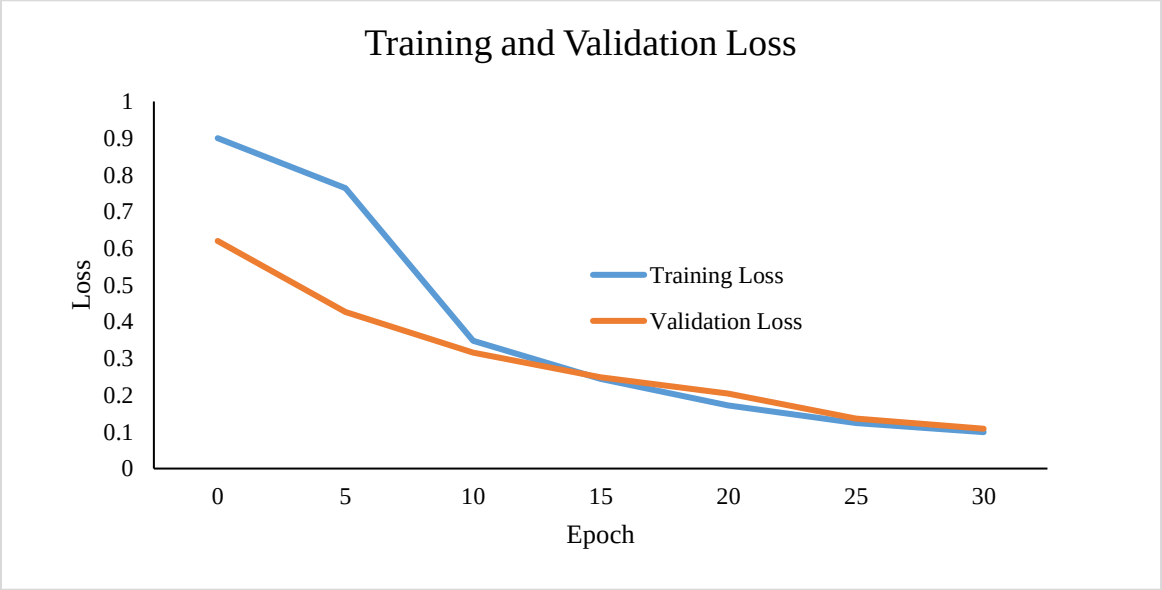

Figure 5.8 ROBERTa MODEL Training and Validation Loss

In the Confusion Matrix Figure (5.9), it is evident that the Question Predictor performs well overall. Demonstrating a good number of correct predictions across all Bloom's Taxonomy levels, as reflected by the strong values along the diagonal elements in the matrix. However, there are still

some mispredictions between Remember (BT1) and Understand (BT2), as well as between Remember (BT1) and Analyze (BT4), and between Remember (BT1) and Analyze (BT3). Moreover, Remember (BT1) and Understand (BT2) exhibit higher misprediction rates.

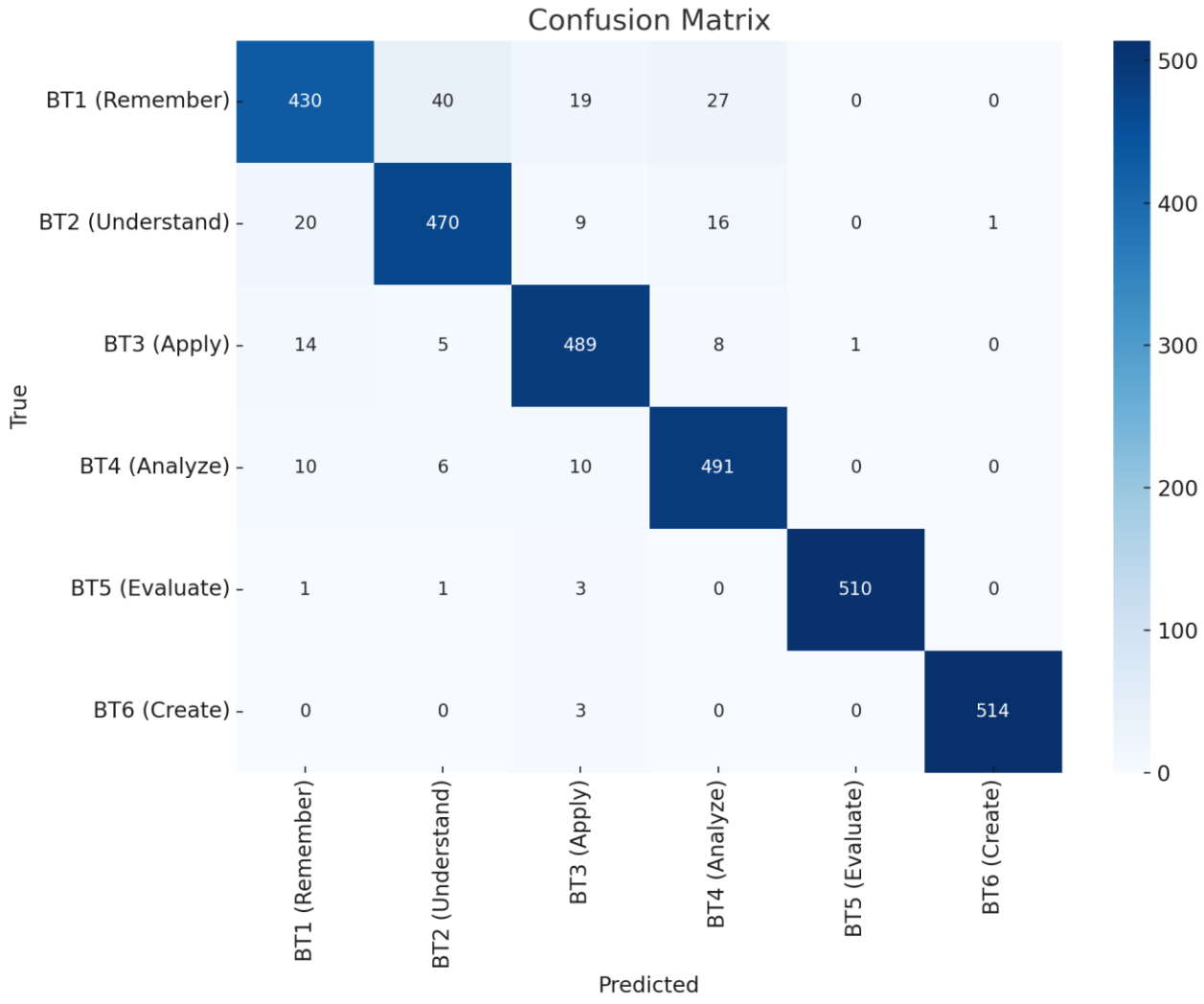

Figure 5.9 Confusion Matrix for RoBERTa

The table 5.5 summary of the result obtained from the Roberta model achieved an overall accuracy of 94.87%, with a precision of 94.32% and a recall of 93.34%. Table 5.6 shows question prediction report and we observe that It performed exceptionally well on higher-order cognitive levels, such as BT5 and BT6, with high precision and recall (0.99 precision, 1.00 recall and 1.00 F-1 Score for BT5, 0.99 precision, 1.00 recall and 1.00 F-1 Score for BT6), showcasing its ability to classify advanced thinking skills effectively.

Table 5.5: Performance matrices Result.

	Training Accuracy	**Testing Accuracy**	**Precision**	**Recall**	**F1-Score**
RoBERTa	95.84%	94.87%	94.32%	93.34%	94.28%

Table 5.6: Question Prediction Report for RoBERTa

Cognitive Level	**Precision**	**Recall**	**F1-Score**
BT1 (Remember)	0.90	0.88	0.89
BT2 (Understand)	0.91	0.93	0.92
BT3 (Apply)	0.95	0.94	0.94
BT4 (Analyze)	0.94	0.95	0.95
BT5 (Evaluate)	0.99	1.00	1,00
BT6 (Create)	0.99	1.00	1.00

5.2.4 Question Prediction with LSTM

The LSTM model was trained to predict questions according to Bloom's Taxonomy by leveraging pre-trained embeddings. Training involved tokenizing the questions, extracting features, and optimizing the model over 30 epochs. From Figure 5.10, we observe that after the final epoch, the model achieves 0.9538 training accuracy and 0.8909 validation accuracy**.** In Figure 5.11, we can see that the validation loss initially increases, starting from 0.8, and eventually decreases to 0.2118 by the final epoch. The model overall demonstrated moderate effectiveness in distinguishing between both lower- and higher-order cognitive levels.

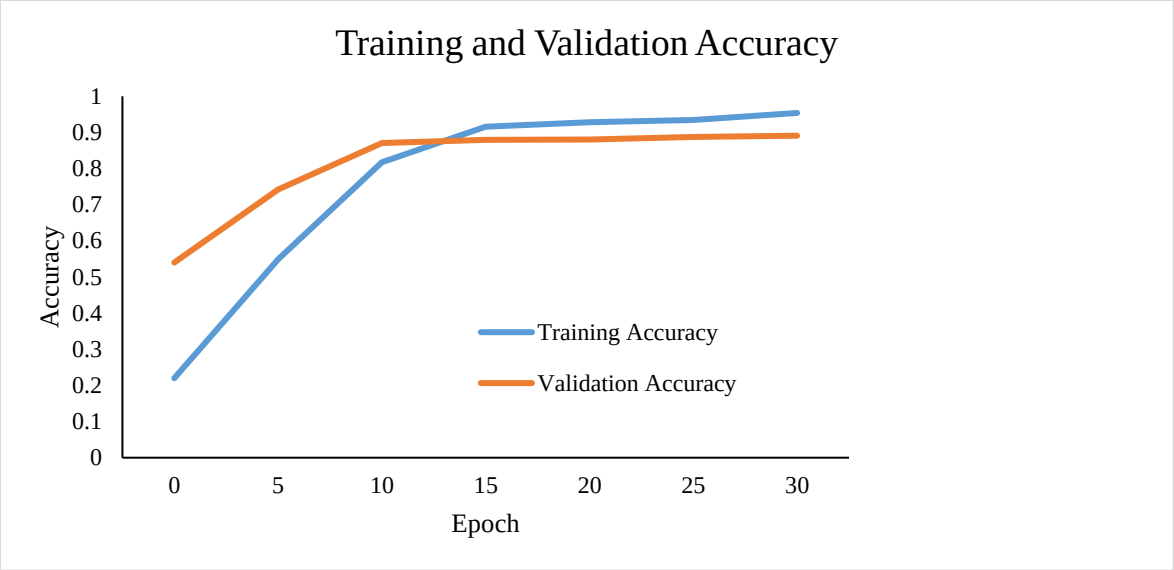

Figure 5.10 LSTM MODEL Training and Validation Accuracy

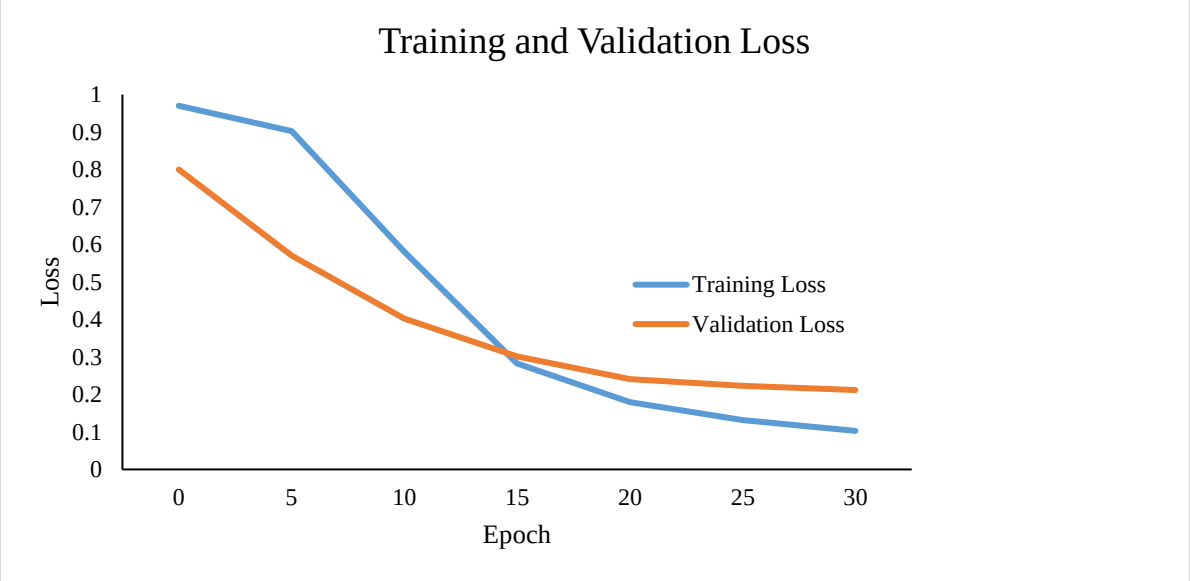

Figure 5.11 LSTM Training and Validation Loss

As shown in the Confusion Matrix (Fig. 5.12), the model exhibited moderate prediction accuracy across all categories. Some degree of misprediction was observed between 'Remember' and 'Understand', as well as between 'Apply' and 'Analyze', indicating overlapping features among these cognitive levels.

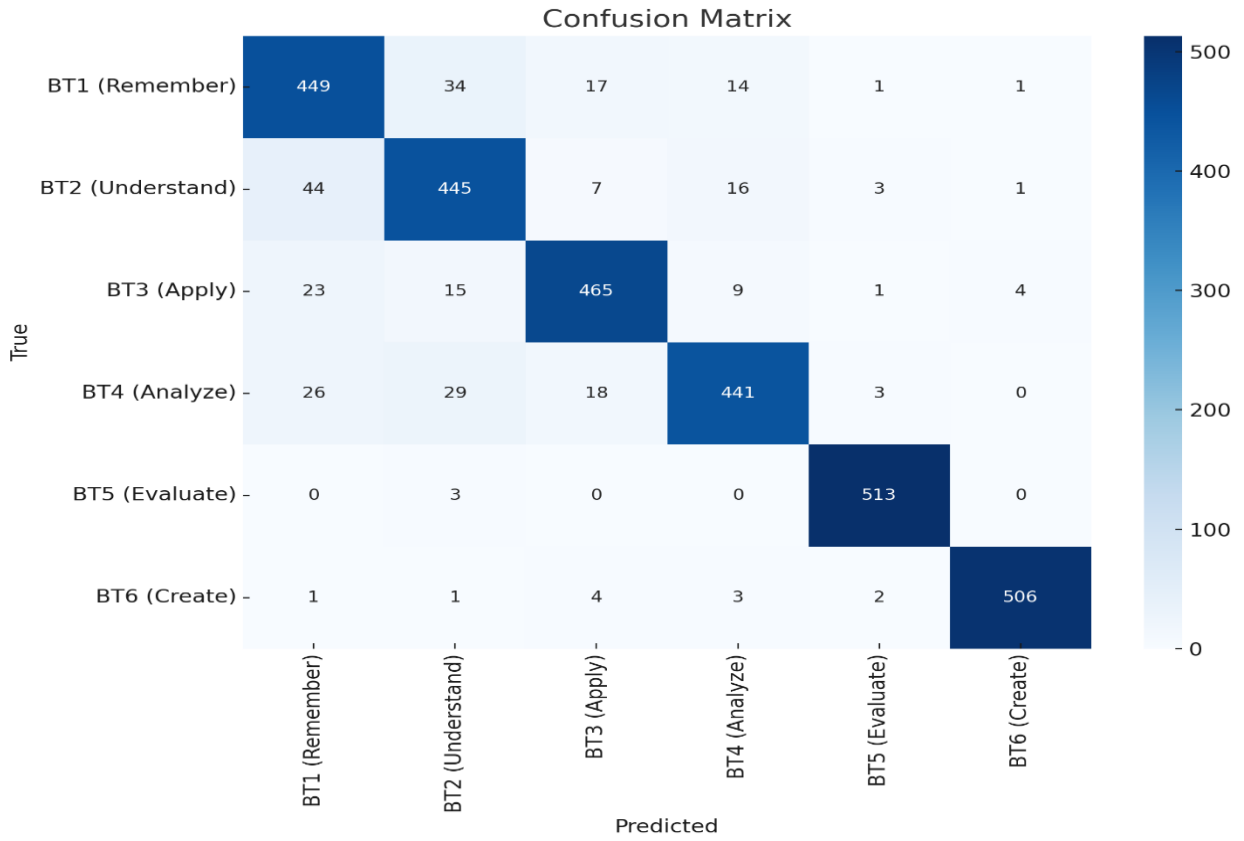

Figure 5.12 Confusion Matrix for LSTM

Table 5.7 presents a summary of the results obtained from the LSTM model, which achieved an overall accuracy of 89.09%, with a macro-average precision of 88% and recall of 87%**,** indicating its relatively moderate performance across Bloom's cognitive levels. Table 5.8 presents the prediction report, which shows that the model performed exceptionally well in predicting BT6 (Create) level questions, achieving a precision of 0.99, a recall of 0.98**,** and an F1-score of 0.98**,** demonstrating its strength in identifying higher-order thinking skills. However, performance was less accurate for lower-order levels (e.g., "Remember" and "Understand").

Table 5.7: Performance matrices Result.

	Training Accuracy	**Testing Accuracy**	**Precision**	**Recall**	**F1-Score**
LSTM	95.38%	89.09%	88.56%	87.545	87.42%

Table 5.8: Question Prediction Report for LSTM

Cognitive Level	Precision	Recall	F1-Score
BT1 (Remember)	0.83	0.87	0.85
BT2 (Understand)	0.84	86	0.85
BT3 (Apply)	0.89	0.88	0.87
BT4 (Analyze)	0.89	0.87	0.88
BT5 (Evaluate)	0.98	0.98	0.98
BT6 (Create)	0.99	0.98	0.98

5.2.5 Question Prediction with SVM

The SVM model was trained to predict questions according to Bloom's Taxonomy by utilizing pre-extracted embeddings. Table 5.9 presents the overall summary of the results, showing that the SVM model achieved an overall accuracy of 87.75% with a precision of 87% and a recall of 86%**,** indicating relatively moderate performance across Bloom's cognitive levels.

Table 5.9: Performance matrices Result.

	Training Accuracy	**Testing Accuracy**	**Precision**	**Recall**	**F1-Score**
SVM	90.23%	87.75%	87.54%	86.42%	86.43%

Table 5.10 presents the Question prediction summary report, which shows particularly strong results for BT5 (Evaluate) and BT6 (Create) level questions, achieving a precision of 0.99, a recall

of 0.99**,** and an F1-score of 0.99. And BT1, BT2, BT3.BT4 shows lower performance among all classes. BT1 has a lower precision of 0.80 and a recall of 0.81. After that, BT2 has a precision of 0.86, with recall and F1 scores of 0.85 and 0.84, respectively. BT3 and BT4 also have lower precision, recall, and F1 scores. For BT3, precision is 0.88, recall is 0.87, and the F1 score is 0.86. For BT4, precision is 0.87, recall is 0.88, and the F1 score is 0.87.

Table 5.10: Question Prediction Report for SVM

Cognitive Level	Precision	Recall	F1-Score
BT1 (Remember)	0.80	0.81	0.81
BT2 (Understand)	0.86	0.85	0.84
BT3 (Apply)	0.88	0.87	0.86
BT4 (Analyze)	0.87	0.88	0.87
BT5 (Evaluate)	0.99	0.99	0.99
BT6 (Create)	0.99	0.99	0.99

As illustrated in the Confusion Matrix (Figure 5.13), the model demonstrated moderate prediction performance across all cognitive categories. However, some misclassification was observed between BT1 (Remember) and BT2 (Understand).BT3 (Apply).BT4 (Analyze). Remember, understand, and analyze show maximum misclassification. 41 BT2 question were incorrectly classified as BT1 and 28 BT1 questions were misclassified with BT3.Similarly, 27 BT1 questions were predicted as BT4, and 24 BT3 questions were misclassified as BT1. These errors suggest that overlapping characteristics between levels, especially among lower-order skills, contribute to confusion. However, higher-order thinking level shows good performance with minimal misclassification. However, the overall performance of the model is not satisfactory because lower-order thinking levels exhibit poor performance and often misclassify with each other.

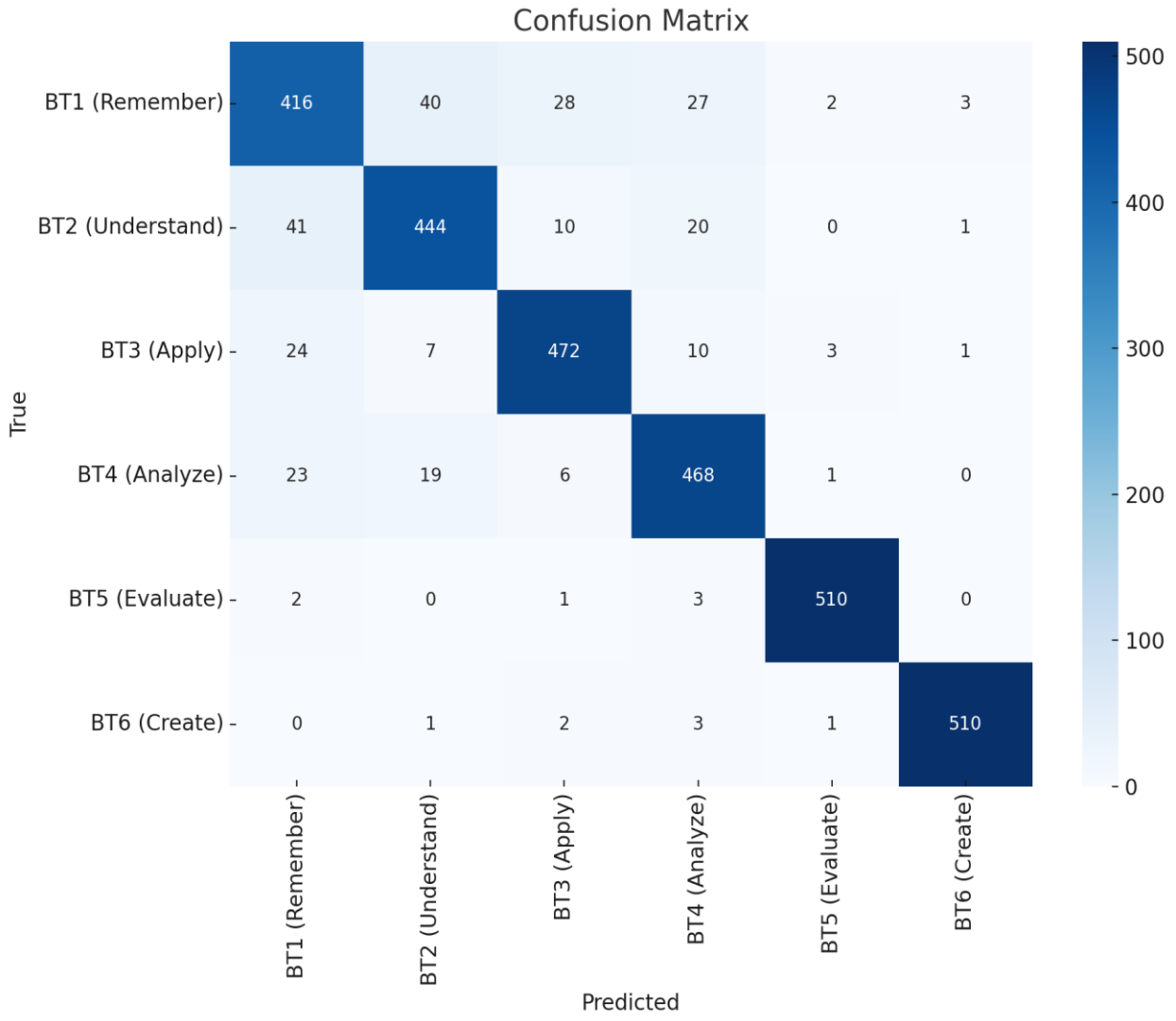

Figure 5.13 Confusion Matrix for SVM

5.2.6 Question Prediction with Hybrid Model

The Hybrid model is trained on a dataset of questions labeled according to Bloom's cognitive levels, with preprocessing steps including tokenization, feature extraction, and data augmentation to enhance model robustness. In Figure 5.14, we can see that the training started at 0.20 and finished at 0.9, achieving an accuracy of 0.9639 after 30 epochs. Hyperparameters used in training included a learning rate of 1e-5 for ROBERTa, 1e-3 for LSTM, and a batch size of 32. The training process consisted of 10 epochs for fine-tuning ROBERTa and 20 epochs for LSTM, CNN, and TF-IDF. In Figure 5.15, we observe that the training loss started at 0.7621 and steadily decreased,

with the final validation loss starting from 0.45 and reaching 0.079 in the final epoch, indicating strong model generalization.

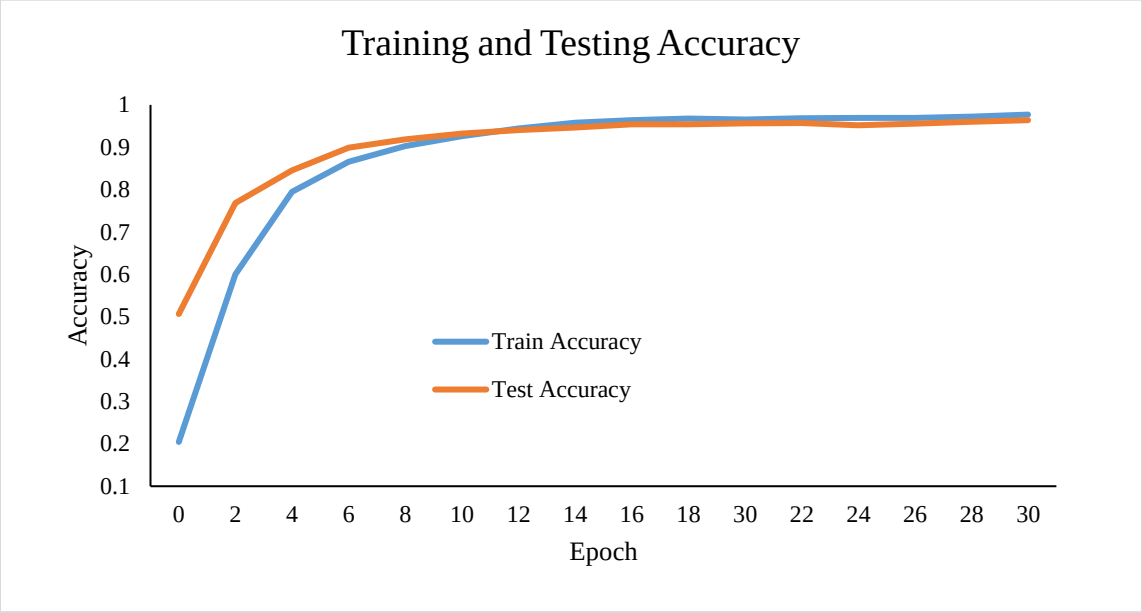

Figure 5.14 Hybrid Model learning behavior with different number of epochs.

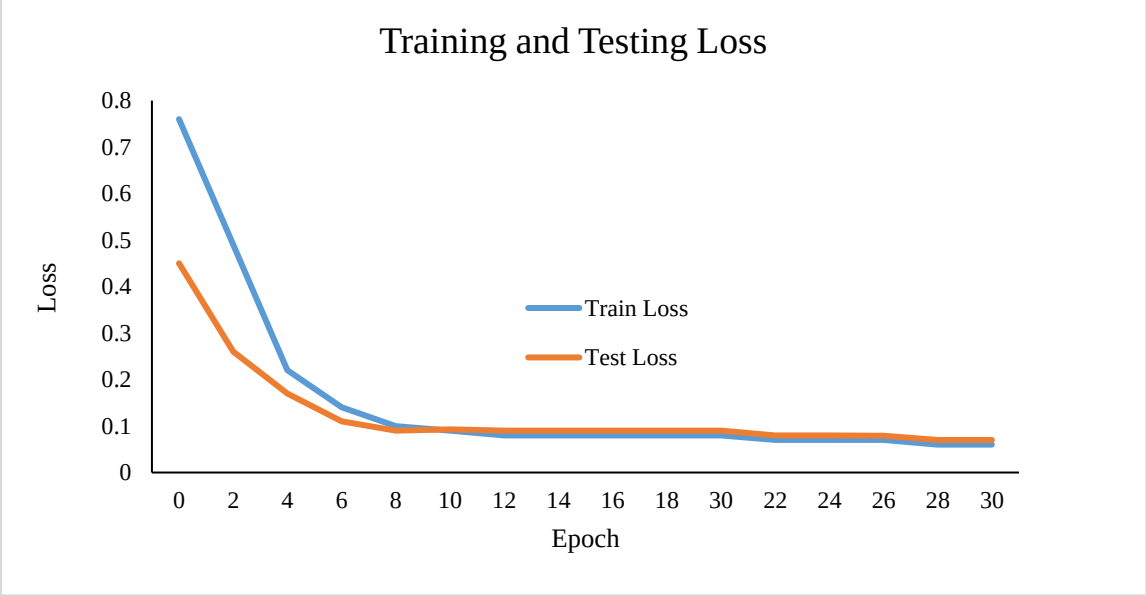

Figure 5.15 Hybrid Model Loss with different number of epochs

In the Confusion Matrix Figure (5.16), it is evident that the classifier performs well overall. Demonstrating a high number of correct predictions across all Bloom's Taxonomy levels as reflected by the strong values along the diagonal elements in the matrix. However, there are still some misclassifications between Remember (BT1) and Understand (BT2), as well as between Remember (BT1) and Analyze (BT4).

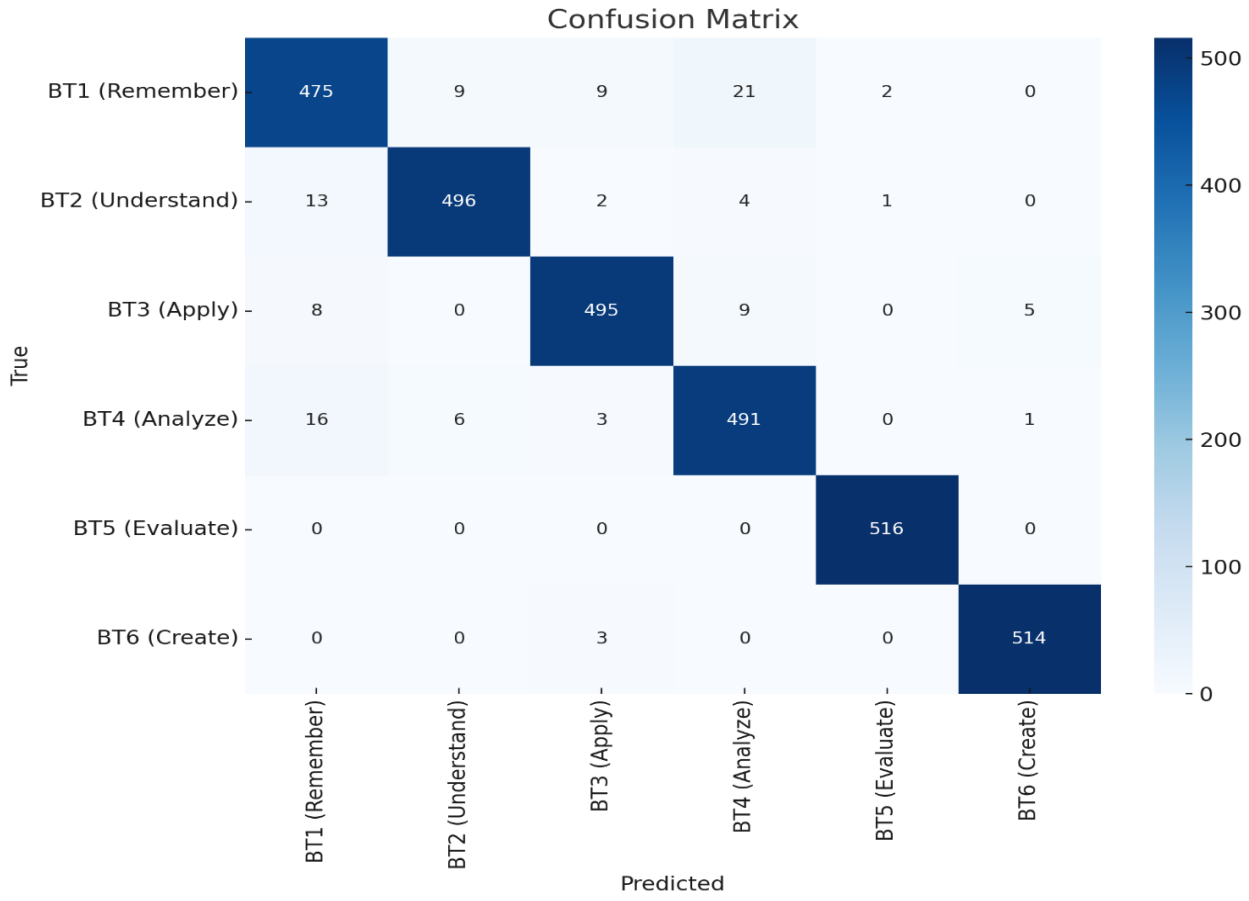

Figure 5.16 Confusion Matrix for Hybrid model

Table 5.11 gives an overall summary of the results obtained from the model, which achieved an accuracy of 96.39%, a precision of 96.21%, a recall of 95.81%, and an F1-score of 96.13%, demonstrating its effectiveness in predicting questions across all cognitive levels of Bloom's Taxonomy, especially for higher-order thinking skills. Moreover, Table 5.12 presents the Question Prediction report for the Hybrid Model at each Cognitive level, where BT5 (Evaluate) and BT6 (Create) demonstrate higher performance.

Table 5.11: Performance matrices Result.

	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score
Hybrid	97.77%	96.39%	96.21%	95.81%	96.13%

Table 5.12: Question Prediction Report for Hybrid Model

Cognitive Level	Precision	Recall	F1-Score
BT1 (Remember)	0.93	0.92	0.92
BT2 (Understand)	0.97	0.96	0.97
BT3 (Apply)	0.97	0.96	0.96
BT4 (Analyze)	0.94	0.95	0.94
BT5 (Evaluate)	0.99	0.99	0.99
BT6 (Create)	0.99	1.00	1.00

5.2.7 New Sample Testing Result for Question Prediction

New samples of various questions are tested to predict their category according to Bloom's taxonomy of cognitive levels. The Output tables are shown below:

Table 5.13 Question Prediction Sample Testing Output by proposed model (Hybrid)

SL NO	Questions	Original Level	Predicted Level	Correct/Incorrect	Accuracy	Error
1	What is PCG?	BT1	BT1	Correct		
2	Define ECG.	BT1	BT1	Correct		

3	Tell about Half-cell potential.	BT1	BT1	Correct		
4	List out the electrodes used for recording EMG and ECG	BT1	BT1	Correct		
5	What should be the characteristics of a bio-potential amplifier? Identify the origin of bio potential	BT1	BT1	Correct		
6	What do you under-stand by electrophore-sis?	BT1	BT1	Correct		
7	Discuss latency as re-lated to EMG	BT2	BT2	Correct	**Accuracy = 92.5%**	**Error = 7.5%**
8	Summarize the genera-tion of PCG signals and discuss the meas-urement of PCG.	BT2	BT2	Correct		
9	Describe the methods of measurement of car-diac output.	BT2	BT2	Correct		
10	Give the difference be-tween an external and an internal defibrilla-tor.	BT2	BT1	Correct		
11	Classify shortwave and microwave diathermy	BT2	BT1	Correct		
12	Illustrate the im-portance of the biolog-ical amplifier	BT3	BT3	Correct		

13	Calculate the stroke volume in milliliters if the cardiac output is 5.2 liters/minute and the heart rate is 76 beats/minute.	BT3	BT3	Correct		
14	Compare the signal characteristics of ECG and PCG	BT4	BT2	Incorrect		
15	Explain the term phonocardiogram	BT2	BT2	Correct		
16	Point out the components of blood.	BT4	BT4	Correct		
17	Analyse the measurement of the blood pH value	BT4	BT4	Correct		
18	Explain a single-channel ECG bio telemetry system with a neat diagram.	BT2	BT2	Correct		
19	Elaborate on the multiple channel telemetry systems with neat diagrams	BT4	BT4	Correct		
20	Assess the important bands of frequencies in the EEG and their importance	BT5	BT5	Correct		
21	Justify the use of the Einthoven triangle	BT5	BT5	Correct		

22	Summarize the different types of pacemakers	BT2	BT2	Correct		
23	Justify the use of the Einthoven triangle	BT5	BT5	Correct		
24	Generalize the problems associated with the implant telemetry circuits? Explain the subcarrier biotelemetry.	BT3	BT1	Incorrect		
25	Formulate the vital capacity of a person who has a total lung capacity of 5.95 liters, if the volume of the lung is filled with air.	BT6	BT3	Incorrect		
26	Compose the Nernst equation	BT6	BT6	Correct		
27	Develop the EEG waveform in detail and its signal frequency bands.	BT6	BT6	Correct		
28	Name any 4 physical principles based on which blood flow meters are constructed.	BT1	BT1	Correct		
29	Summarize the classification of Pacing modes	BT2	BT2	Correct		

30	Differentiate the types of diathermy	BT2	BT2	Correct		
31	Illustrate the block diagram of short wave and ultrasonic diathermy and explain.	BT3	BT3	Correct		
32	Classify the plethysmographs from plethysmography.	BT2	BT2	Correct		
33	Explain the safety precaution to be taken while handling radio isotopes	BT2	BT2	Correct		
34	Discuss the application of Bio-Telemetry	BT2	BT2	Correct		
35	Summarize the classification of Pacing modes	BT2	BT2	Correct		
36	Develop a new telemetry system for ICU patients using IoT and AI technologies.	BT6	BT6	Correct		
37	Design a block diagram for a wearable ECG monitor with real-time cloud integration.	BT6	BT6	Correct		
38	Illustrate the block diagram of short-wave and ultrasonic diathermy and explain it.	BT3	BT3	Correct		

39	Critique the effectiveness of different imaging modalities (CT vs. MRI) for soft tissue analysis..	BT5	BT5	Correct		
40	Analyse the measurement of blood pH value.	BT4	BT4	Correct		

5.3 Comparison of Question Prediction Models

The performance of six models—Hybrid, BERT, DistilBERT, RoBERTa, LSTM, and SVM—was evaluated for predicting questions according to Bloom's Taxonomy, using metrics including training and testing accuracy, precision, recall, and F1-score.

We can visualize from Figures 5.17 and 5.18 that the Hybrid model outperformed all others, with the highest training and accuracy (97.77%) and testing accuracy (96.39%). Transformer-based models, including BERT, DistilBERT, and RoBERTa, all demonstrated strong performance, with testing accuracies above 94%. Among them, RoBERTa shows the highest testing accuracy (94.87%). Moreover, LSTM demonstrated moderate performance with an accuracy of 89.09%, while SVM showed the lowest performance with an accuracy of 87.75%.

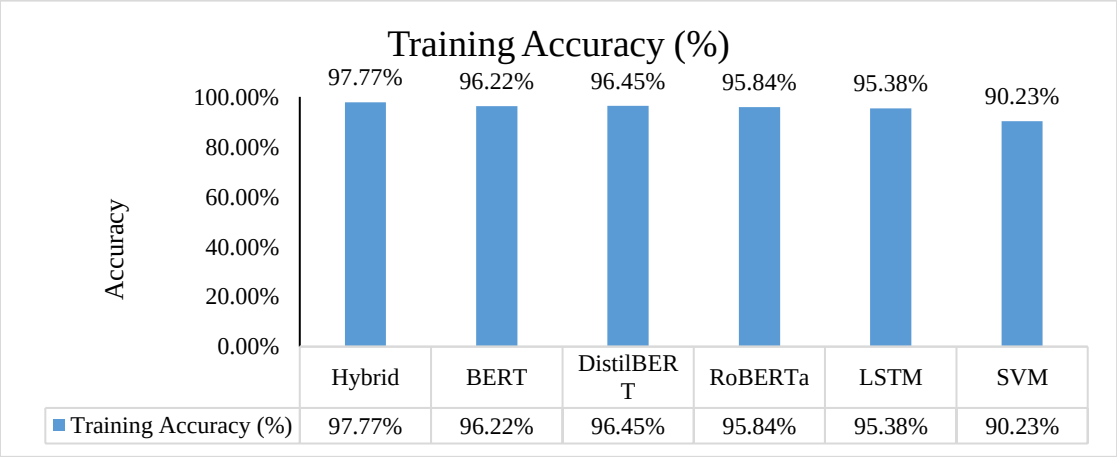

Figure 5.17: Question Prediction Model Training Comparison

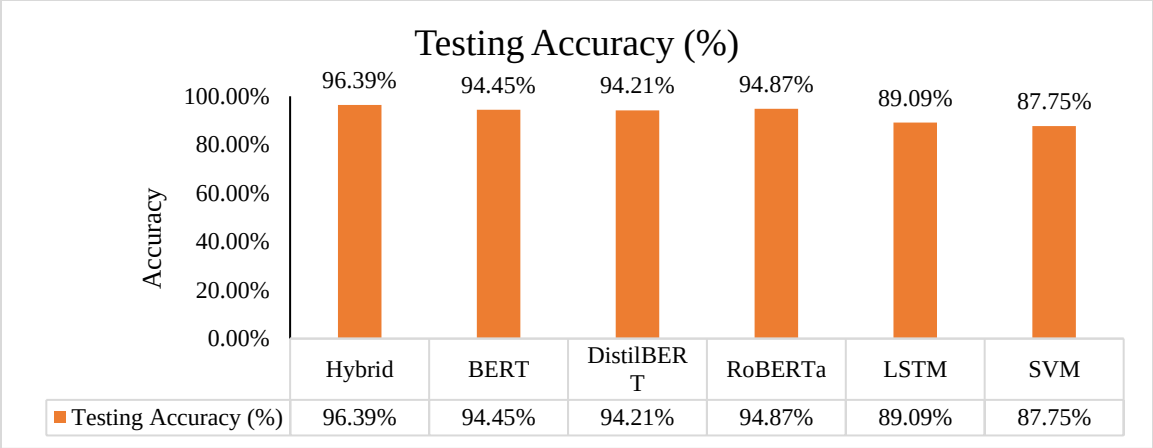

Figure 5.18: Question Prediction Model Testing Comparison

We can observe from Figures 5.19, 5.20, and 5.21 that the Hybrid model also performed well in terms of precision (96.21%), recall (95.81%), and F1-score (96.13%), indicating strong generalization and balanced classification across cognitive levels among its transfer models. DistilBERT slightly outperformed BERT in precision (94.08% vs. 94.00%) and F1-score (94.005% vs. 93.83%), despite being a lighter model. Moreover, RoBERTa shows strong performance with a strong F1-score (94.28%). The LSTM model showed relatively lower performance compared to transformer models. With precision, recall, and F1-score in the 87 to 88 percent range. Suggests it was moderately effective but less capable of capturing the semantic nuances across Bloom's cognitive levels compared to transformer architectures. SVM has lower performance with a precision of 87.54%, a recall of 86.42%, and an F1-score of 86.43%.

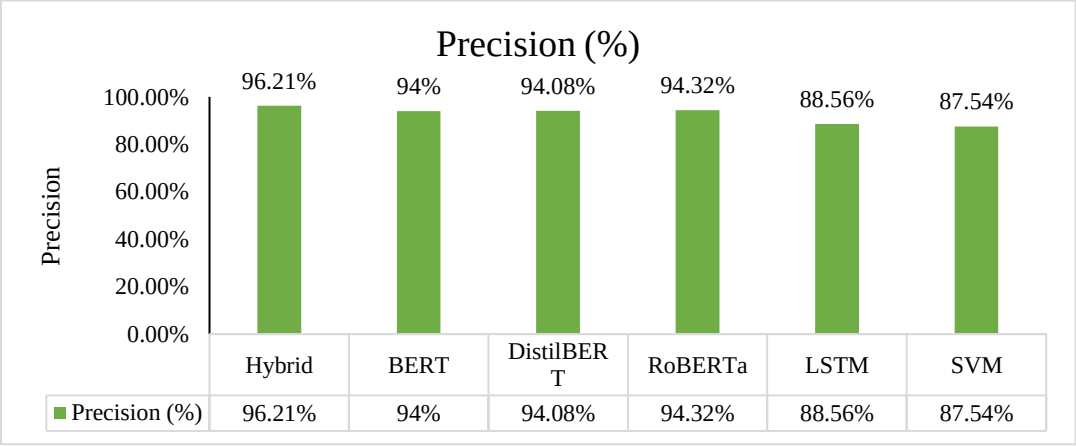

Figure 5.19: Question Prediction Model Precision Comparison

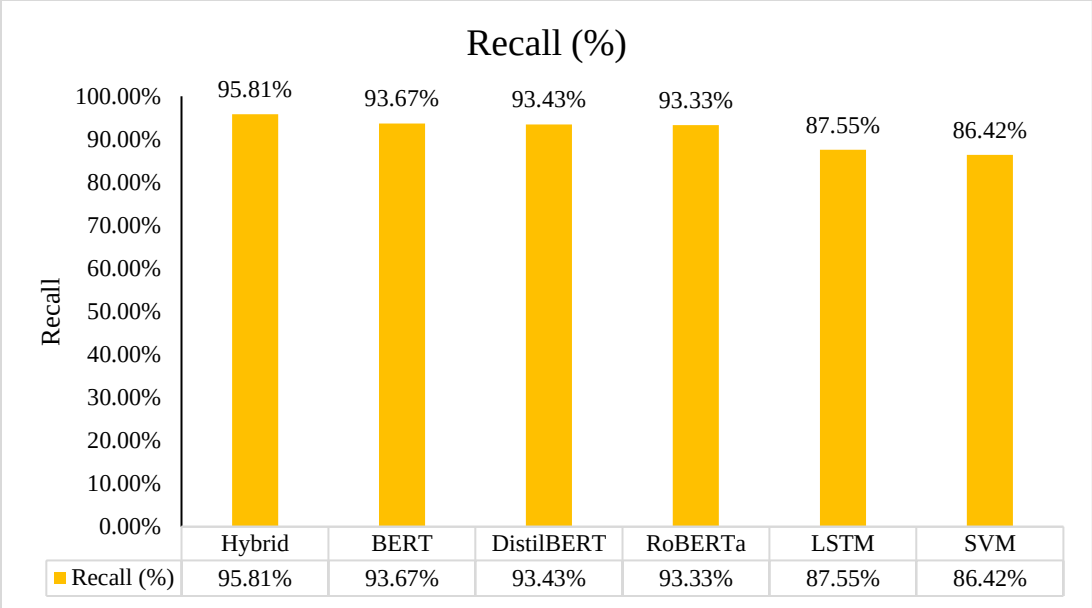

Figure 5.20: Question Prediction Model Recall Comparison

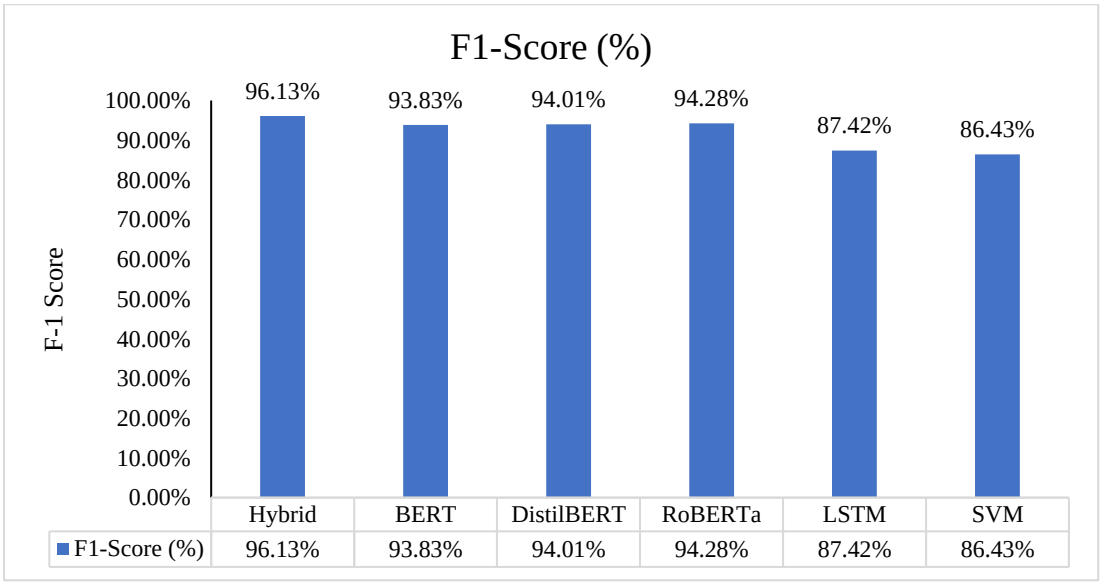

Figure 5.21: Question Prediction Model F-1 Score Comparison

Moreover, Table 5.14 presents a Summary of the question prediction model's performance, and among all models, the Hybrid model demonstrates a strong overall performance with accuracy of 96.39% and Precision Recall and F-1 Score is (96.21%, 95.81%, 96.13%) .

Table 5.14: Summary of Question Prediction Model Performance

Model	Training Accuracy (%)	Testing Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Hybrid	97.77	96.39	96.21	95.81	96.13
BERT	96.22	94.45	94	93.67	93.83
Distil-BERT	96.45	94.21	94.08	93.43	94.005
RoBERTa	95.84	94.87	94.32	93.325	94.28
LSTM	95.38	89.09	88.56	87.545	87.42
SVM	90.23	87.75	87.54	86.42	86.43

5.4 Performance Analysis of Distinct Models for Question Generation

This section evaluates the performance of various models in generating exam questions aligned with Bloom's Taxonomy. The evaluation is based on key metrics, including Training Accuracy, Validation Accuracy, BLEU Score, Expert Evaluation Score, and Diversity Score. These metrics provide a comprehensive evaluation of the model's capacity to generate high-quality, contextually relevant questions.

5.4.1 Question Generation with T5 Model

The T5 model was fine-tuned for the task of question generation, focusing on optimizing accuracy and contextual relevance. The model was trained for 40 epochs with a batch size of 32, a learning rate of 3e-5, and 1,000 warm-up steps to ensure stable convergence. In Figure 5.22, we can see that after completing training, the model began training from 0.25. At the final epoch, it achieved a training accuracy of 0.9689 and a testing accuracy of 0.9536, demonstrating its capacity to efec-tively learn from the data and generalize to new, unseen instances. From Figure 5.23, we observe

that at the end of the epoch, the model achieves a training loss of 0.0635 and a validation loss of 0.0792, indicating that it performed well during training and did not overfit, maintaining a good balance between learning and generalization.

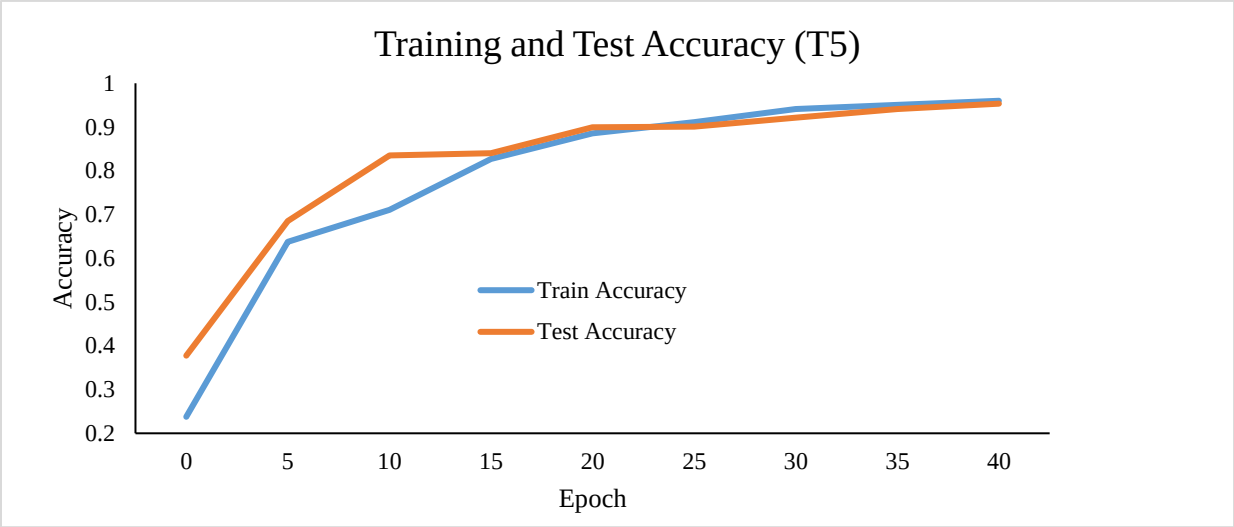

Figure 5.22: T5 MODEL Training and Testing Accuracy

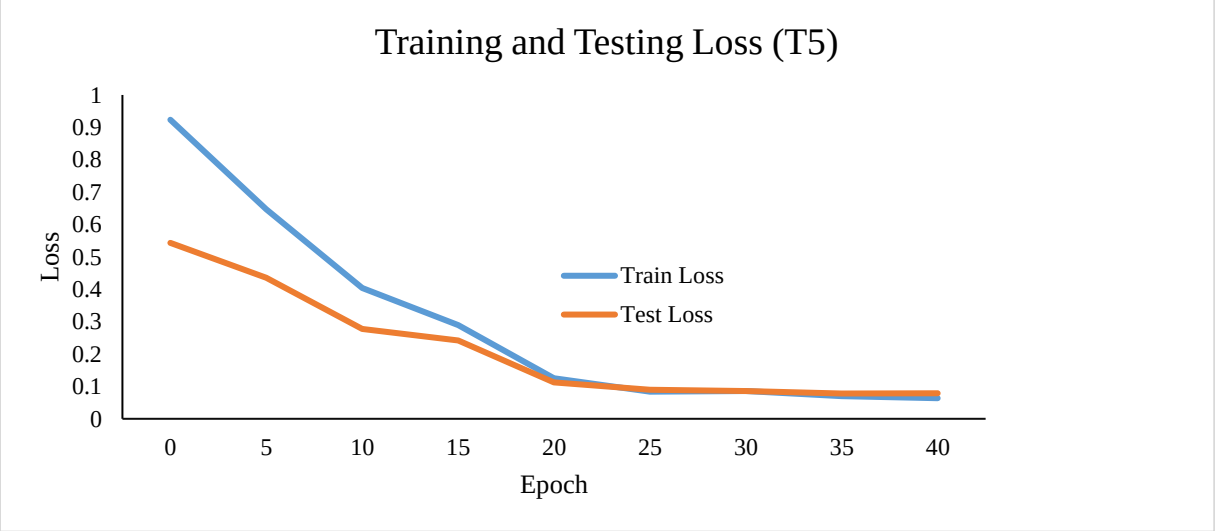

Figure 5.23: T5 MODEL Training and Testing Loss

Table 5.15 shows the summary of performance metrics. T5 shows strong performance, achieving a BLEU score of 0.8628, an Expert Evaluation Score of 0.9311, and a diversity score of 0.8245,

indicating that the model generates highly accurate, relevant, and comprehensive, as well as diverse, questions. Moreover, underscoring the quality of the generated questions in terms of context and alignment with Bloom's taxonomy.

Table 5.15: Performance metrics

	Training Accuracy	**Testing Accuracy**	**BLUE Score**	**Expert Evaluation(Heuristic)**	**Diversity Score**
T5	96.89%	95.36%	0.8628	0.9311	0.8245

5.4.2 Question Generation with BART Model

The BART model was fine-tuned for the task of question generation, with a primary focus on enhancing accuracy, contextual relevance, and diversity in the generated questions. The model was trained for 40 epochs with a batch size of 32, a learning rate of 3e-5, and 1,000 warm-up steps to ensure stable convergence.

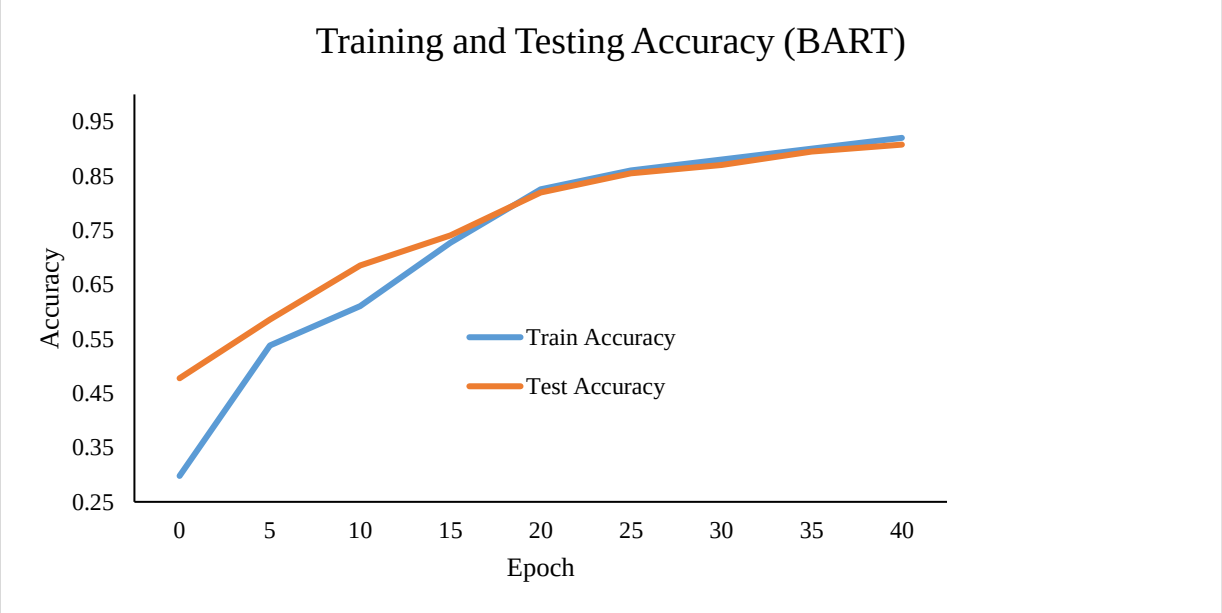

Figure 5.24 BART MODEL Training and Testing Accuracy.

As illustrated in Figure 5.24, after training, the model achieved a training accuracy of 92.23% and a testing accuracy of 90.76%, indicating that it effectively learned from the data and generalized well to new, unseen instances. Figure 5.25 shows that the training loss was 0.1135 and the validation loss was 0.1434, indicating the model's successful learning without overfitting, while maintaining a good balance between fitting the data and generalization.

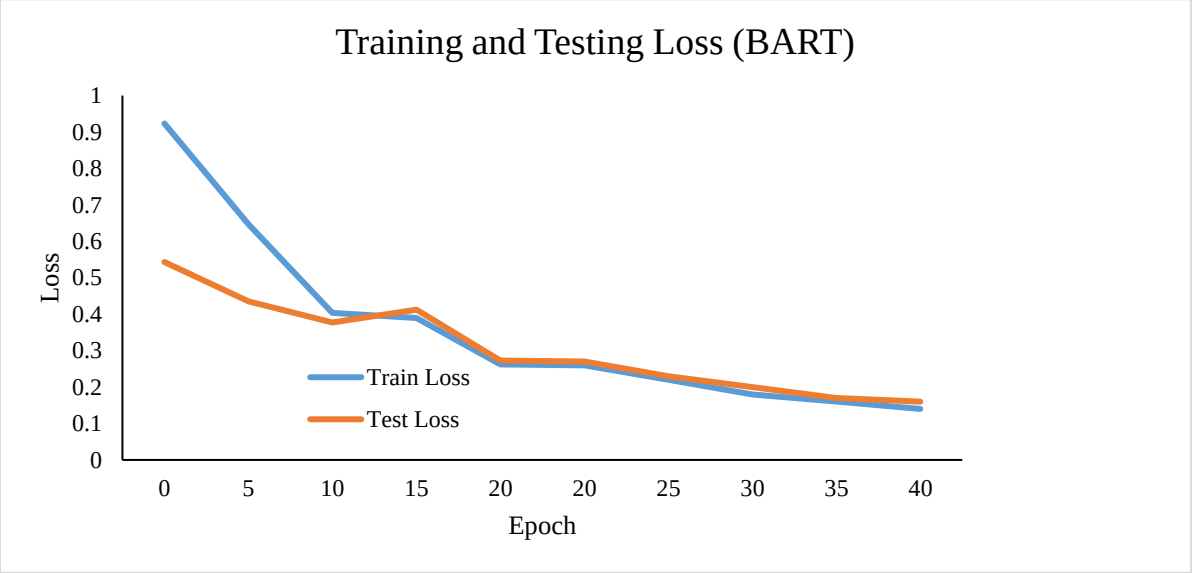

Figure 5.25 BART MODEL Training and Testing Accuracy

Summary table 5.16 of Evaluation metrics highlighted the BART model's strong performance, with a BLEU score of 0.8028, an Expert Evaluation Score of 0.8911, and a diversity score of 0.7521, showing the model's ability to generate highly accurate, relevant, and comprehensive questions in terms of context and alignment with Bloom's taxonomy.

Table 5.16: Performance metrics

	Training Accuracy	**Testing Accuracy**	**BLUE Score**	**Expert Evaluation(Heuristic)**	**Diversity Score**
BART	92.23%	90.76%	0.8028	0.8911	0.7521

5.4.3 Question Generation with Distill BART Model

The Distil BART model was fine-tuned for the task of automated question generation, with an emphasis on optimizing accuracy, contextual relevance, and diversity in the generated questions. The model was trained for 40 epochs with a batch size of 32, a learning rate of 3e-5, and 1,000 warm-up steps to ensure stable convergence. Upon completion of training, the model achieved a training accuracy of 92.46% and a testing accuracy of 90.32 % as illustrated in Figure 5.26, demonstrating its ability to learn effectively from the data and generalize well to unseen examples.

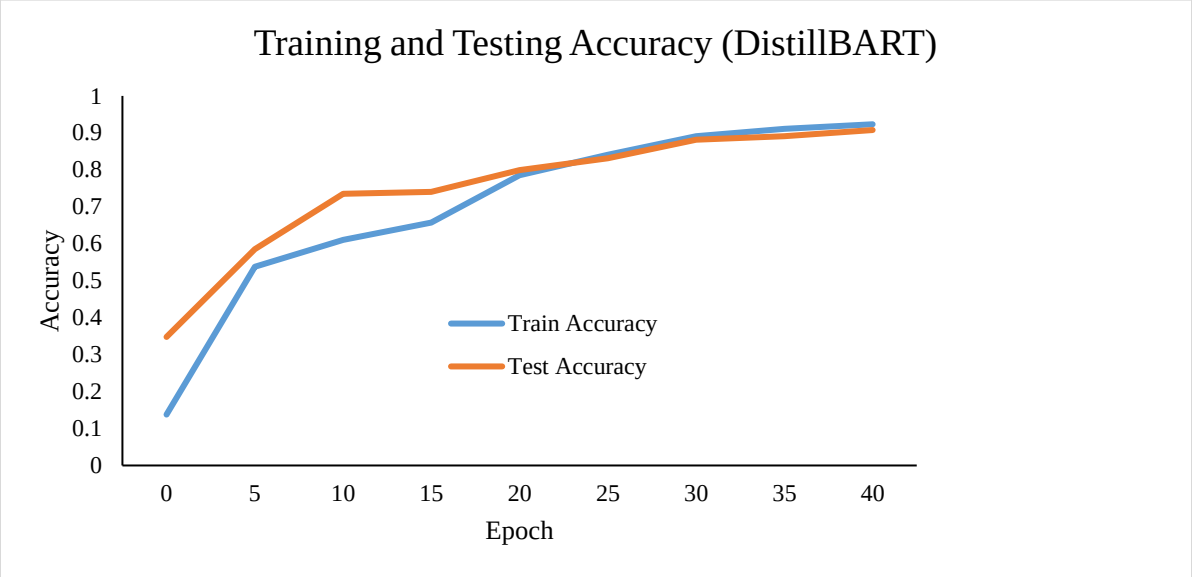

Figure 5.26 DistillBART MODEL Training and Testing Accuracy

Figure 5.27 illustrates that the training loss reached 0.1335, while the validation loss stood at 0.1599, indicating that the DistillBART model achieved effective learning during training while maintaining strong generalization capabilities and avoiding significant overfitting. It is further supported by evaluation metrics, which highlight the model's robust performance across both seen and unseen data. The consistent decline in both training and validation loss across epochs, as shown in the figure, confirms the model's stability and convergence throughout the training process.

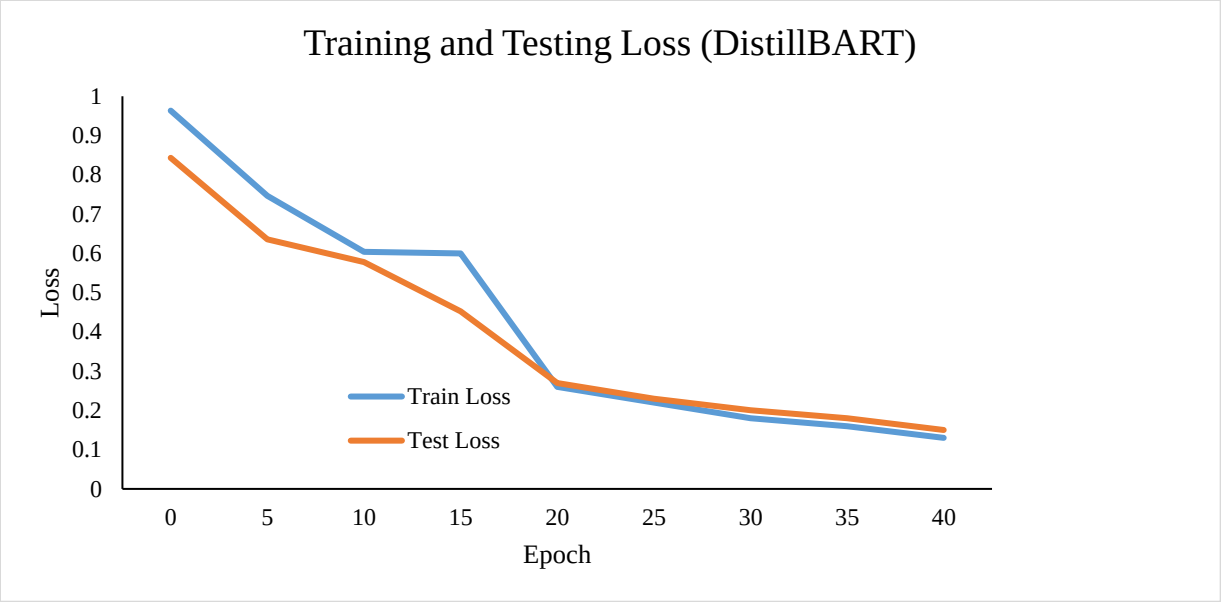

Figure 5.27: DistillBART MODEL Training and Testing Loss

The summary table 5.17 of performance metrics shows that the model achieved a BLEU score of 0.8158, an F1 score of 0.8952, a precision of 0.9145, and a recall of 0.9081, indicating that the Distil BART model generated highly accurate, relevant, and comprehensive questions. Additionally, the model received an Expert Evaluation Score of 0.8711, underscoring the quality of the generated questions in terms of context and alignment with Bloom's taxonomy. The model also achieved a diversity score of 0.7423, suggesting a good level of variation in the generated questions

Table 5.17: Performance metrics

	Training Accuracy	**Testing Accuracy**	**BLUE Score**	**Expert Evaluation(Heuristic)**	**Diversity Score**
Distill BART	92.46%	90.32%	0.8158	0.8711	0.7423

5.4.4 Question Generation with XLNet Model

The XLNet model was fine-tuned for the task of automated question generation, with a focus on optimizing accuracy, contextual relevance, and diversity in the generated questions. The model was trained for 40 epochs with a batch size of 32, a learning rate of 3e-5, and 1,000 warm-up steps

to ensure stable convergence. Upon completion of training, the model achieved a training accuracy of 84.12% and a testing accuracy of 75.77%, as illustrated in Figure 5.28, demonstrating its ability to learn from the data while effectively generalizing to unseen examples. However, the gap between training and validation accuracy suggests that there may be room for improvement in generalization.

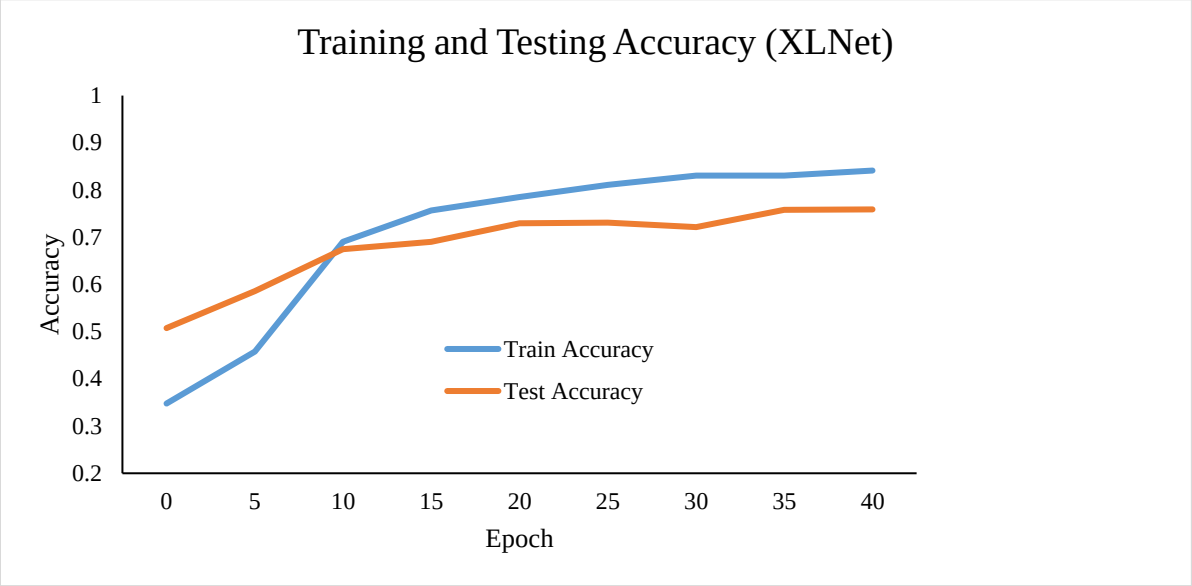

Figure 5.28: XLNet MODEL Training and Testing Accuracy

Figure 5.29 shows that the training loss was 0.1935, while the validation loss reached 0.3454, indicating that although the model performed effectively on the training data, its generalization to unseen data was relatively limited. The graph further illustrates a consistent decline in training loss over the epochs, reflecting successful optimization and learning during training. In contrast, the validation loss exhibits a relatively flat trend initially, followed by a slower decline and eventual stabilization around epoch 30. The noticeable gap between training and validation loss from epoch 15 onwards suggests a degree of overfitting.

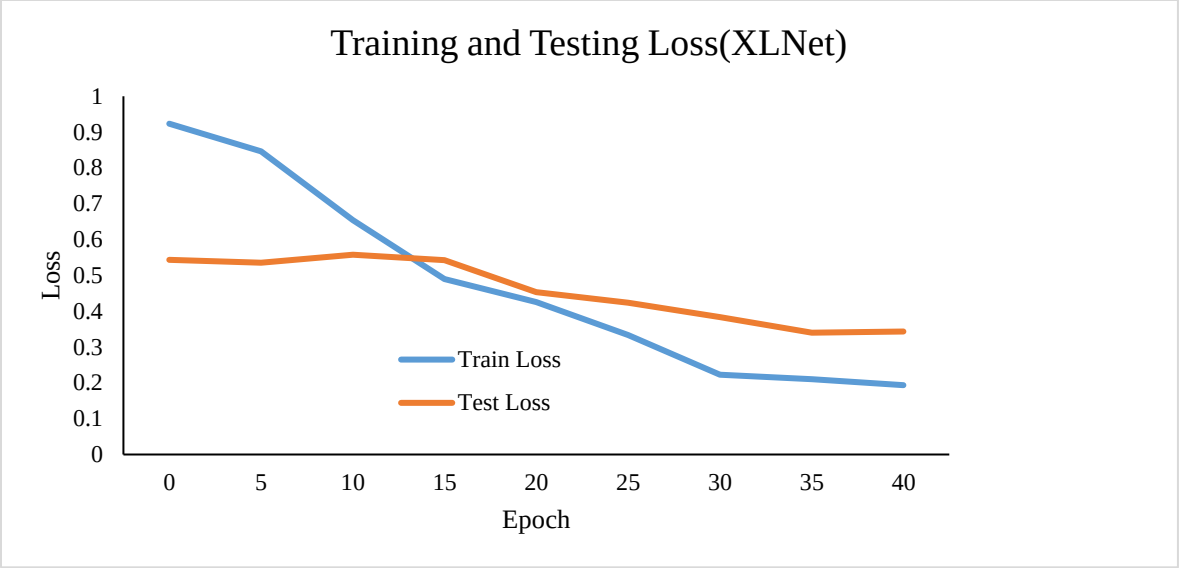

Figure 5.29: XLNet MODEL Training and Testing Loss

The summary table 5.18 of Evaluation metrics highlights the model's performance: XLNet achieved a BLEU score of 0.7628, indicating that the XLNet model generated reasonably accurate and relevant questions. The model received an Expert Evaluation Score of 0.7611, reflecting moderate alignment with Bloom's taxonomy in terms of question quality. Additionally, the model achieved a diversity score of 0.6755, indicating a reasonable level of variation in the generated questions.

Table 5.18: Performance metrics

	Training Accuracy	**Testing Accuracy**	**BLUE Score**	**Expert Evaluation(Heuristic)**	**Diversity Score**
XLNet	84.12%	75.77%	0.7628	0.7611	0.6755

5.4.5 Question Generation with BERT2BERT Model

The BERT2BERT model was fine-tuned for the task of automated question generation, with a focus on optimizing accuracy, contextual relevance, and diversity in the generated questions. The model was trained for 40 epochs with a batch size of 32, a learning rate of 3e-5, and 1,000 warm-

up steps to ensure stable convergence. As shown in Figure 5.30, upon completion of training, the model achieved a training accuracy of 75.82% and a validation accuracy of 70.86%, as depicted in Figure 4.30. Reflects that while the model learned from the training data, there is a significant gap in its ability to generalize to unseen examples.

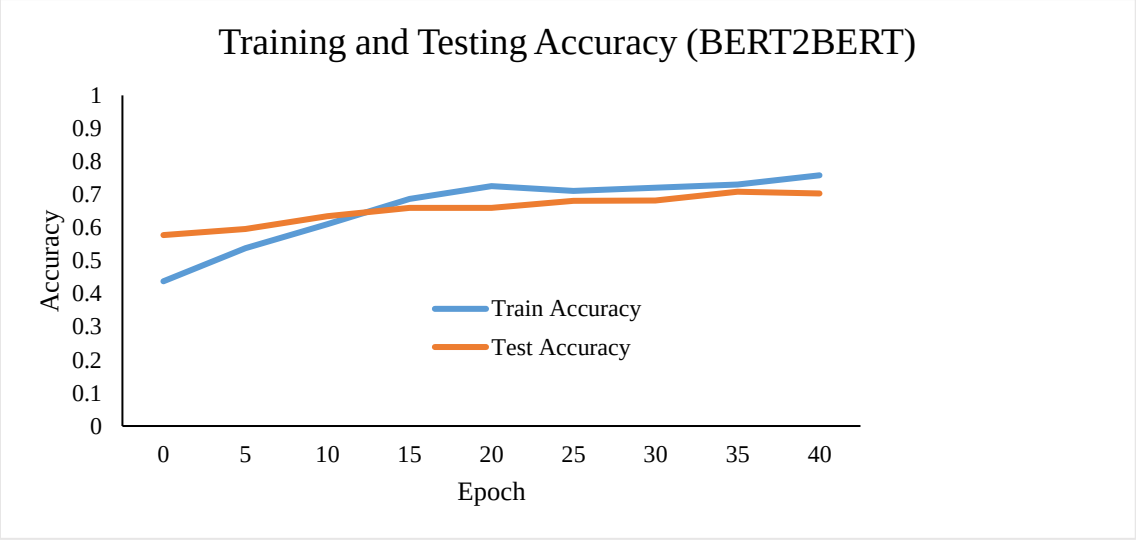

Figure 5.30 BERT2BERT MODEL Training and Testing Accuracy

As illustrated in figure 5.31 the training loss was 0.2635, and the validation loss was 0.3212, showing that the model was able to minimize error during training but struggled with generalization, as reflected by the relatively low validation accuracy.

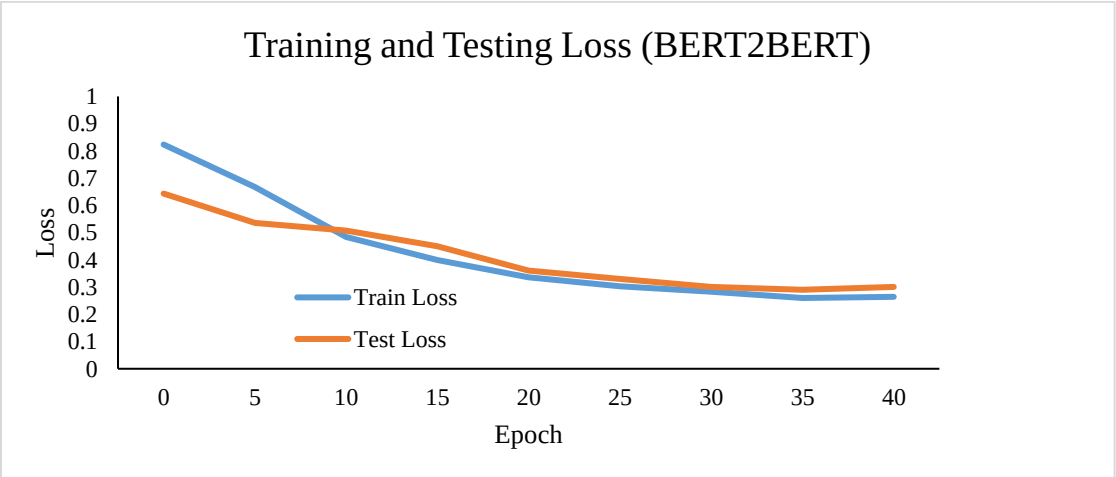

Figure 5.31: BERT2BERT MODEL Training and Testing Loss

Table 5.19 further underscores the evaluation metrics, which indicate that the BERT2BERT model achieved a BLEU score of 0.7028, suggesting that it generated moderately accurate and relevant questions. The model received an Expert Evaluation Score of 0.7211, suggesting a moderate alignment with Bloom's taxonomy in terms of question quality. Additionally, the model achieved a diversity score of 0.6193, indicating some variation in the generated questions.

Table 5.19 Performance matrices

	Training Accuracy	**Testing Accuracy**	**BLUE Score**	**Expert Evaluation(Heuristic)**	**Diversity Score**
BERT2BERT	75.82%	70.86%	0.7028	0.7211	0.6193

5.4.6 New Sample Testing Result for Question Generation

We Test some new sample of Content of different subject. And Model generating question from content according to different Bloom's Cognitive Level. The Output Table 5.20 shown below:

Table 5.20 Question Generation New Sample Testing Output

SI No	**Content**	**Bloom's Level**	**Actual or Reference Question**	**Generated Question**	**Correct or Incorrect**	**Accuracy**	**Error**
1	Biomedical engineers develop technologies that improve healthcare, such as prosthetics, medical imaging devices, and diagnostic tools.	Remember	Who are biomedical engineers and what is their role in healthcare?	How would you define biomedical engineers and their role in improving healthcare?	Correct	**Accuracy = 87.18%**	**Error =12.82%**
2		Understand	How would you compare the effective-	How would you compare the effective-	Incorrect		

	Their work blends biology, medicine, and engineering.		ness of bio-medical engi-neers in im-proving healthcare?	ness of bio-medical engi-neers in im-proving healthcare?			
3		Apply	How would you apply en-gineering principles to develop a healthcare so-lution?	What actions would you take to develop a technologies to improve healthcare?	Correct		
4		Analyze	How can bio-medical engi-neers be cate-gorized based on their skills or domains?	How can you classify differ-ent biomedical engineers based on their knowledge and experience?	Correct		
5		Evaluate	What factors would you consider in evaluating the success of bi-omedical en-gineering so-lutions?	What criteria would you use to assess the ef-fectiveness of biomedical en-gineers?	Correct		
6		Create	Can you propose a design	What changes would you	Correct		

			improvement to an existing biomedical device to enhance diagnostic performance?	make to improve a biomedical device for better diagnostic efficiency?			
7	Newton's laws of motion describe the relationship between a body and the forces acting on it. These laws form the foundation for classical mechanics. They explain how objects move and react to forces. Engineers and scientists use them in everything from designing rockets to cars.	Remember	Can you define Newton's laws of motion and their key principles?	How would you define Newton's laws of motion?	Correct		
8		Understand	How would you explain the difference between Newton's laws and other motion theories?	How would you compare the laws of motion with those of the same body?	Incorrect		
9		Apply	How would you demonstrate Newton's laws using a rocket launch scenario?	What actions would you take to demonstrate the laws of motion in a rocket?	Correct		

10		Analyze	How do Newton's three laws relate to each other and to motion behavior?	What is the theme of Newton's Law of motion?	Incorrect		
11		Evaluate	What conditions would cause Newton's laws to be inadequate or require modifica-tion?	What is the most important factor in determining the validity of Newton's laws of motion?	Correct		
12		Create	How would you modify Newton's laws for extreme conditions like near-light speed?	What changes would you make to revise a Newton's laws of motion?	Correct		
13	Photosynthesis is the process by which green plants convert sunlight into energy. They use carbon dioxide	Remember	What is photosynthesis and what role does it play in plant life?	How would you define photosynthesis?	Correct		
14		Understand	What is the role of carbon dioxide in	How would you clarify the meaning of	Correct		

	and water to produce glucose and oxygen. This process is essential for life, as it forms the base of the food chain and supplies oxygen to the atmosphere.		photosynthesis?	"carbon dioxide" in the process of photosynthesis?			
15		Apply	How would you demonstrate the process of photosynthesis in a classroom experiment?	What actions would you take to demonstrate photosynthesis?	Correct		
16		Analyze	How does photosynthesis contribute to the survival of ecosystems?	What is the function of Photosynthesis?	Correct		
17		Evaluate	What factors would you use to evaluate the efficiency of photosynthesis in different environments?	What criteria would you use to assess the success of photosynthesis?	Correct		
18		Create	Can you invent a synthetic process that mimics	Can you propose an alternative way of photosynthesis for energy production?	Correct		

			photosynthesis for energy generation?				
19	Inflation is the rate at which the general level of prices for goods and services rises. It reduces the purchasing power of money over time. Central banks often adjust interest rates to control inflation and stabilize the economy.	Remember	What is inflation and how does it affect economic growth?	How would you recognize the impact of inflation on economic growth?	Correct		
20		Understand	How would you explain the effect of inflation on service prices compared to goods?	How would you compare inflation to inflation in terms of service prices?	Incorrect		
21		Apply	How would a central bank adjust interest rates to control inflation?	What actions would you take to adjust interest rates to stabilize inflation?	Correct		
22		Analyze	How can inflation be categorized based on its causes and effects on the economy?	How can you classify different types of inflation based on their impact on economic growth?	Correct		
23		Evaluate	What factors should be	What criteria would you use	Correct		

			considered when assessing inflation in a low-inflation market?	to assess inflation in a market with low inflation?			
24		Create	What policy changes could you propose to influence inflation in a low-inflation economy?	What changes would you make to inflation rates in a country with low inflation?	Incorrect		
25	A computer algorithm is a set of instructions designed to perform a specific task. Algorithms are the backbone of software, from search engines to artificial intelligence. Efficiency and accuracy are key when designing algorithms.	Remember	What is an algorithm, and what role does it play in solving problems?	What do you remember about the role of algorithmic algorithms in a specific task?	correct		
26		Understand	How would you explain the term “efficiency” in algorithm design?	How would you clarify the meaning of "efficiency" in the context of algorithm design?	Correct		
27		Apply	How would you create an algorithm to automate a	What actions would you take to develop a computer algo-	Correct		

			simple decision-making task?	rithm to perform a specific task?			
28		Analyze	How can you categorize different types of algorithms based on time and space complexity?	How can you classify different types of algorithms based on their performance?	Correct		
29		Evaluate	What cri-teria would you use to evaluate the effectiveness of a computer algorithm?	What judgement would you make about computer algorithm?	correct		
30		Create	How would you redesign an existing algorithm to improve performance on a specific task?	How do you improve an algorithm to perform a specific task?	Correct		
31	Atoms are the smallest units of matter that retain the properties of an element. They consist of	Remember	What are atoms and how do they behave in a chemical reaction?	What do you remember about the properties of atoms in a chemical reaction?	correct		

32	protons, neu-trons, and elec-trons. Chemical reactions in-volve the rear-rangement of at-oms to form new substances.	Under-stand	How would you ex-plain or compare the prop-erties of different at-oms?	How would you compare the properties of atoms	Correct		
33		Apply	How would you use atomic struc-ture to predict outcomes of a chemical re-action?	How do you use the atoms in chemical re-action?	Incor-rect		
34		Analyze	How do dif-ferent atomic particles (pro-tons, neu-trons, elec-trons) influ-ence chemical reactions?	How would you compare the effects of neutrons and electrons on a chemical reac-tion?	Correct		
35		Evaluate	What factors would you consider to determine whether a chemical re-action involv-ing a certain	What criteria would you use to assess the success of an element in a chemical reac-tion?	correct		

			element is effective?				
36		Create	How would you modify a molecule to enhance its chemical reactivity or bonding properties?	What changes would you make to revise a molecule's properties to improve their chemical properties?	Correct		
37	Climate change refers to long-term shifts in temperatures and weather patterns. It's primarily driven by human activities like burning fossil fuels. The effects include rising sea levels, extreme weather, and biodiversity loss.	Remember	What is climate change, and what are its major effects on global economies?	How would you recognize the impact of climate change on global economies?	Correct		
38		Understand	How would you compare the impact of climate change with other global environmental challenges?	How would you compare climate change and other factors?	correct		
39		Apply	What steps would you take to assess the economic	What actions would you take to analyze the	Correct		

			impact of climate change?	impact of climate change on global economies?			
40		Analyze	How would you categorize different climate patterns and their causes?	How can you classify different types of climate patterns	Correct		
41		Evaluate	What criteria would you use to evaluate the severity of climate change in different regions?	How do you evaluate climate change?	correct		
42		Create	What new policy recommendations would you propose to reduce greenhouse gas emissions?	What changes would you make in climate change policies to reduce greenhouse gas emissions?	Correct		
43	Genetic engineering alters an organism's DNA to improve or introduce desired traits. It's	Remember	What is genetic engineering and what is its primary purpose?	What is Genetic Engineering?	Correct		

44	used in agriculture, medicine, and research. Genetically modified crops can resist pests and increase yields.	Understand	What is the central concept or idea behind genetic engineering?	What is the main idea of Genetic Engineering?	Correct		
45		Apply	How would you design a method to genetically modify a plant for better yield?	What actions would you take to develop a genetic engineering strategy to improve a plant's DNA?	Correct		
46		Analyze	What are the main components or techniques used in genetic engineering?	What are the features of Genetic Engineering?	Incorrect		
47		Evaluate	What factors should be considered in evaluating the effectiveness of genetic engineering in agriculture?	What criteria would you use to assess the success of genetic engineering in agriculture?	Correct		
48		Create	How would you create a	Design a solution for genetic engineering	Correct		

			policy framework to regulate and promote genetic engineering for crop improvement?	policies to improve yields?			
49	Robotics involves designing machines that can perform tasks automatically. Robots are used in factories, medicine, space exploration, and even homes. Combining mechanics, electronics, and software, robotics is a key part of modern automation.	Remember	What is robotics and how does it support automation?	How would you define robotics and its role in automation?	Correct		
50		Understand	How would you explain the differences between types of robots used in industries like manufacturing and healthcare?	How would you compare the different types of robotics in different industries?	Correct		
51		Apply	How would you design a basic robotics system for home automation?	What actions would you take to develop a robotics system in a home?	Correct		
52		Analyze	How would you cate-go-rize robotics	How can you classify differ-	Correct		

			sys-tems based on their application areas?	ent types of robotics according to its role?			
53		Evaluate	What factors should be considered to assess how effective robotics is in completing daily tasks?	What criteria would you use to assess the effectiveness of robotics in everyday tasks?	Correct		
54		Create	How do you redesign an existing robotic to improve its performance for a specific task?	What changes would you make to improve the design of a robotic machine for a job?	Correct		

5.5 Comparison of Generation Models

This section evaluates the performance of various transformer-based models for automated question generation, aligned with Bloom's Taxonomy. We evaluate these models using key performance metrics, including Training Accuracy, Validation Accuracy, BLEU Score, Expert Evaluation Score, and Diversity Score. These metrics offer a comprehensive analysis of each model's ability to generate high-quality, contextually relevant, and diverse questions.

In Figures 5.32 and 5.33, we can notice that the T-5 Model achieves the highest training and testing accuracy of 96.89% and 95.36%, respectively. These results indicate that T-5 not only has a strong learning capacity but also excellent generalization of unseen data. BART followed with a Training

Accuracy of 92.23% and a testing Accuracy of 90.76%. Indicating its capability in the question generation task. DistillBART, a more efficient version of BART, demonstrates good performance, achieving a Training Accuracy of 92.46% and a testing Accuracy of 90.32%. XLNet had a Training Accuracy of 84.12%, but a drop in testing Accuracy (75.77%). Indicates it did not generalize to unseen data. BERT2BERT showed the lowest performance among all the models with Training Accuracy (75.82%) but dropped to a low testing Accuracy of 70.86%.

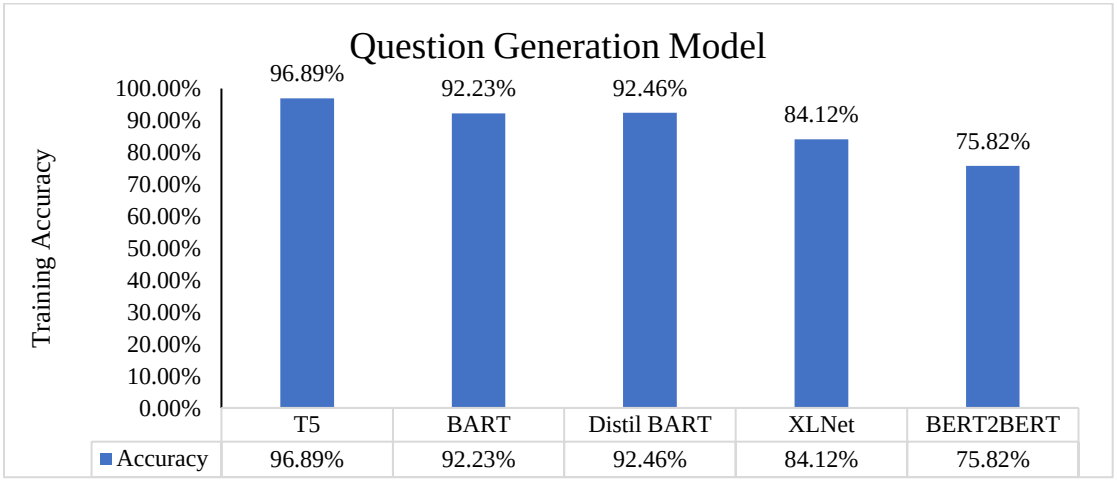

Figure 5.32 Question Generation Model Training Comparison

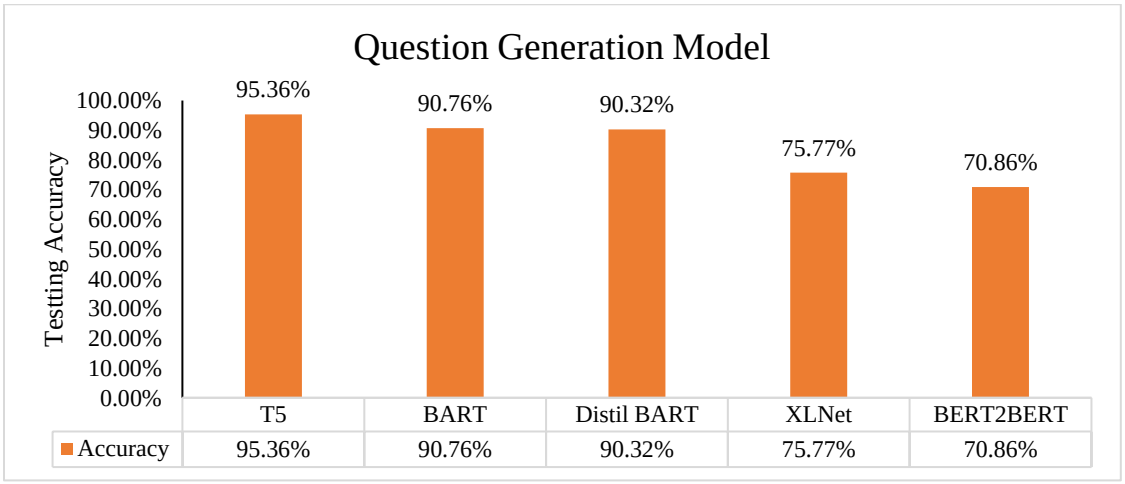

Figure 5.33 Question Generation Model Testing Comparison

Figures 5.34, 5.35, and 5.36 show that the T5 model achieved a BLEU score of 0.8628 and a high Expert Evaluation Score of 0.9311, demonstrating its ability to generate contextually relevant and

diverse questions, with a Diversity Score of 0.8245. BART scored a BLEU Score of 0.8028 and an Expert Evaluation Score of 0.8911, showing good fluency and relevance. But slightly less diversity (Diversity Score of 0.7521) compared to T5.DistilBART, a more efficient version of BART achieved a BLEU Score of 0.8158 and a Diversity Score of 0.7423, offering solid performance but slightly less quality than T5 and BART, with an Expert Evaluation Score of 0.7611. XLNet achieved a BLEU Score of 0.7628, with a Diversity Score of 0.6755, indicating its struggle in generating diverse questions. BERT2BERT showed the lowest BLEU Score of 0.7028 and had a Diversity Score of 0.6193, suggesting limited diversity and quality in question generation, with an Expert Evaluation Score of 0.7028.

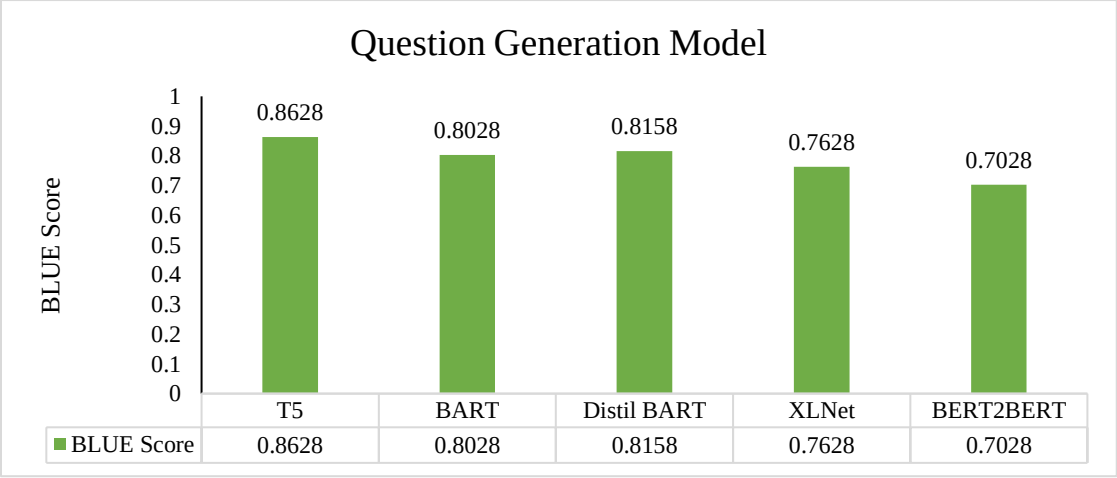

Figure 5.34 Question Generation BLUE Score Comparison

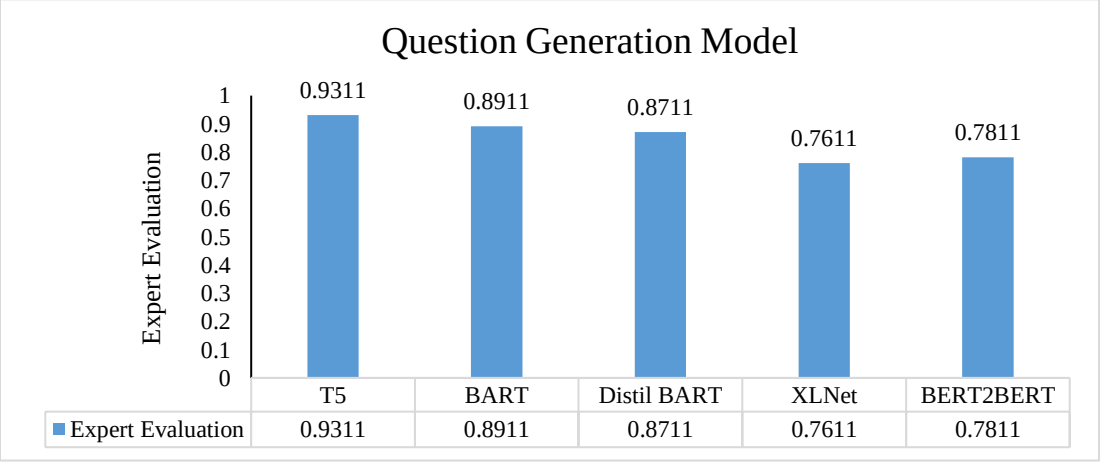

Figure 5.35 Question Generation Model Expert Comparison

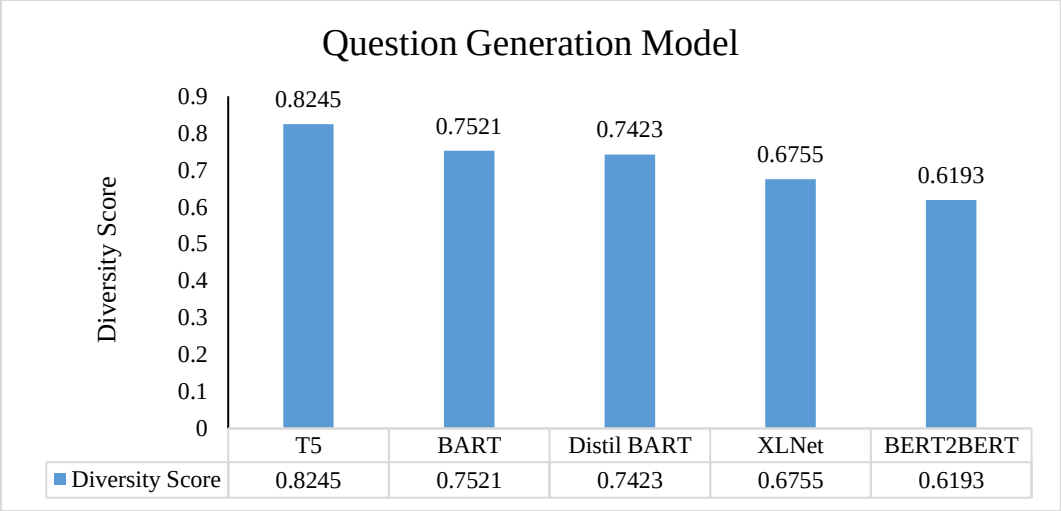

Figure 5.36 Question Generation Model Diversity Comparison

Table 5.21 shows a Summary of the Generation model performance, and among all of the T5 models, it shows strong performance overall with the accuracy 95.36%, BLEU Score 0.8628 and Expert Evaluation and Diversity Score is 0.9311 and 0.8245.

Table 5.21: Summary of Generation model Performance

Model Name	Training Accuracy (%)	Testing Accuracy (%)	BLEU Score	Expert Evaluation Score	Diversity Score
T5	96.89	95.36	0.8628	0.9311	0.8245
BART	92.23	90.76	0.8028	0.8911	0.7521
Distil BART	92.46	90.32	0.8158	0.8711	0.7423
XLNet	84.12	75.77	0.7628	0.7611	0.6755
BERT2BERT	75.82	70.86	0.7028	0.7211	0.6193

CHAPTER 6

CONCLUSION

6.1 INTRODUCTION

This study developed and evaluated machine learning models to automate the prediction and generation of exam questions based on Bloom's Taxonomy. The primary objective was to evaluate the ability of these models to categorize questions accurately across Bloom's cognitive levels and to generate contextually relevant, high-quality questions for educational assessments. The research explored several machine learning techniques, including a hybrid model, as well as transformer-based models such as BERT, DistilBERT, ROBERTa, T5, and others. These approaches aimed to provide scalable and efficient solutions for educators, automating the processes of question creation and prediction in alignment with Bloom's Taxonomy of Cognitive Domains.

6.2 CONCLUSION

This study effectively demonstrated the potential of automated systems for both question prediction and generation, grounded in Bloom's Taxonomy. The hybrid model achieved the highest performance in question prediction, with an accuracy of 96.39%. This model outperformed others, such as BERT and DistilBERT, highlighting its robustness and effectiveness in classifying questions across all cognitive levels of Bloom's Taxonomy. In the task of question generation, the T5 model excelled, attaining a testing accuracy of 95.36% and a BLEU score of 0.8628. This exceptional performance confirms the model's capacity to generate high-quality, diverse, and contextually accurate questions, surpassing other transformer-based models, such as BART, DistillBART, XLNet, and BERT2BERT.

The findings underscore the significant promise of automated systems for creating and classifying educational content, offering practical and scalable solutions for educators. The proposed models not only outperformed existing approaches but also paved the way for more efficient, effective, and high-quality educational assessments. By automating the process of question creation and Prediction, educators can save time and resources, ensuring greater consistency and accuracy in assessments.

6.3 FUTURE WORK

Although the proposed system shows strong performance in both question prediction and generation tasks, several opportunities exist for further enhancement.

Expanding the dataset to encompass a broader range of academic disciplines and multilingual content could significantly improve the model's generalizability across diverse educational contexts. Furthermore, incorporating feedback from educators and learners through a human-in-the-loop framework could enhance the pedagogical relevance of generated questions, ensuring alignment with instructional goals rather than relying solely on automated evaluation metrics. In short, future improvements should focus on broader dataset coverage, better adaptability, and integration of human feedback for more effective educational use.

REFERENCES

[1] Y. Timakova and K. A. Bakon, "Bloom's Taxonomy-Based Examination Question Paper Generation System," *International Journal of Information Systems and Engineering*, vol. 6, no. 2, pp. 76–92, Nov. 2018, doi: 10.24924/ijise/2018.11/v6.iss2/76.92.

[2] Jennifer O. Contreras, Shadi M. S. Hilles, and Z. Bakar, "Essay Question Generator based on Bloom's Taxonomy for Assessing Automated Essay Scoring System," *2021 2nd International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, Cameron Highlands, Malaysia, 2021, pp. 232–237. doi: 10.1109/ICSCEE50312.2021.9498166

[3] M. G. Ben-Jacob, "Assessment: Classic and Innovative Approaches," *Open Journal of Social Sciences*, vol. 5, no. 1, pp. 46–51, Jan. 2017. [Online]. Available: https://doi.org/10.4236/jss.2017.51004

[4] G. T. Thing, H. Mohamed, N. A. Akmal, and H. S. C. Mason, "Questions Classification According to Bloom's Taxonomy Using Universal Dependency and Word Net," *Test Engineering and Management*, vol. 82, pp. 4374–4385, Sep. 2020.

[5] B. T. G. S. Kumara, A. Brahmana, and I. Paik, "Bloom's Taxonomy and Rules Based Question Analysis Approach for Measuring the Quality of Examination Papers," *International Journal of Knowledge Engineering*, vol. 5, no. 1, pp. 20–24, Jul. 2019, doi: 10.18178/ijke.2019.5.1.111.

[6] S. T. Idris, S. F. H. Sepita, and H. Safitri, "Profile of the Ability of Prospective Biology Teachers in Making Question Instruments Using Bloom's Taxonomy," *REID (Res. Eval. Educ.)*, vol. 7, no. 2, pp. 177–185, 2021.

[7] D. A. Abduljabbar and N. Omar, "Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination," *Journal of Theoretical and Applied Information Technology*, vol. 78, no. 3, pp. 447–455, Aug. 2015.

[8 I. Assaly and M. S. Oqlah, "Using Bloom's Taxonomy to Evaluate the Cognitive Levels of Master Class Textbook's Questions," *English Language Teaching*, vol. 8, no. 5, p. 100, Apr. 2015. [Online]. Available: https://doi.org/10.5539/elt.v8n5p100

[9] A. Osman and A. A. Yahya, "Classifications of Exam Questions Using Linguistically-Motivated Features: A Case Study Based on Bloom's Taxonomy," in *Proc. 6th Int. Arab Conf. on Quality Assurance in Higher Education (IACQA' 2016)*, Sudan Univ. of Science and Technology, Republic of Sudan, Feb. 2016.

[10] S. Shaikh, S. M. Daudpotta, and A. S. Imran, "Bloom's Learning Outcomes' Automatic Classification Using LSTM and Pretrained Word Embeddings," *IEEE Access*, vol. 9, pp. 117887–117909, Aug. 2021, doi: 10.1109/ACCESS.2021.3106443.

[11] Y Y. Bengio and J.-S. Senecal, "Adaptive Importance Sampling to Accelerate Training of a Neural Probabilistic Language Model," *IEEE Transactions on Neural Networks*, vol. 19, no. 4, pp. 713–722, Apr. 2008.

[12] N. Yusof and C. J. Hui, "Determination of Bloom's Cognitive Level of Question Items Using Artificial Neural Network," in *Proceedings of the 10th International Conference on Intelligent Systems Design and Applications (ISDA)*, pp. 866–870, Nov. 2010.

[13] Lalchhuanmawii and Zoramchhuana, "Analysis of HSSLC (Class XII) Arts Examination Question Papers in Mizoram (India) Using the Cognitive Levels of Bloom's Taxonomy of Educational Objectives," *International Journal of Advanced Research*, doi: 10.21474/IJAR01/17756. [Online]. Available: https://dx.doi.org/10.21474/IJAR01/17756

[14] M. D. Laddha, V. T. Lokare, A. W. Kiwelekar, and L. D. Netak, "Classifications of the Summative Assessment for Revised Bloom's Taxonomy by Using Deep Learning," *International Journal of Engineering Trends and Technology*, vol. 69, no. 3, pp. 211–218, Mar. 2021, doi: 10.14445/22315381/IJETT-V69I3P232.

[15] A. Chanaa and N. E. El Faddouli, "A cognitive level evaluation method based on machine learning approach and Bloom of taxonomy for online assessments," *Journal of Education and Learning (EduLearn)*, vol. 18, no. 2, pp. 553–560, May 2024, doi: 10.11591/edulearn.v18i2.20948.

[16] A. Yaacoub, J. Da-Rugna, and Z. Assaghir, "Assessing AI-Generated Questions' Alignment with Cognitive Frameworks in Educational Assessment," in *Proceedings of the International Conference on Artificial Intelligence in Education (AIED)*, 2023.

[17] S. Dhainje, R. Chatur, K. Borse, and V. Bhamare, "An Automatic Question Paper Generation: Using Bloom's Taxonomy," *International Research Journal of Engineering and Technology (IRJET)*, vol. 10, no. 1, pp. 3397–3400, Jan. 2023. [Online]. Available: www.irjet.net

[18] K. Jayakodi, M. Bandara, I. Perera, and D. Meedeniya, "WordNet and Cosine Similarity Based Classifier of Exam Questions Using Bloom's Taxonomy," *International Journal of Emerging Technologies in Learning (IJET)*, vol. 11, no. 4, pp. 157–164, 2016, doi: 10.3991/ijet.v11i04.5654.

[19] C. Chavan, A. Rajguru, and N. Varghese, "An Automatic Question Paper Generation System Using ML Techniques," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 11, no. 4, pp. 128–135, Apr. 2023.

[20] T. Chandio, H. Mangi, and A. A. Shaikh, "Enhancing the Practical Learning of English Applying Bloom's Taxonomy," *International Journal of Language Teaching and Learning*, vol. 3, no. 2, pp. 55–64, 2022.

[21] M. O. Gani, H. M. Shuvo, and S. Akhter, “Towards Enhanced Assessment Question Classification: A Study Using Machine Learning, Deep Learning, and Generative AI,” *Connection Sci.*, vol. 37, no. 1, 2025, doi: 10.1080/09540091.2024.2445249.

[22] N. Herrmann-Werner et al., “Assessing ChatGPT’s Mastery of Bloom’s Taxonomy using Psychosomatic Medicine Exam Questions,” *medRxiv*, Aug. 2023. [Online]. Available: https://doi.org/10.1101/2023.08.18.23294159

[23] M. Stevani and K. Tarigan, “Evaluating English Textbooks by Using Bloom’s Taxonomy to Analyze Reading Comprehension Questions,” *SALEE*, Aug. 2022. [Online]. Available: https://doi.org/10.35961/salee.v0i0.526.

[24] T. R. Fuadi, E. S. Bahriah, and S. Agung, "Analysis of Final Semester Assessment Questions in Grade XI Chemistry," *Jurnal Pendidikan Kimia*, 2022.

[25] S. Diab and B. Sartawi, "Classification of Questions and Learning Outcome Statements into Bloom's Taxonomy Using Similarity Measurements," *International Journal of Management and Information Technology (IJMIT)*, vol. 9, no. 2, 2017.

[26] P. Sahu and M. Cogswell, "Comprehension Based Question Answering using Bloom’s Taxonomy," *Journal of Intelligent Learning Systems*, arXiv:2106.04653v1 [cs.CL], 2021.

[27] S. Kadiyala and S. Gavini, "Applying Bloom's Taxonomy in Framing MCQs," *Journal of Dr. NTR University of Health Sciences*, 2017.

[28] K. Moholkar and S. H. Patil, "Multiple Choice Question Answer System Using Ensemble Deep Neural Network," in *Proceedings of the 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, vol. 5, pp. 762–766, Mar. 2020

[29] M. Tanuja, K. Sushma, and B. Shobha, "QuestionGuru: An Automated Multiple-Choice Question Generation System," in *Proceedings of the International Conference on Computing, Intelligence and Communication (ICCIC)*, 2023.

[30] W. Wijanarko *et al*., "Automated Question Generation Using Keyphrase-Based Context-Free Grammar for Bloom’s Taxonomy," *Education and Information Technologies*, vol. 25, no. 6, pp. 5201–5220, 2020, doi: 10.1007/s10639-020-10356-4.

[31] V. Sudhir, "Taxonomy-Based Educational Question Generation Using NLP," in *Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2018.

[32] F. Baharuddin and M. F. Naufal, “Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom's Taxonomy,” *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 2, pp. 253–263, Nov. 2023. DOI: 10.20473/jisebi.9.2.253-263

[33] H. Adel, A. Dahou, A. Mabrouk, M. A. Abd Elaziz, M. Kayed, I. El-Henawy, S. Alshathri, and A. A. Ali, "Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm," *Mathematics*, vol. 10, no. 3, p. 447, Jan. 2022. DOI: 10.3390/math10030447

[34] "Text Classification using Transformer Models: RoBERTa and XLNet," ProjectPro. [Online]. Available: https://www.projectpro.io/project-use-case/text-classification-using-transformer-models-roberta-and-xlnet [Accessed: Jul. 20, 2025].

[35] A. Chabane, M. Sahnoun, and B. Bettayeb, "Forecasting KPIs of Production Systems Using LSTM Networks," in *Proc. Int. Conf. Cyber Management and Engineering (CyMaEn)*, Tunisie, May 2021. [Online]. Available: https://dx.doi.org/10.1109/CyMaEn50288.2021.9497278

[36] "Support Vector Machines(SVM) – A Complete Guide for Beginners," Analytics Vidhya. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/ [Accessed: Jul. 20, 2025].

[37] L. Xue and N. Constant, "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," arXiv:2010.11934v3 [cs.CL], 2021.

[38] A. Alokla, W. K. Gad, W. Nazih, M. Aref, and A.-B. M. Salem, "Pseudocode Generation from Source Code Using the BART Model," *Mathematics*, vol. 10, no. 21, p. 3967, Oct. 2022. DOI: 10.3390/math10213967

[39] H. Chouikhi, F. Jarray, and M. Alsuhaibani, "A Sequence-to-Sequence Neural Network for Joint Aspect Term Extraction and Aspect Term Sentiment Classification Tasks," in *Proc. 15th* [39] Int. Conf. Agents and Artificial Intelligence (ICAART), *2023. [Online]. Available: https://doi.org/10.5220/0011620500003393*

[40] S. Zheng, *xlnet_extension_tf* [Online]. Available: https://github.com/stevezheng23/xlnet_extension_tf