A Modular Approach to Automatic Question Generation: Leveraging Large Language Models for Adaptive Learning

Kunlin Luo

School of Computer Science
The University of Auckland

Supervisor: Mano Manoharan and Xinfeng Ye

A thesis submitted in partial fulfilment of the requirements for the degree of MSc (Computer Science), The University of Auckland, 2025.

Abstract

This study introduces a modular framework for automated question generation, designed to address the limitations(include rigid structure, lack of flexibility, inability to handle diverse question types, reliance on static evaluation metrics, and insufficient adaptability to unstructured input formats) of traditional end-to-end models by integrating distinct agents for analysis, generation, evaluation, and refinement. The proposed model emphasizes alignment with educational objectives, supporting diverse question types, such as numerical and free-format text, while producing context-specific answers aligned with learning outcomes. Unlike existing models, which often rely on static evaluation metrics and structured datasets, this framework incorporates an iterative optimization process with an automated feedback to refine question quality. By ensuring feedback-driven question relevance and diversity, the model demonstrates adaptability across domains and unstructured input formats. Experimental results highlight its potential to enhance educational assessments, personalized learning, and intelligent tutoring systems. This work combines modular design, which separates the task into specialized agents, with dynamic optimization, which continuously improves the process based on real-time feedback. The modular design enables learning and improvement, while the dynamic optimization process ensures that the questions are more contextually relevant and pedagogically sound, setting a foundation for future advancements in question generation systems.

Acknowledgement

I would like to express my deepest gratitude to my supervisors, Mano Manoharan and Xinfeng Ye, for their invaluable guidance, unwavering support, and insightful feedback throughout the course of this research. Their expertise, encouragement, and dedication have been instrumental in shaping my academic journey and the completion of this project.

I also wish to thank my family, especially my parents, for their constant love, patience, and understanding. Their belief in me has been a source of strength, and their support has made this achievement possible.

Finally, I would like to express my sincere appreciation to the 10 friends who participated in the experiments for this study. Their time, effort, and collaboration were crucial to the success of this project, and I am grateful for their contributions.

Contents

Chapter 1

Introduction

1.1 Problem Statement

Automatic Question Generation (AQG) is a critical component of modern educational systems. We want instructional efficiency by being able to create a lot of questions. From the learning perspective, this also gives a lot of practice questions students can work on. With numerical questions, AQG could help us set up examinations with isomorphic questions so that students cannot copy answers from those who sit next to them. Traditional education methods often involve one-size-fits-all approaches where students receive the same set of problems regardless of their individual learning needs. This limitation can hinder learning outcomes, especially in environments where students have varied backgrounds and cognitive abilities.

The advent of machine learning, particularly large language models (LLMs), has significantly improved the way automated systems can generate meaningful, context-specific questions that cater to different cognitive levels. However, traditional AQG methods—often relying on rule-based systems or simple templates—fail to fully capture the complexity and variability required for dynamic educational content generation.

This research aims to address these shortcomings by developing a modular framework for AQG, leveraging LLMs and advanced techniques such as retrieval-augmented generation (RAG) and case-based reasoning (CBR). The proposed framework employs distinct agents for the analysis, generation, evaluation, and refinement of questions, allowing for continuous optimization and alignment with educational objectives. By integrating feedback mechanisms and allowing for personalized learning experiences, this approach sets a foundation for improving the effectiveness of AI-driven question generation systems.

1.2 Project Objectives

The main objectives of this project are to:

- **Design a Modular AQG Framework**: Introduce a system that employs separate agents for each task (material analysis, question generation, evaluation, refinement, and answer generation) to address the limitations of traditional, monolithic AQG models.
- **Align Questions with Educational Objectives**: Ensure that generated questions are not only relevant but also align closely with specific learning outcomes and cognitive levels.
- **Support Multiple Types of Questions**: Enable the system to generate various types of questions, including conceptual, numerical, and application-based questions, to ensure a comprehensive assessment of student learning.
- **Optimize Question Quality**: Incorporate an automatic feedback loop that continuously improves the quality of generated questions by evaluating and refining them based on real-time suggestions.
- **Enhance Personalization**: Use RAG and CBR techniques to provide contextually relevant and accurate questions, adapting to the learner's progress and cognitive needs.

1.3 Research Contribution

This work makes several contributions to the field of AQG:

- **Modular Approach**: The introduction of distinct agents for analysis, generation, evaluation, and refinement represents a novel approach that allows for dynamic, adaptable, and continually improving question generation. This modular design contrasts with traditional end-to-end systems.
- **Personalized Learning**: By integrating RAG and CBR techniques, the system tailors questions to individual learning paths, offering more personalized and contextually relevant educational experiences.
- **Automated Feedback and Optimization**: The inclusion of an evaluation agent and refinement agent provides a unique method for continuously improving generated content. This feedback loop optimizes both the questions and the answers to meet the highest educational standards.
- **Alignment with Educational Objectives**: The framework ensures that generated questions are explicitly aligned with the learning objectives, making them more effective for student comprehension and assessment.

This work combines modular design, which separates the task into specialized agents, with dynamic optimization, which continuously improves the process based on real-time feedback, offering an innovative approach that promises to advance the field of AQG. The outcomes of this research will contribute to the future development of intelligent tutoring systems and other educational tools that rely on automated question generation.

1.4 Structure of the Thesis

This thesis is organized as follows:

- **Chapter 2: Background** provides a detailed review of the relevant techniques and approaches used in AQG, including LLMs, RAG, CBR, and the evaluation metrics for AQG systems.
- **Chapter 3: Previous Work** surveys the existing literature on AQG systems, comparing different methods, highlighting their strengths and weaknesses, and identifying the gaps that this research seeks to fill.
- **Chapter 4: Methodology** details the modular framework proposed in this study, explaining each agent's role and how they interact within the system to generate, evaluate, and refine educational content.
- **Chapter 5: Experimental** presents the results from the experiments conducted using the proposed system. This includes a comparison with existing models and an analysis of the performance based on several evaluation criteria.
- **Chapter 6: Conclusion and Future Work** summarizes the findings of this research and suggests possible improvements and directions for future research.

The data sets, code, model generation problems, and scoring results used in the experiments in this thesis are available at: https://github.com/LKL1111/Multi-Agent-Educational-Question-Generation.

Chapter 2

Background

2.1 Introduction to Automated Question Generation (AQG)

Automated Question Generation (AQG) is a key area of research within the field of Natural Language Processing (NLP) that focuses on developing systems capable of generating meaningful and contextually relevant questions from educational texts. AQG plays an essential role in modernizing assessment systems and creating personalized learning experiences. Traditional question generation systems often rely on handcrafted rules, templates, or simple methods that do not account for the complexities of human learning. However, with the advent of large-scale language models and deep learning techniques, AQG systems have evolved to generate more varied, accurate, and context-sensitive questions.

The task of AQG can be broken down into several challenges:

- **Question Type Classification:** Determining the type of question that best fits a given educational objective, such as conceptual, factual, numerical, or application-based questions.
- **Contextual Relevance:** Ensuring that the generated questions are closely tied to the material, helping learners to engage with the content effectively.
- **Cognitive Alignment:** Ensuring that the questions are aligned with the cognitive level required for the learner's development, from basic recall to higher-order thinking as proposed by Bloom's Taxonomy [9].
- **Question Quality:** Generating questions that are clear, relevant, and promote critical thinking while avoiding redundancy or ambiguity.

Recent advancements in AQG systems, particularly those that utilize large language models like GPT-3 [16], T5 [9], and BERT [4], have significantly improved the performance and capabilities of these systems. However, challenges remain in terms of controlling the diversity, quality, and relevance of generated questions.

2.2 Techniques Used in the Project

The proposed AQG system in this project builds upon several advanced techniques in NLP and machine learning, which are essential for generating high-quality questions and answers. These techniques include:

2.2.1 Large Language Models (LLMs)

Large Language Models (LLMs), such as GPT-3, T5, and BERT, have been pivotal in revolutionizing natural language understanding and generation tasks. These models are pre-trained on massive datasets and have the capability to generate coherent, contextually relevant text by predicting the next word or token in a sequence. Fine-tuning these models for specific tasks, such as AQG, allows for more domain-specific and context-aware question generation.

- **GPT-3:** A autoregressive language model that can generate highly coherent and context-sensitive text. Its ability to perform few-shot learning makes it particularly suited for question generation tasks.
- **T5:** A unified framework for text-to-text tasks, T5 is versatile and effective for both question generation and answering, enabling it to perform multiple NLP tasks using the same architecture.
- **BERT:** Unlike GPT-3 and T5, which are autoregressive, BERT is a bidirectional transformer model that focuses on understanding the context of the entire sentence, making it ideal for text classification tasks, including question answering and generating questions that require deep comprehension.

2.2.2 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) [15] is a hybrid approach that combines the strengths of both retrieval-based and generation-based models. The architecture of RAG is shown in figure 2.1. In traditional generation-based systems, the model generates text based on its learned knowledge from pretraining. However, these models lack real-time access to external information, which may limit their accuracy, especially in cases that require up-to-date knowledge or domain-specific facts.

RAG improves upon this by retrieving relevant information from external knowledge bases (such as text corpora or databases) before generating text. For AQG, RAG allows for the retrieval of relevant content (e.g., textbook material or previous learning objectives) that can then

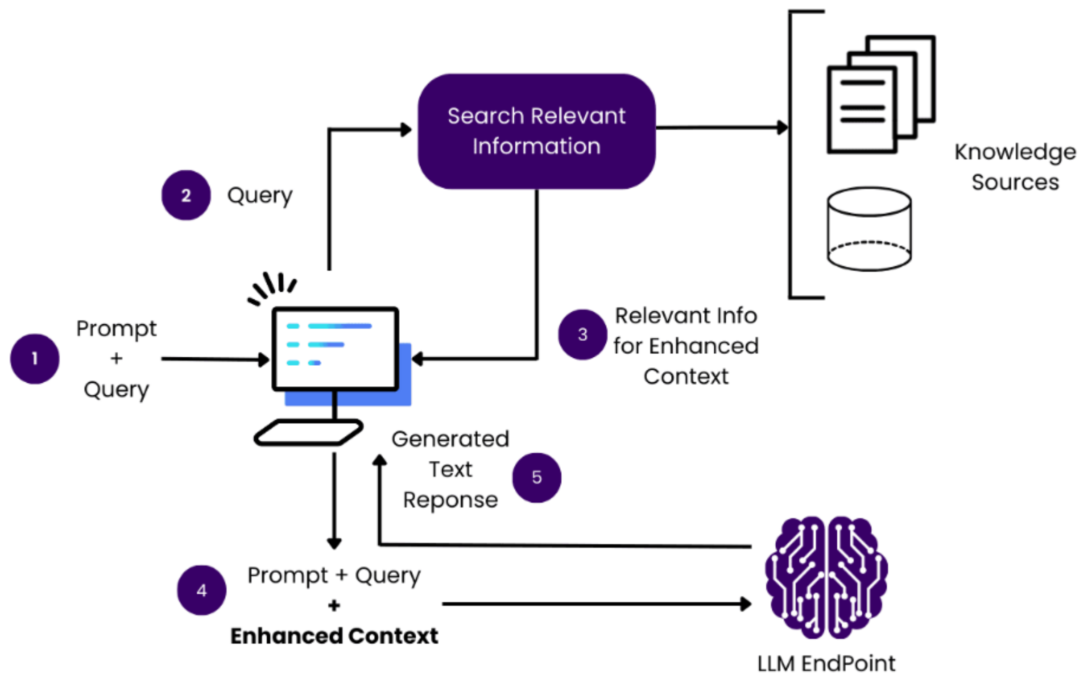

Figure 2.1: The architecture of RAG

be used to generate more accurate and contextually grounded questions. This approach ensures that the generated questions are not only grounded in the input material but also factually correct, as they are informed by external knowledge.

2.2.3 Case-Based Reasoning (CBR)

Case-Based Reasoning (CBR) is a problem-solving methodology that uses past cases to address new problems by finding similar cases and adapting their solutions. In the context of AQG, CBR can be used to enhance the retrieval process by organizing previous questions and answers into cases. When a new learning objective is provided, CBR can retrieve the most relevant prior cases and use them to generate new questions based on their context. This technique helps ensure that the generated questions are not only relevant to the current material but also adapt to the learner's progression.

For example, imagine a learner studying algebra. If the learning objective involves solving quadratic equations, the CBR system can retrieve previous cases where questions related to quadratic equations were generated, such as ”What is the solution to the quadratic equation”

$$ax^2 + bx + c = 0$$

The system can then adapt these past questions to fit the new learning context or the learner’s current difficulty level, such as by altering the coefficients or introducing a word problem context.

By using case-based reasoning, the system can make smarter decisions about which previous knowledge or questions should be reused or modified to suit the current learning objective, ensuring that the learner receives appropriate and contextually relevant questions based on their progression

In the context of educational discussions and community-driven learning, platforms like Ed Discussion offer additional tools such as Bots++, which automates response generation and enhances engagement by providing real-time assistance and personalized feedback to students, making it a valuable feature for automating question and answer generation tasks.

2.2.4 Bloom’s Taxonomy and Cognitive Alignment

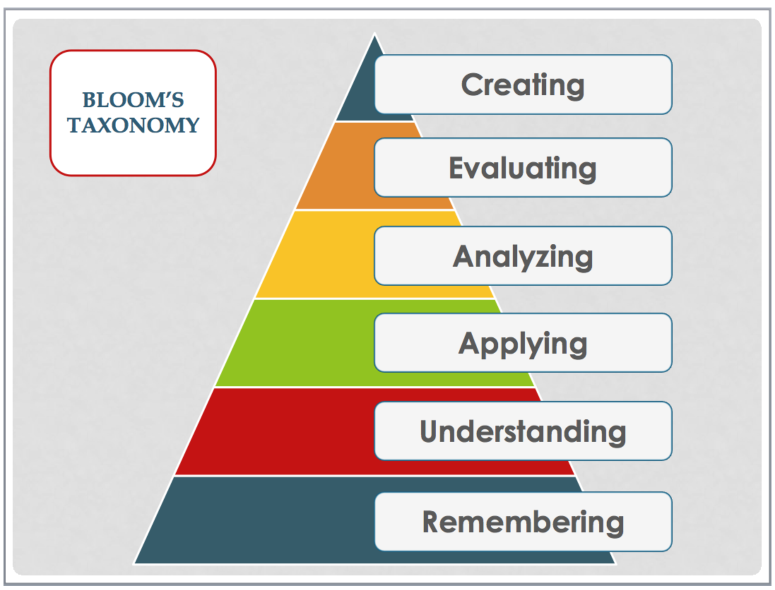

Figure 2.2: Bloom’s Taxonomy

Bloom’s Taxonomy is a framework that categorizes different levels of cognitive skills required for learning, ranging from basic knowledge recall to more complex activities such as application and analysis. The classification includes six levels as shown in the figure2.2:

- **Remembering**: Recall of information.
- **Understanding**: Understanding of concepts.
- **Applying**: Use of knowledge in new situations.
- **Analyzing**: Breaking down information into components.
- **Evaluating**: Making judgments based on criteria.
- **Creating**: Producing new or original work.

In AQG, aligning the questions with Bloom's Taxonomy allows the generated questions to match the learner's cognitive level and promote more effective learning. For instance, conceptual questions can be tied to the ”Understanding” level, while numerical questions might align with the ”Applying” or ”Analyzing” levels. The incorporation of this taxonomy ensures that the system generates questions that are educationally appropriate and aligned with the learners' developmental stages.

2.2.5 Feedback Loops for Continuous Optimization

A key aspect of the proposed AQG system is the incorporation of “feedback loops”. After generating a question, it is evaluated based on quality, relevance, and alignment with the learning objectives. A “Question Evaluation Agent” assesses the generated question, providing feedback on its clarity, structure, and depth. The “Question Refinement Agent” uses this feedback to modify and improve the question. This iterative process allows for continuous improvement, ensuring that the generated questions become more effective and aligned with the desired learning outcomes over time.

Chapter 3

Related Work

3.1 Fundamentals of Automatic Question Generation (AQG)

Before delving into the specific advancements in automatic question generation (AQG), it is essential to understand the evolution of research in this area. Traditional AQG methods relied on rule-based and template-driven approaches, which, while effective in generating structured questions, lacked flexibility and contextual adaptability. With the advent of deep learning and neural language models, AQG systems have witnessed significant improvements in generating diverse and meaningful questions. This chapter provides an overview of the foundational work in AQG, covering early approaches and recent advancements driven by large language models (LLMs).

3.1.1 Early Approaches: Rule-Based and Template-Based Methods

Early research in AQG primarily focused on rule-based and template-based techniques. These methods leverage syntactic structures and manually designed transformation rules to convert declarative sentences into interrogative ones. For example, systems would identify the subject and verb in a sentence and apply predefined rules to form questions. While effective for structured domains, such as factoid question answering and knowledge base query generation, these approaches struggled with complex sentence structures and domain adaptability [1, 2].

3.1.2 Transition to Machine Learning-Based AQG

With the rise of machine learning and NLP advancements, AQG systems began to adopt supervised learning and statistical models to improve question generation quality. Early approaches used sequence classification models to predict whether a sentence contained question-worthy information, followed by phrase transformation models to generate questions [3]. However, these approaches were still heavily dependent on feature engineering and linguistic heuristics, which limited their ability to generalize.

3.1.3 Large Language Models and Neural AQG

The advent of large-scale pre-trained language models (LLMs), such as BERT [4] and GPT-based models, revolutionized the field of AQG by eliminating the need for handcrafted templates and allowing models to learn question structures from large corpora. Unlike previous approaches, LLMs leverage bidirectional contextual understanding to generate diverse and high-quality questions that align with the semantics of the input text.

One key advantage of transformer [20]-based models like BERT is their ability to encode rich semantic representations, which significantly improves question relevance [4]. Moreover, fine-tuning BERT-based models on AQG-specific datasets further enhances their ability to generate coherent and domain-specific questions [5]. Later advancements, such as T5 (Text-to-Text Transfer Transformer), introduced a more flexible paradigm where the AQG task could be framed as a text-to-text transformation problem, allowing for greater versatility [6].

However, LLMs' reliance on large pre-trained datasets may introduce biases (e.g., cultural or gender stereotypes) into generated questions, and their "black-box" nature limits transparency in the question generation process, posing challenges for educational equity and accountability.

3.2 Neural Question Generation (NQG) and Recent Advancements

3.2.1 The Shift to Neural Networks for AQG

With the rapid advancement of deep learning, *neural question generation (NQG)* has become the dominant method in the field of automatic question generation. Traditional methods, which relied on rule-based systems and template matching, struggled to generate semantic-rich and diverse questions. The shift to neural networks, particularly *sequence-to-sequence (Seq2Seq)* models, enabled the automatic generation of more flexible and contextually relevant questions by learning directly from large-scale data. These neural approaches have shown significant improvements in the quality and diversity of generated questions, making them more suitable for complex and unstructured domains.The early advancements in neural question generation laid the groundwork for further research in this field, significantly influencing the direction of modern AQG techniques [7].

3.2.2 The Role of Large Language Models (LLMs) in NQG

The advent of large language models (LLMs) like BERT and GPT has revolutionized the approach to AQG, allowing for the pre-training of deep, bidirectional representations from vast corpora of text. BERT, in particular, has introduced a *masked language model (MLM)*, which helps generate more accurate and semantically coherent questions by conditioning on both left and right context, rather than just one directional context as in previous models like GPT. This has been instrumental in improving the semantic depth of generated questions, especially in domains that require contextual understanding.

For instance, BERT's pre-trained language representations have been successfully used for fine-tuning on AQG tasks, allowing it to generate high-quality conceptual and factual questions aligned with specific learning objectives (LOs) [4]. The transformer-based architectures used in LLMs, such as GPT-3, have made it possible to generate highly coherent and relevant questions without the need for explicit rule-based systems.

3.2.3 Fine-Tuning LLMs for AQG

Fine-tuning LLMs on domain-specific data has proven effective in improving question quality for specific educational contexts. T5 (Text-to-Text Transfer Transformer), a model that operates in an encoder-decoder framework, has been widely adopted for AQG tasks because of its versatility in handling various text generation tasks, including question generation. Fine-tuning these models on educational datasets enables the generation of questions that are aligned with Bloom's Taxonomy and target specific cognitive levels like understanding, application, and analysis [6].

One of the core challenges in AQG lies in content selection—identifying the most relevant part of a passage to ask about. Seq2Seq models equipped with attention mechanisms address this challenge by enabling the model to focus on important sections of the input text when generating questions. The attention mechanism not only enhances the model's ability to select relevant content but also improves the semantic alignment of the generated questions.

3.3 Optimization in Question Generation

3.3.1 The Challenge of Content Selection and Refinement

One of the key challenges in *Automatic Question Generation (AQG)* is determining which parts of a given text are question-worthy. Traditional methods struggled with this aspect, as they relied on syntactic parsing or heuristic rules to identify the relevant content. However, these approaches often failed to capture the semantic depth of the text and overlooked the context in which information should be questioned.

Recent advancements in neural networks, particularly attention-based models, have significantly improved the content selection process. For example, Seq2Seq models equipped with attention mechanisms [20] allow models to focus on the most relevant parts of the input text, thereby improving the semantic relevance of generated questions. These mechanisms work by giving more weight to important words or phrases in the passage when generating a question, ensuring that the generated question reflects the key ideas of the text.

Additionally, fine-tuned models like BERT and T5 have been successfully used to generate questions that are not only relevant but also contextually appropriate for the learning objective. By pre-training on large corpora and fine-tuning on domain-specific educational data, these models are able to refine the content they focus on during question generation. This results in a more accurate and targeted question generation process that better aligns with educational goals [4, 6].

3.3.2 Feedback Mechanisms and Question Refinement

While content selection is crucial, it is equally important to ensure the clarity, difficulty, and relevance of the generated questions. To address this, recent AQG systems have introduced automated feedback loops that assess the quality of the generated questions. These systems [5, 6] evaluate whether a question is clear, whether it targets the right cognitive skill, and whether it is appropriate for the learner's level of understanding.

Question evaluation agents work by assessing the draft questions based on feedback criteria such as:

- **Clarity**: Whether the question is easily understood by students.
- **Relevance**: Whether the question is aligned with the learning objective.
- **Difficulty**: Whether the question is suitable for the intended cognitive level, ranging from simple recall to higher-order thinking tasks such as analysis or application.

For example, in recent work on AQG systems, GPT-based models and BERT-based systems were combined with evaluation agents to automatically assess and refine the generated questions, ensuring that they met the educational criteria before they were finalized [5, 6].

In addition to automated feedback, human expert evaluations continue to play a significant role in ensuring that the generated questions are pedagogically sound. Studies have shown that feedback from teachers or subject matter experts helps refine questions that might otherwise be too vague or inappropriate for the learning context [2, 6].

3.3.3 Generating Answers Along with Questions

Another advancement in AQG is the simultaneous generation of questions and answers. Early AQG systems focused solely on question generation, but more recent models have recognized the importance of generating coherent answers that match the questions in both semantic and factual accuracy. This integration of question-answer pair generation improves the educational value of AQG systems.

Text-to-text transformer models, such as T5, have been used to generate question-answer pairs directly from a given text. These models leverage the encoder-decoder architecture to generate both the question and the corresponding answer in a single pass, which greatly simplifies the process for educators and students.

The integration of answer generation with AQG helps create interactive educational tools, such as automated tutoring systems, where students can practice both asking and answering questions, thereby improving their understanding of the material. This approach also ensures that the generated answers are directly aligned with the text and provide accurate, relevant information [5, 6, 10].

3.4 Multiple-Choice Question (MCQ) Generation

3.4.1 The Importance of MCQs in Education

Multiple-Choice Questions (MCQs) are among the most widely used assessment tools in education due to their efficiency and ease of automated scoring. MCQs are effective in measuring students' knowledge, comprehension, and even application of concepts when designed properly. However, manual creation of MCQs is a time-consuming process that demands careful crafting of the question stems, options, and distractors, along with maintaining balance in the difficulty and cognitive level of the questions.

Given their importance in educational assessments, automating the process of MCQ generation has been a significant area of research. Early works [8] in MCQ generation involved rule-based systems and template-based approaches, which were limited in their ability to generate contextually diverse and high-quality distractors. However, with the rise of neural networks and large language models (LLMs), the generation of high-quality MCQs has seen a significant improvement. LLMs like GPT-3 and T5 have shown promising results in generating MCQs that are contextually accurate and grammatically correct [8].

3.4.2 LLMs for MCQ Generation

Large Language Models (LLMs), such as GPT-3 and T5, have shown promising results in generating MCQs automatically from text. The use of these models, particularly fine-tuned versions, enables the generation of diverse question stems and plausible distractors. Recent work has demonstrated the potential of using LLMs to automatically generate MCQs that are both contextually accurate and grammatically correct.

In multiple-choice question generation, a key challenge lies in distractor generation, which involves creating plausible but incorrect options that challenge the student's understanding. LLMs like GPT-3 and GPT-4 have been shown to generate high-quality distractors by leveraging contextual information from the passage or subject area [11, 12]. These models generate distractors that are both contextually relevant and diverse, thus preventing the creation of overly simplistic or easily guessable answers.

3.4.3 Challenges in MCQ Generation

Despite the advancements, MCQ generation still faces several challenges:

- **Difficulty in Creating Plausible Distractors**: Even with powerful LLMs, generating high-quality distractors that are grammatically correct and semantically plausible remains a difficult task. Poorly designed distractors can lead to biased assessments, undermining the validity of the generated MCQs.

- **Ensuring Cognitive Alignment**: It is essential for MCQs to align with the cognitive level of the content being assessed, as per Bloom's Taxonomy. This requires models to generate MCQs that not only assess recall but also higher-order thinking skills, such as application and analysis. Recent advancements have attempted to address this by fine-tuning models like GPT-3 on educational datasets that contain a mix of question types across various cognitive levels [12].

3.4.4 Recent Advances in MCQ Generation

Recent advancements have leveraged transformer-based models, such as T5 and GPT-4, to create more sophisticated MCQs that assess a wider range of cognitive skills. For example, T5 has been used for multi-task learning, where it is trained to generate not only MCQs but also short-answer and open-ended questions. This allows for greater diversity in the generated assessments, enabling educational systems to use AI to automatically generate personalized tests for students, based on their learning progress and cognitive abilities [12].

In a recent study on MCQ generation using GPT-3 for programming education, the model was able to produce relevant and challenging MCQs with high discriminatory power. The MCQs generated by GPT-3 were shown to have similar or even better psychometric properties compared to human-crafted questions, particularly in terms of item difficulty and discrimination [11].

3.4.5 Enhancements in MCQ Generation using Fine-Tuning

Fine-tuning LLMs on domain-specific datasets has also been a critical development in improving the quality of generated MCQs. For instance, fine-tuning on pharmacology or medical education datasets has enabled LLMs like GPT-3 to generate domain-specific MCQs that are both contextually accurate and pedagogically sound [11]. Furthermore, feedback loops involving human expert evaluations of generated questions are being integrated into the process to ensure the validity and pedagogical alignment of the generated MCQs [12, 13].

3.5 Question-Answer Pair Generation

3.5.1 The Integration of Question and Answer Generation

The integration of question generation and answer generation has become a significant area of research in the field of *Automatic Question Generation (AQG)*. Traditionally, AQG systems focused primarily on question generation, with separate systems developed for answer extraction. However, recent advancements in deep learning have enabled the simultaneous generation of both questions and answers, creating a more cohesive and interactive educational tool.

The dual task of question-answer pair generation is essential for improving the learning experience by providing not only the questions but also the answers that align directly with the content. By generating both in a single pass, these systems can better ensure semantic coherence and contextual relevance between the generated question and its corresponding answer. This dual-generation approach enhances the utility of AQG systems, particularly in domains such as automated tutoring systems, where students require immediate feedback after answering a question.

3.5.2 Models for Generating Q-A Pairs

Recent research has focused on leveraging transformer-based models, such as T5 (Text-to-Text Transfer Transformer), to handle the task of question-answer pair generation. These models have demonstrated significant improvements in generating coherent, diverse, and contextually accurate question-answer pairs. In particular, T5's encoder-decoder architecture is well-suited for AQG tasks, as it can be framed as a text-to-text transformation problem where both questions and answers are treated as sequences to be generated based on the input passage.

One of the major breakthroughs in this area is the development of models that can generate question-answer pairs from unstructured text. In the context of educational materials, these models are trained to generate questions that are aligned with learning objectives, and to simultaneously produce relevant answers that provide correct information based on the text. This process ensures that the generated content is both educationally valuable and contextually accurate, which is crucial in automated assessment systems and interactive learning environments [10].

3.5.3 Question-Answer Pair Generation in Educational Systems

The integration of question-answer generation is particularly important in adaptive learning systems and automated tutoring environments. These systems not only generate questions based on a learner's progress but also provide immediate feedback through the automatic generation of corresponding answers. This approach is beneficial in self-paced learning environments, where students can interact with the system and receive real-time answers to their questions, without requiring constant teacher supervision.

Automated question-answer pairs also play a critical role in exam preparation, as they allow students to practice and reinforce concepts without the need for manually created assessments. By generating a large number of questions and answers, these systems help create personalized learning experiences, adapting to the student's individual needs and helping them focus on areas that require further improvement.

3.5.4 Evaluation of Q-A Pair Generation Models

Evaluating the effectiveness of Q-A pair generation models involves assessing both the quality of the questions and the accuracy of the answers. One of the primary challenges in evaluating Q-A pairs is ensuring that the generated answers are not only factually correct but also contextually appropriate. Traditional evaluation metrics, such as BLEU, ROUGE, and METEOR [3], have been widely used to evaluate the fluency and relevance of generated text. However, they may not fully capture the semantic correctness or educational value of the generated question-answer pairs.

To address this issue, recent studies have introduced human expert evaluations for assessing the pedagogical quality of the generated content. For example, studies have shown that teacher evaluations are crucial in ensuring that the generated answers are aligned with the intended learning objectives and are accurate in providing the correct information. These human evaluations often involve assessing the clarity, relevance, and appropriateness of the generated question-answer pairs within an educational context [2, 5].

In addition to human evaluations, automatic evaluation methods have been explored using large language models (LLMs), such as GPT-4, to assess the quality of generated content. These methods have shown potential in providing feedback on the pedagogical alignment of the generated questions and answers, though they are still being refined for wider application [10, 12].

3.6 AI-Based Learning Pathways and Content Generation

3.6.1 The Role of AI in Personalized Learning

One of the key areas where AI has made significant strides is in the creation of personalized learning pathways. Traditional education systems often use one-size-fits-all approaches, where all students are exposed to the same material at the same pace. However, research has shown [1] that students have varying levels of prior knowledge, learning preferences, and cognitive abilities. Personalized learning addresses these differences by tailoring the learning experience to the individual needs of each student. AI-powered systems have the ability to track a student's progress, identify areas where they need more support, and generate custom-tailored content to help them succeed.

AI-based systems are able to dynamically adapt the learning content, questions, and difficulty levels based on a student's performance [1]. For instance, automated question generation (AQG) can generate questions aligned with the student's current learning progress and cognitive level. Moreover, AI can help in identifying and suggesting the next best steps in a student's learning path by analyzing their interaction with the system.

Recent studies have highlighted the potential of large language models (LLMs), such as GPT-4, in personalized education. These models can generate context-specific questions that adapt to the learner's level, ensuring that the material is challenging but not overwhelming. The ability of LLMs to generate high-quality content in real-time allows for a more engaging and interactive learning environment. By fine-tuning these models on domain-specific datasets, they can generate learning materials tailored to specific courses, subjects, or even individual learners' progress [1, 12].

3.6.2 AI for Adaptive Content Generation

Beyond generating personalized learning questions, AI is increasingly being used to create adaptive learning systems that generate learning content dynamically based on student performance. Adaptive learning platforms use algorithms to adjust the difficulty and content based on students' strengths and weaknesses. These platforms analyze the student's interaction history to decide what content to present next, ensuring that the student is always exposed to material that is appropriately challenging.

For example, systems like Intelligent Tutoring Systems (ITS) [9] use AI to continuously assess student responses and dynamically adjust the learning material. AI can generate real-time feedback to help students progress through their learning journey, ensuring they are constantly challenged at the right level.

Recent work has shown that AQG models can be integrated with these adaptive learning platforms to generate content that aligns with students' learning needs. For instance, question generation models fine-tuned on students' previous responses can create questions that help address their learning gaps, providing targeted assessments that help reinforce their understanding of weak areas [1, 13].

3.6.3 AI-Powered Assessment and Evaluation

AI's ability to automate the process of assessment generation also extends to the creation of adaptive quizzes and learning evaluations. These assessments can be tailored to test different cognitive skills, from basic recall to higher-order thinking, in line with Bloom's Taxonomy. By generating personalized assessments, AI can help identify the cognitive areas where a student needs further development.

For example, Bloom's Taxonomy divides cognitive skills into categories such as remembering, understanding, applying, and evaluating. AI-powered AQG systems can generate questions that assess each of these cognitive levels, ensuring that the learning content is well-balanced and cognitively appropriate for the learner's stage of development. This can help students progress through different cognitive levels and gain a deeper understanding of the subject matter.

Furthermore, AI-based evaluation systems can provide real-time feedback on the student's performance, giving both students and educators insights into areas that need attention. For instance, systems using models like GPT-4 can generate personalized assessments, along with detailed feedback on the answers, offering insights into areas of improvement [1, 13].

3.7 Retrieval-Augmented Question Generation (RAG) and Case-Based Reasoning (CBR)

3.7.1 Introduction to Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is an emerging paradigm that combines the strengths of both retrieval-based and generation-based approaches in natural language processing (NLP). While traditional generative models like GPT-3 can produce text based on learned representations, they often lack access to external knowledge, which can lead to hallucinations or factually incorrect responses. RAG addresses this challenge by incorporating external knowledge into the generation process, effectively improving the accuracy and factual correctness of the generated content.

In question generation tasks, RAG-based models first retrieve relevant information from an external knowledge base or corpus and then use that information to generate a question. This approach ensures that the questions generated are contextually grounded and factually accurate, as the model has access to a broad set of retrieved facts to base its questions on. For example, a question about a historical event would be generated by retrieving information from a database of historical events, ensuring the question's accuracy.

One of the primary advantages of RAG is its ability to enhance the contextual relevance of the generated questions. This is particularly important in fields like education and legal question answering, where accurate and relevant questions are crucial for assessing a learner's understanding. Recent studies have shown that RAG-based models outperform purely generative models in domains that require precise and up-to-date knowledge [15].

3.7.2 Case-Based Reasoning (CBR) for Enhancing Retrieval in RAG

Case-Based Reasoning (CBR) is a methodology that uses past cases (or experiences) to solve new problems by finding similar cases and adapting their solutions. In the context of RAG-based AQG, CBR can play a crucial role in enhancing the retrieval process by organizing the knowledge in a way that allows for more effective matching of relevant cases to the query. This method helps improve the quality and relevance of the retrieved information, leading to more accurate and contextually appropriate questions.

In the legal domain, CBR has been effectively integrated with RAG to improve the generation of legal questions and answers. The CBR-RAG system first retrieves relevant legal cases from a case database, then uses these cases as context for generating more relevant legal questions. This method allows the legal question-answering system to not only retrieve relevant laws or precedents but also to generate questions based on the most similar historical cases, making the questions more contextual and factually accurate [15].

3.7.3 Integrating CBR and RAG in AQG Systems

Integrating CBR and RAG into AQG systems offers several advantages. By using retrieved cases as contextual grounding, the system can ensure that the questions are not only grammatically correct but also semantically relevant. Additionally, by applying case-based reasoning, the system can generate questions that span multiple cognitive levels, from fact-based questions to application or analysis questions.

This integration is especially beneficial in knowledge-intensive domains, such as law, medicine, and engineering, where precise and contextually grounded questions are essential for accurate assessments. For instance, in legal AQG, the system retrieves relevant legal precedents and generates questions that test the understanding of legal principles. Similarly, in medical education, RAG-based AQG systems can generate questions about clinical cases by retrieving medical records or diagnostic guidelines.

The RAG-CBR integration has also been successfully tested in question answering (QA) systems, where the retrieval of relevant information significantly improves the accuracy and specificity of the generated answers. By retrieving the most relevant information and using that to generate both questions and answers, these models have shown to outperform standard generative models in terms of factual accuracy and contextual relevance [14, 15].

3.8 Synthetic Data Generation for Training AQG Models

3.8.1 The Need for High-Quality Training Data in AQG

The development of high-quality automatic question generation (AQG) systems requires large amounts of training data, which traditionally comes from question-answer pairs. However, manually curating these datasets is both time-consuming and resource-intensive, making it challenging to scale up for diverse educational domains. One way to address this problem is through synthetic data generation, where automated systems generate question-answer pairs based on unstructured or semi-structured data, reducing the reliance on manual data creation.

In the context of AQG, synthetic data generation can help create training sets that include a wide variety of questions and answers, covering different cognitive levels (as per Bloom's Taxonomy) and domain-specific topics. This is especially important for adaptive learning systems, where the questions must be tailored to the learner's current level of understanding. By using large language models (LLMs) like GPT-3 or T5, we can generate vast amounts of data that can be used to fine-tune AQG models, enabling them to produce high-quality questions across different domains and contexts.

3.8.2 Using LLMs for Synthetic Question-Answer Pair Generation

Recent studies have shown that LLMs can be effectively used to generate synthetic question-answer pairs for training AQG systems. One of the key advantages of using LLMs for synthetic data generation is their ability to understand the semantic content of a given passage and generate contextually relevant questions. For example, models like GPT-3 and T5 can process input text and generate both questions and answers in a seamless manner, based on the content of the text.

In the case of GPT-3, the model can be prompted with a passage and asked to generate a set of questions along with their corresponding answers. These generated pairs can then be used to train AQG systems, providing valuable data that would otherwise be difficult to generate manually. Furthermore, these models can be fine-tuned on domain-specific data to generate questions that align with particular educational content, ensuring that the questions are both relevant and pedagogically sound.

3.8.3 Data Augmentation and Diversity in Generated Questions

Another challenge in AQG is ensuring the diversity of the generated questions. Synthetic data generation through LLMs can help address this issue by producing a wide variety of questions that span different cognitive levels (e.g., recall, application, analysis) and question types (e.g., multiple-choice, open-ended, fill-in-the-blank). By training AQG models on these synthetic datasets, we can significantly improve the diversity and coverage of the questions generated.

One advantage of using synthetic data generation is that it can provide augmented datasets for areas where labeled data is scarce. For example, in technical fields like engineering or medicine, where creating domain-specific question sets can be particularly difficult, LLMs can generate a large volume of diverse questions that can be used to train domain-specific AQG models. This helps overcome the limitations of traditional data collection methods and accelerates the development of high-quality educational tools.

3.8.4 Evaluation of Synthetic Question-Answer Pairs

Evaluating the quality of synthetic question-answer pairs is crucial to ensure that the generated data is useful for training AQG models. Human evaluations have been widely used to assess the relevance and pedagogical quality of generated questions and answers. However, these evaluations are often time-consuming and expensive. To address this, some recent studies have used automatic evaluation metrics like BLEU, ROUGE, and METEOR to assess the fluency and semantic correctness of the generated question-answer pairs. While these metrics provide useful insights into the syntactic quality of the generated content, they may not fully capture the pedagogical value of the questions.

In addition to traditional evaluation methods, LLMs can also be used to assess the quality of the generated data by comparing the generated questions with existing human-crafted questions. For example, models like GPT-4 have been used to assess the validity of the generated question-answer pairs and provide feedback on how well they align with the intended learning objectives [14].

3.9 Evaluation Methods for AQG Models

3.9.1 The Importance of Evaluation in AQG

Evaluation plays a critical role in the development of Automatic Question Generation (AQG) systems. Since AQG models aim to generate questions that are not only grammatically correct but also pedagogically valuable, it is essential to evaluate the quality of generated questions on multiple levels. Traditional automatic evaluation metrics such as BLEU, ROUGE, and METEOR have been used to assess the fluency and semantic relevance of generated text. However, these metrics have limitations when it comes to evaluating the educational quality of the questions, as they focus on surface-level text similarity rather than on the pedagogical value of the questions.

In recent years, human expert evaluations have been widely adopted to assess the relevance, clarity, and educational alignment of generated questions. Additionally, there has been a push towards developing more specialized evaluation methods tailored to AQG tasks. These methods assess the cognitive alignment of the generated questions with frameworks like Bloom's Taxonomy, as well as the question difficulty and its alignment with the intended learning objectives.

3.9.2 Traditional Evaluation Metrics

While human evaluations are considered the gold standard for evaluating AQG systems, traditional automatic evaluation metrics still play a crucial role in assessing the basic fluency and semantic coherence of generated questions. Commonly used metrics like BLEU, ROUGE, and METEOR compare the generated question to reference questions and provide a score based on text overlap and similarity. These metrics can give a quick overview of the quality of the generated questions in terms of grammar and vocabulary, but they often fail to capture the pedagogical alignment or cognitive complexity of the questions.

For instance, BLEU is commonly used to evaluate machine-generated text by measuring the overlap between n-grams of the generated and reference texts. While it can measure basic text similarity, it does not account for whether the question truly assesses the intended learning objective or whether the distractors in multiple-choice questions (MCQs) are plausible. Thus, while automatic metrics are useful for measuring fluency, they fall short in evaluating the educational value of AQG models [2, 5].

3.9.3 Human Expert Evaluations

Human expert evaluations remain the most reliable method for assessing AQG models, as they take into account both the pedagogical quality and contextual appropriateness of generated questions. In these evaluations, domain experts (such as teachers or subject matter experts) review the questions generated by the AQG system and provide feedback on whether the questions are relevant, clear, and aligned with the intended learning objectives.

For example, in a study on MCQ generation using GPT-3, human experts were asked to assess the quality of the questions in terms of difficulty, clarity, and alignment with Bloom's Taxonomy. These evaluations are particularly important in fields like education and legal question answering, where generating questions that are pedagogically sound is just as important as generating grammatically correct questions. Recent research has shown that expert feedback is crucial for improving the validity of machine-generated questions, especially in specialized domains like medicine and law [12, 15].

3.9.4 Evaluation Using Bloom's Taxonomy

One of the key frameworks used in evaluating the cognitive alignment of generated questions is Bloom's Taxonomy, which classifies questions into six cognitive levels: remembering, understanding, applying, analyzing, evaluating, and creating. AQG systems that are aligned with Bloom's Taxonomy generate questions that assess a range of cognitive skills, from basic recall of facts to more advanced skills like problem-solving and analysis.

For example, an AQG system might generate simple factual questions (e.g., ”What is the capital of France?”) that test remembering, or it might generate higher-order questions (e.g., ”How would you apply the principles of physics to solve this problem?”) that test applying and analyzing. Evaluating how well an AQG model generates questions across these levels is crucial for understanding its effectiveness in assessing students' knowledge and cognitive abilities. Recent studies have used Bloom's Taxonomy as a benchmark for evaluating AQG systems, helping ensure that the generated questions are not only relevant but also aligned with the intended learning outcomes [9, 12].

3.9.5 Automatic Evaluation using Large Language Models (LLMs)

As AQG models become more sophisticated, there has been growing interest in using large language models (LLMs), such as GPT-4, to assess the quality of generated questions and answers. LLMs can be used to evaluate AQG outputs by mimicking human evaluation processes and providing insights into the relevance, clarity, and pedagogical quality of the generated content. For example, GPT-4 has been used to evaluate MCQs and other types of generated questions, offering a statistical evaluation based on relevance and factual accuracy.

These LLMs can also assist in providing feedback to AQG systems, helping improve question clarity and educational value. While LLM-based evaluations still need further refinement to fully match human judgment, they provide a promising avenue for automating the evaluation process and making AQG systems more scalable [10, 16].

3.10 Future Directions in AQG

3.10.1 Addressing Current Limitations in AQG Systems

While automatic question generation (AQG) has seen significant advancements, several limitations persist, which require further research and innovation. One of the primary challenges in AQG is the generation of high-quality distractors in multiple-choice questions (MCQs). Even with the use of large language models (LLMs) like GPT-3 and T5, generating plausible distractors remains difficult [8]. Poorly designed distractors can lead to biased assessments, reducing the validity of the generated questions. To address this, future work should focus on:

- Developing more sophisticated retrieval mechanisms to improve distractor generation by using external knowledge sources, such as semantic networks and contextual databases.
- Implementing context-sensitive models that generate diverse distractors based on the content's semantic structure rather than relying solely on language patterns [12, 13].

Another limitation is the difficulty in aligning questions with Bloom's Taxonomy or other cognitive frameworks. Although AQG models are improving in generating fact-based questions, they often struggle with higher-order cognitive tasks like analysis, synthesis, and evaluation. To overcome this challenge:

- Improved fine-tuning strategies are needed to train AQG models on cognitive-level specific datasets that can guide the generation of questions targeting different cognitive levels.
- Hybrid models that combine retrieval-based generation with deep learning could be explored to ensure that questions not only adhere to Bloom's Taxonomy but also align with the intended learning objectives [9, 10].

3.10.2 Enhancing Domain-Specific AQG Systems

Another area for future development is the generation of domain-specific questions. While general-purpose AQG models have shown promising results, the need for specialized systems for fields like medicine, law, and engineering is still prominent. Domain-specific AQG systems can enhance educational content by providing questions that are both contextually accurate and aligned with professional standards. These systems can be achieved by:

- Fine-tuning models like GPT-3 or T5 on specialized datasets, such as clinical case studies, legal case reports, and engineering principles.
- Integrating domain-specific knowledge through ontology-based systems or semantic web technologies to provide precise and contextually relevant questions [12, 15].

3.10.3 Integration with Adaptive Learning Systems

As adaptive learning systems gain traction, integrating AQG systems with these platforms will be crucial for creating personalized educational experiences. Adaptive learning platforms adjust the learning path based on students' strengths and weaknesses, and AQG systems can play a vital role by dynamically generating questions tailored to the learner's current level of understanding. Future research should focus on:

- Integrating AQG models with real-time learner data, such as responses and performance history, to generate personalized assessments that target areas for improvement.
- Developing real-time feedback mechanisms where AQG systems generate both questions and answers based on learner interactions with the platform, ensuring that the learning experience remains engaging and effective [9, 10].

3.10.4 Leverage Few-Shot Learning and Prompt Engineering

One of the promising approaches to improving AQG models is through few-shot learning and prompt engineering. Recent studies have shown that LLMs like GPT-3 and GPT-4 can achieve high performance on AQG tasks with minimal fine-tuning by using carefully designed prompts [17]. For example, GPT-4 demonstrates stronger performance in generating high-order cognitive questions (e.g., analysis-level questions) with a 12% higher score on Bloom's Taxonomy alignment compared to LLaMA3-8B in preliminary tests, though it requires more computational resources. This opens up exciting opportunities for:

- Developing prompt engineering protocols that allow AQG systems to generate high-quality questions without the need for large-scale training datasets, significantly reducing the computational cost.
- Exploring few-shot learning techniques to adapt AQG models to new topics or domains with minimal training, thus enabling the generation of domain-specific questions in areas where labeled data is scarce [18, 19].

Chapter 4

Methodology

In this study, we propose a modular framework for automated question generation (AQG) that uses large language models (LLMs) like LLaMA3 to generate high-quality questions and answers from educational text. The framework consists of five key agents, each responsible for a specific task within the process: Material Analysis and Segmentation Agent, Question Generation Agent, Question Evaluation Agent, Question Refinement Agent, and Answer Generation Agent. These agents are designed to collaborate, ensuring the generated content aligns with the educational objectives while maintaining high quality and relevance.

4.1 System Architecture and Workflow

As shown in Figure 4.1, the system consists of several interconnected agents that work together to generate high-quality questions and answers.

4.1.1 Material Analysis and Segmentation Agent

The first step in our system involves the Material Analysis and Segmentation Agent, which takes raw educational content, such as a textbook chapter, and divides it into manageable segments. The process is shown in Figure 4.2.This agent also extracts key learning objectives and associated keywords from the segmented text. This segmentation is crucial as it allows the following agents to operate on more focused, relevant sections of text, ensuring that the generated questions align closely with the educational goals. For example, in our implementation, the content of a textbook chapter on encryption methods, like the Caesar Cipher, would be divided into segments focused on different encryption techniques, each associated with specific objectives like understanding the concept of a cipher, applying the cipher to encrypt text, and solving encryption challenges.

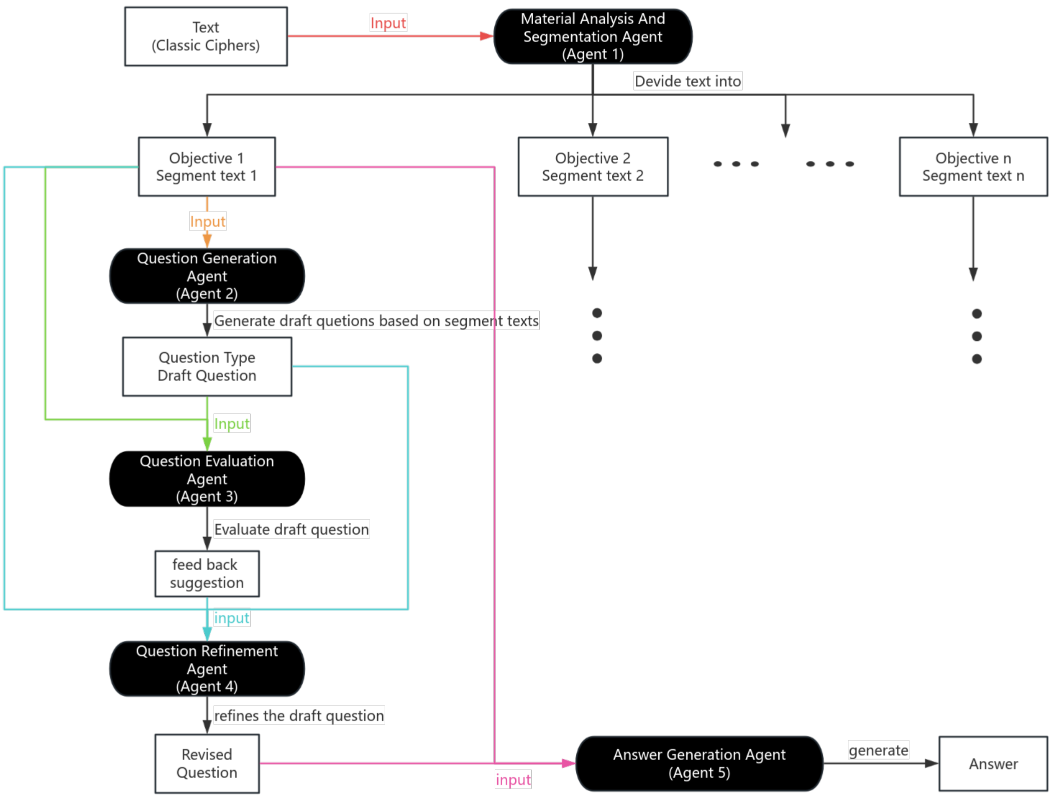

Figure 4.1: Model flow chart

4.1.2 Question Generation Agent

Once the material is segmented, the Question Generation Agent comes into play. The visualization process is shown in Figure 4.3. This agent is responsible for generating the initial draft questions based on the segments and the learning objectives derived by the Material Analysis Agent. The agent first determines the appropriate type of question (conceptual, numerical, or application-based) based on the content and the objective, and then constructs the draft question. This step is critical as it ensures that the generated questions are contextually relevant and aligned with the learning goals. For instance, if the objective is to understand the application of the Caesar Cipher, the agent may generate an application-based question like, ”Given the ciphertext ’AWW’ and a key of 2, apply the Caesar Cipher to find the original plaintext.”

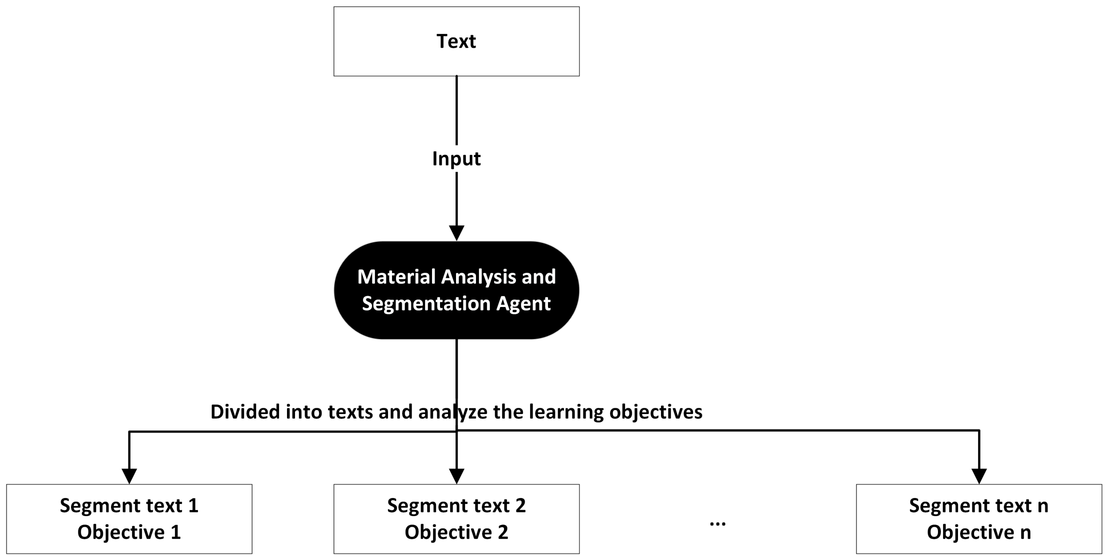

Figure 4.2: Material Analysis and Segmentation Agent Workflow

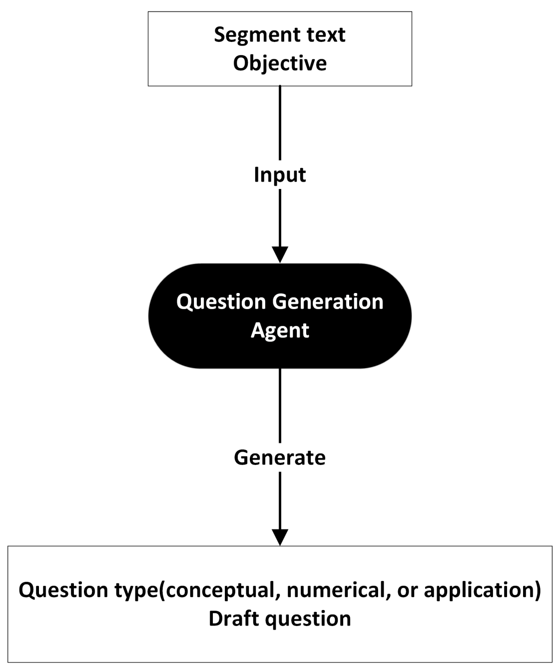

Figure 4.3: Question Generation Agent Workflow

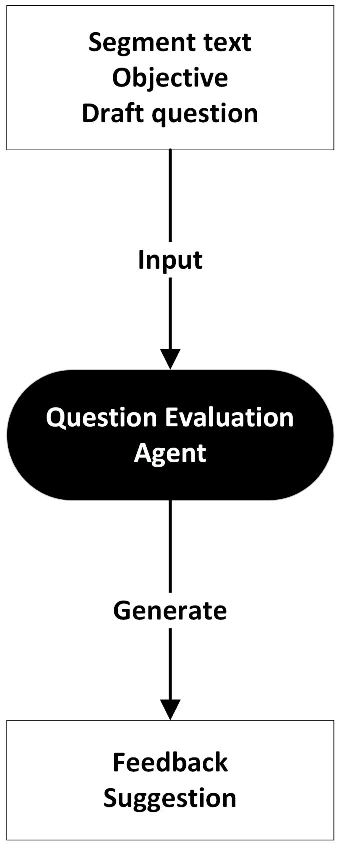

Figure 4.4: Question Evaluation Agent Workflow

4.1.3 Question Evaluation Agent

Following the question generation, the Question Evaluation Agent evaluates the quality of the draft question. The process is shown in Figure 4.3. This agent assesses the clarity, difficulty, and relevance of the question based on predefined criteria. It also provides feedback and suggests improvements, such as rephrasing the question to make it more challenging or to better target specific learning objectives. For example, if the generated question is too vague or does not fully address the intended learning objective, the evaluation agent might suggest a more specific version that asks for more detailed reasoning.

4.1.4 Question Refinement Agent

After receiving the feedback, the Question Refinement Agent refines the question. The process is shown in Figure 4.5. This agent incorporates the suggestions from the evaluation agent,

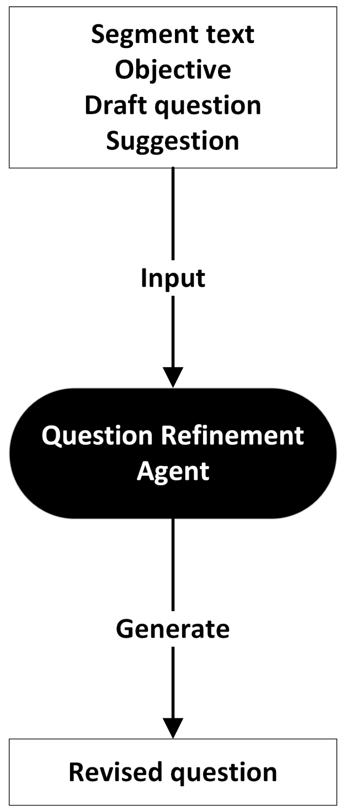

Figure 4.5: Question Refinement Agent Workflow

improving the clarity and depth of the question. For example, if the evaluation agent suggests that the question lacks clarity, the refinement agent will rephrase the question to ensure it is more precise. This iterative process helps in producing questions that are well-structured, challenging, and educationally valuable.

4.1.5 Answer Generation Agent

The final step in the process is handled by the Answer Generation Agent, which generates the corresponding answers to the refined questions. The visualization process is shown in Figure 4.6. This agent leverages the original text segment and the keywords extracted earlier to craft accurate and detailed responses. It ensures that the answers are comprehensive and logically aligned with the content, providing students with clear and informative feedback. For example, for a question on the application of the Caesar Cipher, the answer would explain the encryption process in detail, illustrating how the cipher shifts letters based on the given key.

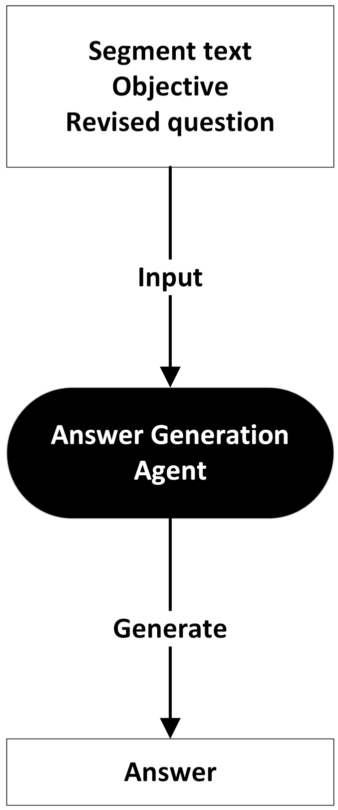

Figure 4.6: Answer Generation Agent Workflow

4.2 Choice of Language Model: LLaMA3

The backbone of this system is the large language model LLaMA3 [21], which we selected for its robust natural language processing capabilities. LLaMA3, developed by Meta, is designed to handle a wide range of text generation tasks, including question answering, summarization, and translation. Its ability to process vast amounts of data and generate coherent, contextually relevant text makes it ideal for AQG tasks where high-quality, domain-specific content is required.

LLaMA3 offers several advantages over other LLMs. Its architecture allows for effective fine-tuning on specific educational datasets, making it highly adaptable to the nuanced requirements of AQG. Moreover, LLaMA3 has been shown to outperform other models like GPT in certain tasks, particularly in terms of factual accuracy and generating informative responses [21]. This makes it a suitable choice for generating educational content, where accuracy and alignment with learning objectives are paramount.

By using LLaMA3, we benefit from its advanced capabilities in text generation and understanding. It is fine-tuned to ensure that the generated questions and answers are both relevant and insightful, which is crucial when dealing with complex topics like cryptography and encryption methods. The LLaMA3 model helps ensure that each component of the system, from text segmentation to question refinement, produces high-quality output that enhances the overall learning experience.

4.3 Model Code and Prompt Design

The success of this modular system heavily relies on carefully crafted prompts that guide each agent's task. These prompts ensure that the agents produce relevant and high-quality outputs, which are critical for achieving the desired educational outcomes.

- **Material Analysis and Segmentation**:The prompt for this agent instructs the model to divide the text into segments and identify learning objectives. The prompt is designed to extract both the educational content and the key keywords, which help inform the subsequent question generation.
- **Question Generation**:The prompt for this agent focuses on determining the type of question (conceptual, numerical, or application-based) and generating the draft question. The model is provided with examples for each question type to ensure it produces the appropriate kind of question for each segment and objective.
- **Question Refinement**:The refinement prompt directs the model to modify the draft question based on the feedback from the evaluation agent, ensuring that the final question is both clear and challenging.
- **Answer Generation**:The answer generation prompt instructs the model to create a precise and accurate answer for each refined question, based on the educational content from the text.

By combining these carefully designed prompts with the powerful capabilities of LLaMA3, the system ensures that each step in the question generation process contributes to the overall goal of producing high-quality educational content.

Each agent's prompt contains a role description, task description, output format, and example input and output. Full prompts are provided here:

- **Material Analysis and Segmentation Agent Prompt:**

```
You are an AI assistant that analyzes an educational text.

Tasks:
```

```
1. Divide the text into segments, each corresponding to a
central content.(Each segment is a piece of the original text)
2. Infer the teaching objectives from the segments.
3. For each segment, extract key concepts/keywords.

Attention:
1. Do not use unformatted punctuation marks such as
'*' in your output.
2. The output of the sample should not appear in the output.
3. After output, output <END> on a new line.

Output Format:
Number of Objectives: n
Objective i: the extracted objective
Segment i:
the text segment
Keywords i:
comma-separated keywords
<END>

Example Input:
An early cipher (circa 45 BC, when we were all little kids).
Key in the range (1 .. 25). A → C; B → D; C → E; ... Z → B etc
when key = 2. A → E; B → F; C → G; ... Z → D etc when key = 4.
y = (x + key) mod 26.

Examples: YOU → AQW (key = 2); YOU → CSY (key = 4). A = 65,
the ASCII value of the character A. Y = 89, the ASCII value
of the character Y. Assume key = 4. Y then becomes
65 + (((89-65)+4) mod 26). 89-65 = 24;

This says Y is the 24th character in the alphabet.
Add 4 (mod 26): you get 2 - i.e., the 2nd character in
the alphabet: C (A being the 0th character). To get the
ASCII value of C, add 65 to 2. C = 67.

Example Output:
Number of Objectives: 2
Objective 1: Understand the Caesar Cipher encryption method
```

```
Segment 1:
• An early cipher (circa 45 BC, when we were all little kids).
• Key in the range (1 .. 25)
• A → C; B → D; C → E; ... Z → B etc when key = 2
• A → E; B → F; C → G; ... Z → D etc when key = 4
• y = (x + key) mod 26
Examples: YOU → AQW (key = 2); YOU → CSY (key = 4)
Keywords 1:\nCaesar Cipher, encryption, key, mod 26, examples
Objective 2: Learn how to implement the Caesar Cipher
Segment 2:
• A = 65, the ASCII value of the character A.
• Y = 89, the ASCII value of the character Y.
• Assume key = 4
• Y then becomes 65 + (((89-65)+4) mod 26)
• 89-65 = 24; This says Y is the 24th character in
the alphabet
• Add 4 (mod 26): you get 2 - i.e., the 2nd character
in the alphabet: C (A being the 0th character).
• To get the ASCII value of C, add 65 to 2. C = 67.
Keywords 2:\nimplementation, ASCII, key, Caesar Cipher,
encryption, decryption\n
<END>

Text:{text}
```

- **Question Generation Agent Prompt**:

```
You are an expert educational AI assistant.

Your tasks:
1. Determine which of the following three types of
questions would be most beneficial for achieving the
teaching objective, based on the provided text segment,
objective, and keywords.
2. After determining the question type, generate an
initial draft question that matches the chosen type
and helps achieve the learning objective.

Question Types and Their Standards:
- Conceptual Questions: Focus on understanding concepts,
```

```
principles, or the reasoning behind processes. Use
this when the objective and text emphasize understanding
'what', 'why', or 'how' something works conceptually.
example of Conceptual Questions: What is the range of keys
in a Caesar Cipher?
- Numerical Questions: Involve direct numeric computations,
ASCII values, letter positions, frequencies, percentages,
or key values. Use this when the text and keywords
strongly hint at numeric relationships or calculations.
example of Numerical Questions: What is the ASCII value
of the letter 'Y'?
- Application Questions: Involve applying the learned
concept to a concrete scenario, such as decrypting a
given ciphertext, demonstrating brute-forcing, or using
frequency analysis to determine a plaintext letter. Use
this if the objective suggests practical usage or
scenario-based tasks. If there are mathematical
formulas in the text, priority should be given to
application question
The generated questions should demonstrate specific
application scenarios, example of Application Questions:
Given the ciphertext 'AWW' and a key of 2, apply the
Caesar Cipher to find the original plain.

Attention:
1. All questions are Q&A type, no multiple-choice, no fill-in.
2. Output strictly according to format, specially attention
to the use of newline characters
3. Do not output explanations beyond the required format
as much as possible.

Output Format:
Question Type: [Question Type]
<END>

Draft Question: [Draft Question]
<END>

Example Input:
```

```
Objective: Understand the concept of Caesar Cipher
Keywords: cipher, shift, alphabet
Segment:
This segment explains the Caesar Cipher and why
shifting letters by a given key changes their position in
the alphabet.
Example Output:
Question Type: Conceptual Questions
<END>

Draft Question: Explain how the Caesar Cipher uses modular
arithmetic (mod 26) to shift letters.
<END>

Now, based on these standards, determine the best question
type and output it.
Then generate a draft question of that type aligned with the
objective and the segment.

Objective: {objective}
Keywords: {', '.join(keywords)}
Segment:
{segment_text}
```

- **Question Evaluation Agent Prompt**:

```
You are an AI teaching expert.

Your tasks:
Evaluate the draft question in terms of clarity,
difficulty,
and relevance, then provide one suggestion.

Output format:
Feedback: [Feedback]
Suggestion: [Suggestion]
<END>

Example Input:
```

```
Draft Question: Explain how Caesar Cipher works.
Example Output:
Feedback: The question is clear and relevant.
Suggestion: Add more specificity about the shifting
process.
<END>

Now start your task
Objective: {objective}
Keywords: {', '.join(keywords)}
Draft Question:{question}
Segment text:
{segment_text}
```

- **Question Refinement Agent Prompt**

```
You are an AI assistant.

Your tasks:
Refines the draft question based on the suggestion of the
evaluation agent and the corresponding segment text.

attention:
1. Output strictly according to format, specially attention
to the use of newline characters
2. Do not output explanations beyond the required format.

Output format:
Revised Question:[Revised Question]
<END>

Example Input:
Draft Question: Explain Caesar Cipher.
Suggestion: Specify how the key affects letter shifting.
Segment text:
The Caesar Cipher is one of the simplest and most well-known
encryption techniques. It works by shifting the letters
of the plaintext by a fixed number of positions down or up
the alphabet. The key in a Caesar Cipher determines the
number of positions each letter in the plaintext will be
```

```
shifted. For example, if the key is 3, the letter "A"
becomes "D", "B" becomes "E", and so on.
Example Output:
Revised Question: How does the chosen key affect the letter
shifting process in a Caesar Cipher?
<END>

Now start your task.
Draft Question:
{draft_question}
Suggestion:{suggestion}
Segment text:
{segment_text}
```

- **Answer Generation Agent Prompt**

```
You are an AI assistant

Your task:
Generating an answer based on the provided segment
and the question.

Output format:
Answer: [Answer]
<END>

Attention:
1. Output strictly according to format, specially attention
to the use of newline characters.
2. Do not output explanations beyond the required format.

Example Input:
Question: How does Caesar Cipher shift letters?
Segment:
The Caesar Cipher is an encryption technique that shifts
letters of the alphabet by a fixed number of positions,
determined by the key. Each letter is mapped to its
corresponding position in the alphabet
(e.g., A = 0, B = 1, etc.),
```

```
and the key is added to this position. If the shift moves
past "Z", it wraps around to the beginning of the alphabet.
For instance, if the key is 3, then "A" becomes"D",
"B" becomes "E", and so on. This simple method creates a
substitution cipher that can easily be decrypted if the
key is known.
Example Output:
Answer:
The key in the Caesar Cipher determines the number of
positions each letter in the plaintext is shifted.
For example, if the key is 3, each letter is moved
3 positions forward in the alphabet. If the shift
exceeds "Z", it wraps around to the beginning of the
alphabet. The key essentially controls the shift amount
for each letter.
<END>

Now start your task.
Question:{question}
Segment:
{segment_text}
```

4.4 An Example

- **Step 1: User Input**

  The Internet is a computer network that interconnects billions of computing devices throughout the world. Not too long ago, these computing devices were primarily traditional desktop PCs, Linux workstations, and servers that store and transmit information such as Web pages and e-mail messages. Increasingly, however, nontraditional Internet "things" such as laptops, smartphones, tablets, TVs, gaming consoles, thermostats, home security systems, home appliances, watches, eye glasses, cars, traffic control systems, and more are being connected to the Internet. These devices are interconnected using various communication protocols, which allow for the efficient transfer of data across vast distances. The Internet infrastructure relies on numerous physical media like fiber optics, coaxial cables, and wireless transmission methods to maintain its speed and reliability. Additionally, the Internet is built on the foundation of several key protocols, including TCP/IP, which ensures that data packets are correctly routed from one device to another. Furthermore, the Internet allows for the creation and deployment of applications that are accessible by users worldwide. These applications range from simple web browsing to complex cloud computing services. Understanding the components and working of the Internet is crucial for understanding how data is exchanged across the globe.

- **Step 2: Material Analysis And Segmentation Agent** analyzes the input text and divides it into segments while extracting the learning objectives and relevant keywords.

  Output:

  Number of Objectives: 3

  Objective 1: Understand the basic structure of the Internet and its components.

  Segment 1:

  The Internet is a computer network that interconnects billions of computing devices throughout the world. Not too long ago, these computing devices were primarily traditional desktop PCs, Linux workstations, and servers that store and transmit information such as Web pages and e-mail messages. Increasingly, however, nontraditional Internet "things" such as laptops, smartphones, tablets, TVs, gaming consoles, thermostats, home security systems, home appliances, watches, eye glasses, cars, traffic control systems, and more are being connected to the Internet.

  Keywords 1: ['Internet', 'computing devices', 'desktop PCs', 'Linux workstations', 'smartphones', 'laptops', 'tablets', 'Internet of Things']

  Objective 2: Learn about the infrastructure and protocols that enable the Internet.

  Segment 2:

  These devices are interconnected using various communication protocols, which allow for the efficient transfer of data across vast distances. The Internet infrastructure relies on numerous physical media like fiber optics, coaxial cables, and wireless transmission methods to maintain its speed and reliability. Additionally, the Internet is built on the foundation of several key protocols, including TCP/IP, which ensures that data packets are correctly routed from one device to another.

  Keywords 2: ['communication protocols', 'data transfer', 'TCP/IP', 'fiber optics', 'coaxial cables', 'wireless transmission']

  Objective 3: Understand the role of Internet applications and services.

  Segment 3:

Furthermore, the Internet allows for the creation and deployment of applications that are accessible by users worldwide. These applications range from simple web browsing to complex cloud computing services. Understanding the components and working of the Internet is crucial for understanding how data is exchanged across the globe.

Keywords 3: ['applications', 'cloud computing', 'web browsing', 'global data exchange']

<END>

- **Step 3: Question Generation Agent** determines the question type (Conceptual, Numerical, or Application-based) and generates draft questions based on the text segments, objectives, and keywords.

  Input:

  Objective 1, Segment 1, Keywords 1.

  Output:

  Question Type: Conceptual Questions Generated Question: What are the primary types of devices traditionally connected to the Internet, and how have modern devices, such as smartphones and smart appliances, expanded the scope of the Internet?

  <END>

  Input:

  Objective 2, Segment 2, Keywords 2.

  Output:

  Question Type: Conceptual Questions Generated Question: How does TCP/IP facilitate data transfer across the Internet, and what role do physical media like fiber optics and coaxial cables play in ensuring reliable communication?

  <END>

  Input:

  Objective 3, Segment 3, Keywords 3.

  Output:

Question Type: Application Questions Generated Question: How would the global data exchange be affected if the Internet lacked applications like cloud computing services, and how would users be impacted?

<END>

- **Step 4: Question Evaluation Agent** evaluates the generated draft questions, provides feedback, and suggests improvements.

  Input:

  Question 1, Segment 1, Objective 1, Keywords 1.

  Output:

  Feedback: The question is clear and relevant, asking about the traditional and modern devices connected to the Internet.

  Suggestion: None required; the question is well-structured.

  <END>

  Input:

  Question 2, Segment 2, Objective 2, Keywords 2.

  Output:

  Feedback: The question is clear and asks about the role of TCP/IP and physical media, but could be more specific about how each type of physical medium contributes to data transfer.

  Suggestion: Include a specific example of how fiber optics vs. coaxial cables affect data speeds.

  <END>

  Input:

  Question 3, Segment 3, Objective 3, Keywords 3.

  Output:

  Feedback: The question is relevant, but the term ”global data exchange” could be clarified to ensure students understand its context in Internet services.

  Suggestion: Reword the question to specify ”how the Internet’s ability to exchange data globally relies on web-based applications like cloud computing.”

  <END>

- **Step 5: Question Refinement Agent** refines and optimizes the draft questions based on the feedback to improve their quality.

  Input:

  Question 1, Suggestion 1, Segment 1, Objective 1, Keywords 1.

  Output:

  Revised Question: What are the primary types of devices traditionally connected to the Internet, and how has the introduction of modern devices like smartphones and smart appliances expanded the scope of the Internet, leading to the concept of the Internet of Things (IoT)?
  `<END>`

  Input:

  Question 2, Suggestion 2, Segment 2, Objective 2, Keywords 2.

  Output:

  Revised Question: How does TCP/IP facilitate data transfer across the Internet, and in what way do physical media like fiber optics and coaxial cables ensure the reliability and speed of data communication, particularly in large-scale networks?
  `<END>`

  Input:

  Question 3, Suggestion 3, Segment 3, Objective 3, Keywords 3.

  Output:

  Revised Question: How would global data exchange be impacted if the Internet lacked applications such as cloud computing, and what role do these applications play in enabling businesses and individuals to access data and services across the globe? `<END>`

- **Step 6: Answer Generation Agent** generates accurate answers based on the refined questions and the relevant text.

  Input:

  Question 1, Segment 1, Objective 1, Keywords 1.

  Output:

Answer: Traditionally, devices such as desktop PCs, Linux workstations, and servers were primarily connected to the Internet. However, modern devices, including smartphones, laptops, tablets, and smart appliances, have expanded the scope of the Internet by enabling a vast network of interconnected devices, collectively known as the Internet of Things (IoT).
`<END>`

Input:

Question 2, Segment 2, Objective 2, Keywords 2.

Output:

Answer: TCP/IP ensures data transfer by breaking data into packets and transmitting them across the Internet, where they are reassembled at their destination. Physical media such as fiber optics provide high-speed connections, while coaxial cables are used for less demanding applications. Fiber optics offer faster data transmission and more reliable connections over long distances compared to coaxial cables, making them the preferred choice for backbone networks.
`<END>`

Input:

Question 2, Segment 2, Objective 2, Keywords 2.

Output:

Answer: Without applications such as cloud computing, the Internet would lose its capacity to offer scalable, global access to data and services. Cloud computing allows businesses and individuals to store and retrieve data remotely, enabling seamless communication and data exchange across the world. These applications are essential for industries ranging from e-commerce to remote work solutions.
`<END>`

4.4.1 Summary

- **User Input:** A Caesar Cipher text.
- **Material Analysis And Segmentation Agent:** Segments the text into objectives, extracts keywords.

- **Question Generation Agent:** Generates conceptual and application-based questions.
- **Question Evaluation Agent:** Evaluates questions and provides feedback and suggestions.
- **Question Refinement Agent:** Refines questions based on feedback.
- **Answer Generation Agent:** Generates answers based on the refined questions and corresponding segments.

Chapter 5

Experimental Evaluation

5.1 Experimental Setup

The experiment is designed to evaluate the performance of different model configurations to highlight the importance of each component of the proposed system. Specifically, the Full Model, which includes all agents, is compared to two reduced models: one without the Evaluation and Refinement Agents and one without any agents. The goal is to demonstrate how the integration of agents improves the quality of question generation and how the Text Summarization Agent is employed to handle longer inputs.

5.1.1 Models Evaluated:

- **Full Model (LLaMA 3-8B):** This model includes all five agents (Material Analysis and Segmentation, Question Generation, Question Evaluation, Question Refinement, and Answer Generation) and uses the LLaMA 3-8B variant for generating questions and answers. The system is designed to process large chunks of text and produce high-quality, context-specific questions and answers.

- **No Evaluation and Refinement (LLaMA 3-8B):** This model uses the same LLaMA 3-8B variant, but it omits the Question Evaluation and Question Refinement Agents. This setup allows us to analyze the importance of feedback and optimization in the generation process, comparing it to the Full Model's output.

- **No Agent (LLaMA 3-8B):** This configuration removes all agents except for the core LLaMA 3-8B model, generating questions directly from the input text. This allows us to examine the performance of the system when no segmentation, evaluation, or refinement occurs.

- **Full Model (LLaMA 3-3B):** A smaller variant of the Full Model using LLaMA 3-3B. This model suffers from input length limitations, which is why the Text Summarization Agent is introduced to summarize long texts in batches.

5.1.2 Text Summarization Agent (for Full Model LLaMA 3-3B)

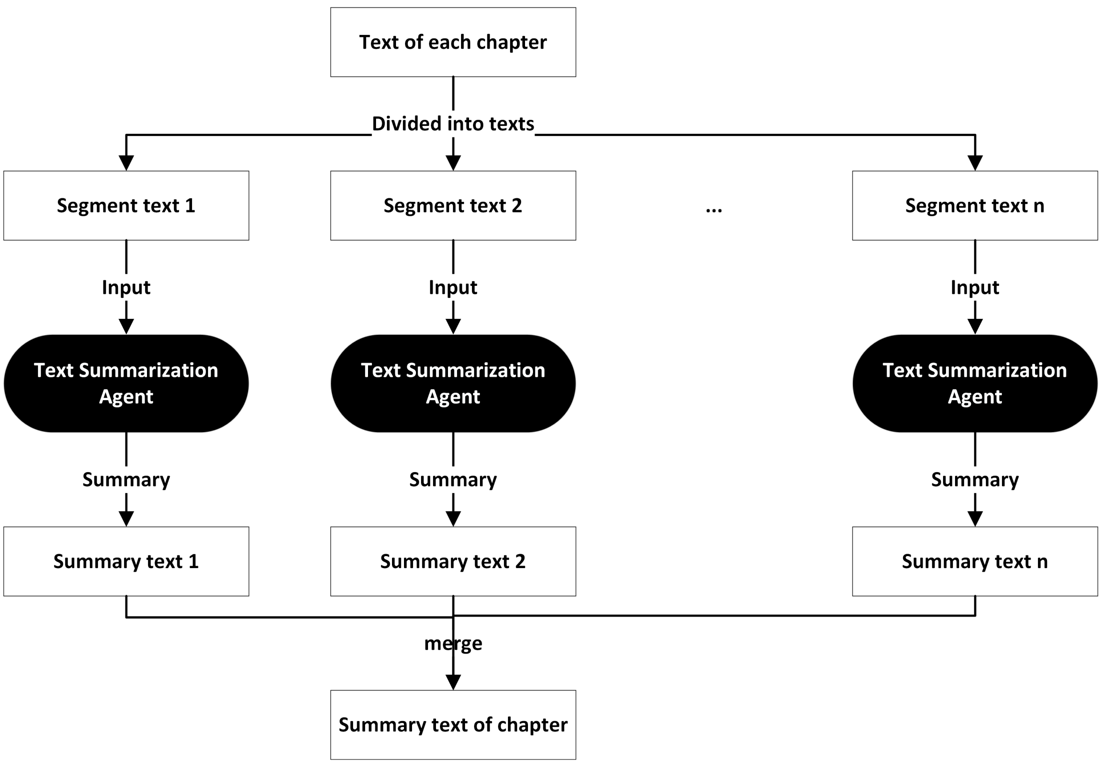

Figure 5.1: Text Summarization Agent Workflow

Due to input length restrictions in the LLaMA 3-3B model, the Text Summarization Agent is employed to handle large input texts. Instead of feeding the entire text at once, which may exceed the model's token limit, the Text Summarization Agent first splits the text into smaller chunks (sub-chapters of the textbook). Each chunk is summarized, and the resulting summaries are concatenated to form a final, more manageable input for subsequent agents.

5.1.3 Function of the Text Summarization Agent:

- **Manual Segmentation:** The textbook text is manually divided into smaller sections (sub-chapters), which ensures that each segment is focused and concise.

- **Text Summarization:** Each segment is then passed through the Text Summarization Agent, which generates a concise summary of the text. The agent does not analyze or segment the content; it only provides a summary of each text segment to reduce the overall token count.
- **Final Input:** After summarizing each segment, the summaries are concatenated to form the final input for the subsequent agents (Material Analysis and Segmentation, Question Generation, etc.). This ensures that the model can handle larger sections of text while respecting token limitations.

The process of the Text Summarization Agent is shown in Figure 5.1. The Text Summarization Agent thus plays a crucial role in enabling LLaMA 3-3B to process large textbooks efficiently while maintaining coherence and preserving the key information in the text.

5.1.4 Data Preparation

For the purpose of this experiment, we used the text of book Computer Networking [22], which was divided into 9 chapters. Each chapter was further split into smaller sections (sub-chapters). We employed four different models in this experiment: Full Model (LLaMA3-8B), Full Model (LLaMA3-3B), No Evaluation and Refinement Model(LLaMA3-8B), and No Agent Model(LLaMA3-8B). Each model processed the text of the book and generated 90 questions (10 questions per chapter, across 9 chapters), leading to a total of 360 questions. These questions were evaluated by human assessors to assess their quality and relevance.

Each model follows a slightly different procedure for generating the questions, as detailed below:

- **Full Model (LLaMA 3-8B)**
  - **Input Text:** The full text of the chapters.
  - **Processing:** Each text is directly input into the system. The Material Analysis and Segmentation Agent divides the text into manageable chunks and identifies learning objectives for each segment. The system proceeds with the Question Generation Agent, which generates questions based on these segmented texts and the learning objectives.
  - **Output:** The model generates 10 questions for each of the 9 chapters (a total of 90 questions). After the questions are generated, the Question Evaluation and Refinement Agents assess and refine the questions, improving their quality and alignment with the educational goals.
  - **Answer Generation:** Finally, the Answer Generation Agent generates answers to the refined questions based on the segment text and the question.

- **Full Model (LLaMA 3-3B)**
  - **Input Text:** Due to the token limitations of LLaMA 3-3B, the text is split into segment text. Each segment text is processed separately.
  - **Text Summarization:** Each segment of the text is summarized using the Text Summarization Agent. This agent takes each sub-chapter, processes it into a concise summary, and ensures the output fits within the token limits of the LLaMA 3-3B model. The summaries of all sub-chapters are then combined into a coherent summary, which is fed into the following agents.
  - **Processing:** The Material Analysis and Segmentation Agent then analyzes the summarized text, extracting learning objectives and keywords for each segment. The Question Generation Agent generates the questions based on the summarized and segmented content.
  - **Output:** The model generates 10 questions per chapter. After the questions are generated, the Question Evaluation and Refinement Agents assess and improve the question quality.
  - **Answer Generation:** Finally, the Answer Generation Agent generates answers to the refined questions based on the segment text and the question.
- **No Evaluation and Refinement Model(LLaMA 3-8B)**
  - **Input Text:** The full text is used as input, just like in the Full Model (LLaMA 3-8B).
  - **Processing:** The Material Analysis and Segmentation Agent divides the text into smaller segments and extracts learning objectives and keywords. The Question Generation Agent generates the questions based on these segments and objectives.
  - **Output:** 90 questions are generated in total. However, unlike the Full Model, this configuration skips the Question Evaluation and Question Refinement Agents, meaning the generated questions are not refined or evaluated for quality. This helps assess the importance of feedback and refinement in the question generation process.
  - **Answer Generation:** The Answer Generation Agent produces answers for the generated questions based on the text.

.

- **No Agent Model(LLaMA 3-8B)**
  - **Input Text:** The full text is used as input, just like in the Full Model (LLaMA 3-8B).

  - **Processing:** In this configuration, no agents are used for analysis, segmentation, evaluation, or refinement. The raw text is fed directly into the LLaMA 3-8B model, which generates questions directly from the input text without any pre-processing.
  - **Output:** 90 questions are generated. Since no agents are used for refining the questions or evaluating their quality, the model generates them without any feedback mechanism.
  - **Answer Generation:** The Answer Generation Agent produces answers for the generated questions, but without any evaluation of the questions, these answers may not always align as closely with the educational objectives.

- **Answer Generation:** The Answer Generation Agent generated answers for the refined questions based on the text.

5.1.5 Evaluation Criteria

Each question and its corresponding answer were evaluated on two primary criteria. The scoring criteria for questions are shown in Table 5.1, and the scoring criteria for answers are shown in Table 5.2

Score	Question Evaluation Criteria
1	Completely unsuitable for the topic, lacks clarity, or does not reflect the learning objectives.
2	Partially related to the topic but is vague, lacks structure, and may not effectively help students understand key concepts.
3	Relevant and clear, but lacks depth or challenge. It covers basic understanding but does not provoke deeper thinking.
4	Well-structured, encourages critical thinking, and helps students explore core concepts more deeply. It is somewhat challenging.
5	Highly insightful, completely aligned with the topic, and designed to guide students toward deep conceptual understanding. It challenges them to synthesize knowledge and apply reasoning effectively.

Table 5.1: Scoring Criteria for Questions

5.1.6 Human Evaluation

A group of 10 human evaluators participated in the assessment of the generated questions and answers:

- **Group 1:** 5 participants with domain expertise in computer networking.

Score	Answer Evaluation Criteria
1	Irrelevant or completely incorrect, failing to address the question meaningfully.
2	Partially relevant, but lacks completeness or clarity.
3	Clear and relevant, but has minor flaws in logic or lacks depth in explanation.
4	Comprehensive and well-explained, covering core concepts effectively.
5	Perfectly explains the question, covering all important details with logical clarity.

Table 5.2: Scoring Criteria for Answers

- **Group 2:** 5 participants with no domain-specific expertise, tasked with evaluating the questions and answers based on their ability to understand the material after reading the provided text.

Ten participants were given three days to read the content of the book computer network. After learning the book, participants rated four sets of question answer pairs (90 questions per group, for a total of 360 questions) based on the scoring criteria and text content over the course of a day.

5.1.7 Automated Evaluation

In this experiment, the Automated Evaluation step leverages ChatGPT-4o(paid) to assess the quality of the generated questions and answers. By training GPT-4 on the provided textbook text along with the established scoring criteria, we allow the model to automatically score 360 question-answer pairs (90 per model configuration). This automated evaluation is compared with human-assigned scores to assess its performance. First, provide GPT with the text of the book ”Computer Networking” and the scoring criteria for both questions and answers. The scoring criteria will define how the generated questions and answers should be evaluated.Then input 360 question-answer pairs into GPT, consisting of 4 sets of 90

5.2 Experimental Results (Human Evaluation)

The experimental results are analyzed across five key areas to evaluate the performance of the proposed models. Our primary goal is to highlight the impact of the modular framework with the evaluation and refinement agents, comparing it with configurations that either lack these agents or use a smaller version of the model. The smaller version needs less resource to run, so we want to see whether quality of the questions and answers are sufficiently good compared with the larger model. The experiments assess the effectiveness of these variations in generating high-quality questions and answers from the Computer Networking textbook. The following sections summarize the results:

- **Chapter-wise Comparison:** This analysis provides a detailed view of the performance across the nine chapters of the textbook. The results indicate how each model performs in generating questions and answers, shedding light on variations across the chapters.

- **Model Comparison:** The performance of the four models—Full Model (LLaMA 3-8B), Full Model (LLaMA 3-3B), No Evaluation and Refinement (LLaMA3-8B), and No Agent (LLaMA3-8B)—is compared to understand the contribution of each component in generating high-quality educational content.

- **Expert vs Non-Expert Evaluation:** The evaluation of the generated questions and answers by participants with and without expertise in the domain reveals insights into how well the models perform for both professional and non-professional users.

- **LLaMA3-8B vs LLaMA3-3B:** This comparison focuses on the effect of using different versions of the LLaMA model, with one being a larger version (3-8B) and the other a smaller one (3-3B). The results highlight the trade-offs between model size and output quality.

- **Comparison with Other Models:** Finally, the performance of the Full Model (LLaMA 3-8B) is compared with that of other state-of-the-art models, providing a benchmark for the proposed approach in terms of quality and relevance of the generated questions and answers.

Each of these comparisons will provide insight into the strengths and weaknesses of our modular approach and demonstrate how the integration of evaluation and refinement agents enhances the quality of educational question generation. The following sections present the results for each of these comparisons in detail.

5.2.1 Chapter-wise Score Comparison

Chapter	1	2	3	4	5	6	7	8	9
No Agent(Q)	2.13	1.87	2.11	1.91	1.91	2.09	2.02	1.94	2.03
No Evaluation & Refinement(Q)	3.09	2.98	2.92	2.99	3.03	2.97	3.07	2.81	3.01
Full Model(LLaMA3-8B)(Q)	4.06	4.00	4.10	3.93	4.06	3.86	4.03	3.90	4.08
Full Model(LLaMA3-3B)(Q)	3.83	3.75	3.77	3.78	3.78	3.74	3.73	3.92	3.63
Mean Question Score	3.28	3.15	3.22	3.15	3.19	3.17	3.21	3.14	3.19
No Agent(A)	2.07	2.06	1.99	1.89	1.83	2.05	1.96	1.92	1.87
No Evaluation & Refinement(A)	2.97	2.88	3.00	3.11	2.96	2.90	2.91	2.96	3.26
Full Model(LLaMA3-8B)(A)	4.00	3.96	3.88	3.98	3.96	4.06	3.93	3.94	3.95
Full Model(LLaMA3-3B)(A)	3.34	3.33	3.39	3.31	3.22	3.30	3.22	3.42	3.11
Mean Answer Score	3.10	3.06	3.07	3.07	2.99	3.08	3.01	3.06	3.05

Table 5.3: Average score for each chapter

In this section, we compare the average scores for each chapter generated by the four models. The results are shown for both question scores (Q) and answer scores (A), and are summarized in the Table 5.3. The Figure5.2 visualizing these results is also included for better understanding.

This figure illustrates the average question and answer scores for each chapter. As seen in the graph, the Full Model (LLaMA 3-8B) consistently produces the highest question and answer scores across all chapters, followed by Full Model (LLaMA 3-3B), No Evaluation and Refinement, and No Agent models.

From this, we can observe that:

5.2.1.1 Differences in Chapter Complexity and Content

One of the key reasons for the variation in scores across chapters is the inherent complexity and depth of the content in each chapter. Some chapters may contain more complex concepts, which naturally require more precise and intricate questions and answers.

It can be clearly seen from Figure 5.2 that chapter1 has the highest average score and Chapter 8 has the lowest average score.

Chapter 1, titled ”What Is the Internet?”, presents fundamental concepts related to the Internet and its infrastructure. The chapter provides clear and concise definitions of key components like end systems, communication links, and packet switches. It is written in a simple, easily understandable language, allowing both models and human evaluators to generate clear,

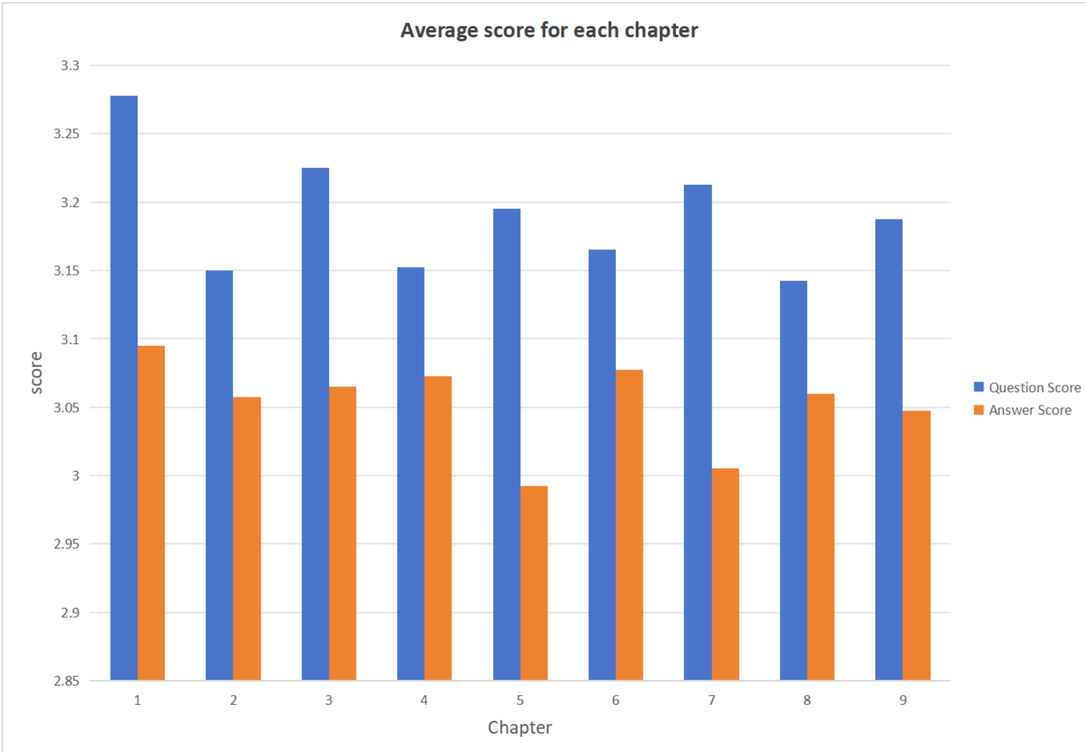

Figure 5.2: Average score for each chapter

relevant questions aligned with specific learning objectives. The content's simplicity aids in generating accurate questions and answers, making it easier for the model to handle. Additionally, the chapter is divided into well-structured segments with defined learning objectives, making segmentation and summarization more effective for question generation. The straightforward nature of the text allows the models to produce high-quality questions, especially for conceptual understanding.

In contrast, Chapter 8, "What Is Network Security?", delves into more complex concepts related to encryption, confidentiality, integrity, and cryptographic protocols. The chapter's technical depth, filled with jargon-heavy terms like ciphertext, encryption, and decryption, presents significant challenges for both the segmentation and question-generation processes. The specialized nature of the content makes it difficult for the models to identify the most relevant learning objectives and generate clear, focused questions. The segmentation of dense material such as cryptography requires careful handling to ensure accuracy without omitting crucial details, which the model struggled with in this chapter. This complexity likely contributed to lower scores, as the generated questions were less coherent and required more refinement. The difficulty in handling technical subjects like network security highlights the need for further improvements in summarization and segmentation, especially for dense, jargon-heavy content.

The higher score for Chapter 1 reflects its clear and accessible language, which makes it easier for both the model and human assessors to evaluate the generated questions and answers. Its straightforward content also helps the model produce coherent and relevant questions aligned with specific learning objectives. In contrast, Chapter 8's lower score is likely due to the higher complexity and specialized technical nature of the content. The models likely struggled more with generating high-quality questions for such dense material, and the segmentation process may have contributed to less effective question generation. These observations underscore the importance of text complexity and structure in determining the effectiveness of automatic question and answer generation systems. Further refinements in handling more complex content and improving segmentation will be crucial for enhancing performance in topics like network security.

5.2.1.2 Text Segmentation Impact

Another contributing factor to the variation in scores is the way text is segmented for input into the model. The Material Analysis and Segmentation Agent divides the content into smaller segments based on specific learning objectives. In some cases, the segmentation might not perfectly align with the most logical breaks in the material, leading to less coherent or harder-to-understand segments for the model. Consequently, this could impact the quality of the generated questions and answers.

For example, if a segment contains multiple related concepts that should have been addressed in one question but are divided into smaller segments, it might result in lower-quality questions and answers. Conversely, a well-structured segment with a clear and focused objective tends to generate higher-quality content, resulting in higher scores. The segmentation process's ability to accurately capture the main concepts and their relationships plays a significant role in determining how well the model can generate meaningful content.

5.2.1.3 Variability in Textual Structure

The text in different chapters of the textbook may have varying structures, which can influence the model's performance in generating questions and answers. Some chapters may be more narrative, explaining concepts through detailed examples or descriptions, while others might include more technical, formula-based content, which is more difficult for a model to handle. The structure and the level of abstraction in the content can affect how well the model understands the material and generates relevant questions and answers.

Chapters with highly technical content (e.g., those discussing specific algorithms or network configurations) may result in questions that are more specific but harder to evaluate. In contrast, chapters with broader conceptual content (e.g., basic principles of networking) are more likely to generate general, easily understandable questions, leading to higher question and answer scores.

5.2.1.4 Influence of Learning Objectives

The learning objectives derived from the text segments also contribute to the score variations. In some cases, the objectives might be more straightforward, such as understanding a particular concept, leading to clearer and more relevant questions. In other instances, the objectives might be more complex, involving the application of multiple concepts or the synthesis of ideas from different sections of the chapter. Such objectives lead to more complex questions and, as a result, could lead to lower scores for both questions and answers.

For example, Objective 1 might focus on basic concepts such as ”Understand the basic structure and components of the Internet.” while Objective 2 could involve more complex problem-solving, such as ”Learn about the Internet infrastructure and its services.” The difference in the complexity of the objectives could explain why some chapters exhibit lower scores.

In conclusion, the variations in chapter-wise scores can be attributed to a combination of factors, including the complexity and depth of the chapter's content, the effectiveness of the segmentation process, the structure of the material, and the nature of the learning objectives. While some chapters may generate simpler, more straightforward questions leading to higher scores, others may require more complex problem-solving, resulting in lower scores. The interaction between these factors is key to understanding why certain chapters receive higher or lower ratings in terms of question and answer quality.

5.2.2 Different Models and Their Impact on Overall Performance Comparison

Model	Question Score	Answer Score
Model With No Agent	2.00	1.96
No Evaluation & Refinement agent	2.99	2.99
Full Model (LLaMA3-8B)	4.00	3.96
Full Model (LLaMA3-3B)	3.77	3.29
Mean	3.19	3.05

Table 5.4: Average score for each Model

In the experimental setup, four distinct models were used to generate and assess questions and answers from the provided educational content. The models were compared based on their

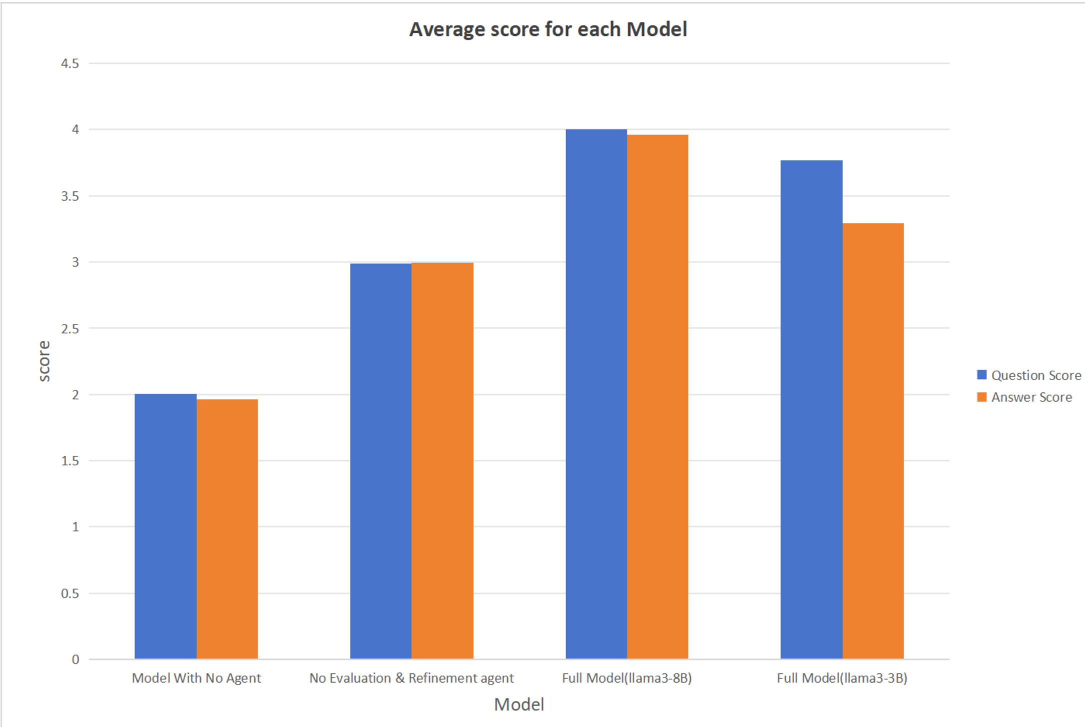

Figure 5.3: Average score for each Model

performance in terms of Question Scores and Answer Scores. The models included: Model With No Agent, Model with no evaluation and refinement agent, Full Model (LLaMA 3-8B) and Full Model(LLaMA 3-3B).

The average scores of the questions generated by the four models are shown in Table 5.4 visualizing these results is also included for better understanding.. And Figure 5.3 these results is also included for better understanding.The results show that the inclusion of different agents has a significant impact on the model's performance, as reflected in the varying question and answer scores across the models.

5.2.2.1 Model With No Agent:

The ”Model With No Agent” shows the lowest scores for both questions and answers, with a question score of 2.00 and an answer score of 1.96. This model lacks any kind of evaluation, refinement, and segmentation of the text into learning objectives. As a result, the questions generated are not as relevant or clear as they should be. Without the ability to properly segment the text and identify learning objectives, the model struggles to generate high-quality questions.

The absence of essential agents, such as the Material Analysis and Segmentation Agent and the Evaluation and Refinement Agents, leads to poorly structured questions that do not effectively test understanding or promote deeper learning. This highlights the critical role these agents play in ensuring the content's alignment with learning goals and its overall quality.

5.2.2.2 No Evaluation and Refinement Agent:

The ”No Evaluation and Refinement Agent” model performs significantly better than model with no agent, with a question score of 2.99 and an answer score of 2.99. This improvement can be attributed to the presence of the Material Analysis and Segmentation Agent, which divides the text into relevant segments and identifies key learning objectives. While this model does not benefit from the full evaluation and refinement processes, the segmentation and objective identification allow it to generate questions that are more relevant to the learning material. However, the lack of refinement and evaluation means that the generated questions are still not as precise or well-structured as those from more advanced models. Although better than the ”No Agent” model, it does not yet reach its full potential for generating high-quality educational content.

5.2.2.3 Full Model (LLaMA3-8B):

The ”Full Model (LLaMA3-8B)” achieved the highest scores, with a question score of 4.00 and an answer score of 3.96. This model includes all agents, such as Material Analysis and Segmentation, Question Generation, Evaluation, Refinement, and Answer Generation. By utilizing a more comprehensive setup, it can generate well-structured, relevant, and contextually accurate questions. The model's large size (8B parameters) enables it to handle complex educational content effectively, generate nuanced questions aligned with the learning objectives, and produce detailed, clear, and accurate answers. The incorporation of the Evaluation and Refinement Agents further optimizes the questions, ensuring they meet the desired quality standards. This model demonstrates the benefits of using a fully integrated system that employs multiple agents to improve both the quality of the questions and the accuracy of the answers.

5.2.3 Full Model (LLaMA3-3B) and Full Model (LLaMA 3-8B) Comparison

Model	Question Score	Answer Score
Full Model (LLaMA3-8B)	4.00	3.96
Full Model (LLaMA3-3B)	3.77	3.29

Table 5.5: Average score for Full Model (LLaMA3-8B) and Full Model (LLaMA3-3B)

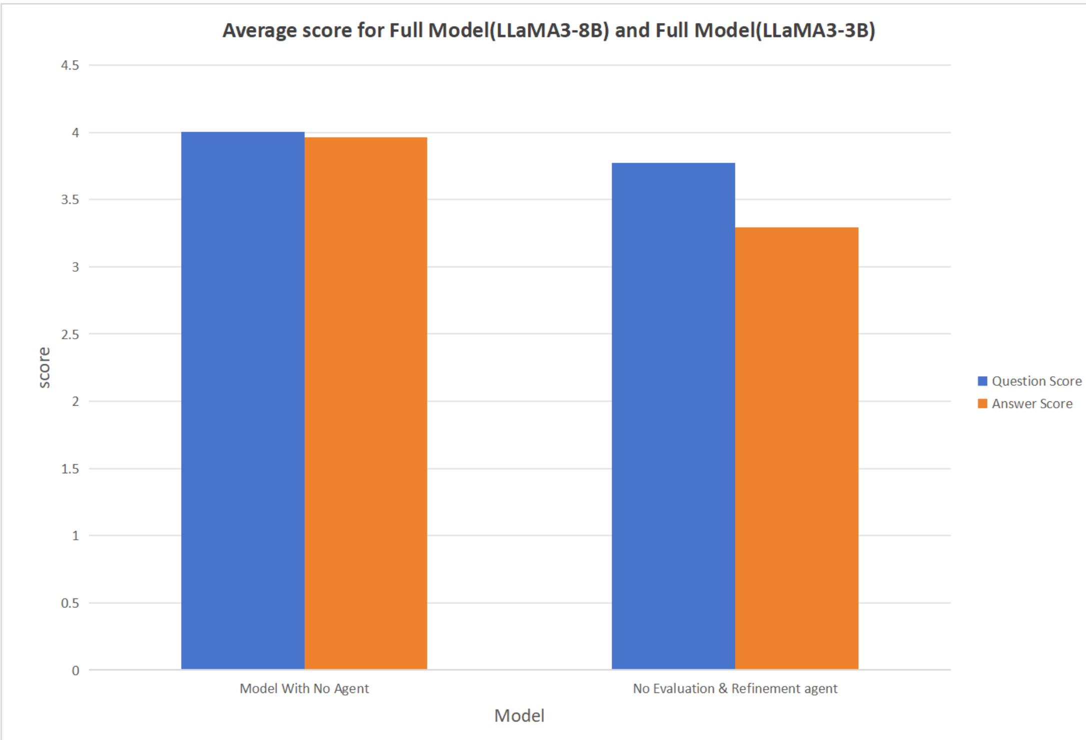

Figure 5.4: Average score for Full Model(LLaMA3-8B) and Full Model(LLaMA3-3B)

The average scores of the questions generated by Full Model(LLaMA3-8B) and Full Model(LLaMA3-3B) are shown in Table 5.5. And The Figure 5.4 visualizing these results is also included for better understanding.

In this section, we analyze the differences in performance between the Full Model (LLaMA 3-3B) and the Full Model (LLaMA 3-8B) by focusing on their question and answer scores. While both models share the same setup, the key difference is the number of parameters, with the 3-8B model having significantly more parameters than the 3-3B model.

When comparing the Full Model (LLaMA 3-3B) and Full Model (LLaMA 3-8B), several factors come into play that help explain their differing performance. While both models utilize similar architectures, the size of the models—specifically the number of parameters—plays a critical role in their output quality. LLaMA 3-8B, being a larger model, generally generates more nuanced and detailed responses, as it is capable of processing more complex contexts, but this comes with its own trade-offs.

5.2.3.1 Model Size and Capacity:

The most significant factor influencing the performance difference between the two models is their size. The LLaMA 3-8B model, with 8 billion parameters, is far better equipped to handle the complexity of the educational content. Larger models are able to capture more intricate relationships in the text, process more context, and generate responses with greater accuracy and depth. The additional parameters enable the model to make more nuanced inferences, producing more relevant and insightful questions and answers.

5.2.3.2 Contextual Understanding:

The larger LLaMA 3-8B model demonstrates a better understanding of the relationships between various pieces of information in the text, including the identification of learning objectives and the generation of questions that align with these objectives. It also handles more complex sentence structures and abstract concepts more effectively than the LLaMA 3-3B model. This improved contextual understanding is crucial in generating high-quality educational content that is not only relevant but also appropriately challenging for learners.

5.2.3.3 Quality of Questions and Answers:

The difference in model size impacts both the quality of the questions and the depth of the answers. The Full Model (LLaMA 3-8B) generates more challenging questions that require a deeper level of comprehension, which is essential for educational settings where critical thinking and problem-solving are emphasized. Additionally, the answers generated by LLaMA 3-8B are more comprehensive, logically sound, and clearly aligned with the learning objectives, ensuring that students receive accurate and detailed explanations.

5.2.3.4 Handling Complexity:

Educational content often involves complex, multifaceted information. The Full Model (LLaMA 3-8B), due to its larger size, can manage such complexity more effectively. The smaller LLaMA 3-3B model may struggle to generate high-quality questions and answers for these complex topics, resulting in less accurate and sometimes overly simplified responses.

5.2.3.5 Token Limitations:

One of the primary factors contributing to the difference in performance between LLaMA 3-3B and LLaMA 3-8B is the token limit of each model. The LLaMA 3-8B model, being larger, has a higher token limit and can handle longer chunks of text in one go. However, the text used in this experiment, derived from the ”Computer Networking” textbook, is extensive, and due to the inherent limitations of input length for both models, the process involves dividing the chapters into smaller sections. For LLaMA 3-8B, these sections must be processed individually, and a Text Summarization Agent is employed to consolidate the text from each section into a summarized form before feeding it into subsequent agents. This extra step introduces additional complexity and time overhead, as the model needs to process and summarize each segment, then combine the results.

In contrast, LLaMA3-3B has a smaller token limit, which might allow it to process smaller sections of text without requiring as much summarization. Although it does not handle as large chunks of text as effectively as LLaMA 3-8B, it can operate more quickly because it avoids the summarization step. As a result, LLaMA3-3B processes each chapter faster, though at the cost of reduced depth and complexity in the generated content. This token limit limitation is crucial for determining how efficiently each model handles large datasets.

5.2.3.6 Processing Speed:

Another significant difference between these models is their processing speed. Due to the smaller number of parameters, LLaMA3-3B is faster in generating responses compared to LLaMA3-8B. The smaller model requires less computation per token, allowing it to process text more quickly. This speed is particularly important when working in scenarios where real-time responses are needed, such as in interactive learning systems or applications requiring instant feedback.

While the larger LLaMA3-8B model offers greater accuracy in terms of question and answer quality, its processing speed is slower due to the larger computational resources required. This speed difference is especially evident in systems where multiple queries need to be answered rapidly, as LLaMA 3-3B will likely provide results faster, even though the questions and answers may not be as precise or detailed.

In conclusion, the Full Model (LLaMA3-8B) provides superior performance in terms of content quality but at the cost of processing speed, while LLaMA3-3B is faster but produces less complex and less detailed responses. The choice between these two models depends on the specific needs of the application. If high-quality content and deep understanding are the main requirements, LLaMA3-8B is the better option. However, if speed is a priority and the content does not need to be as detailed, LLaMA3-3B is more suitable. Additionally, the token limit and processing speed considerations will impact the model's effectiveness depending on the size of the text inputs and the computational resources available.

5.2.4 Professional and Non-professional Participants Comparison

	Question Score	Answer Score
Professional	3.18	3.2
Non-professional	2.91	3.19

Table 5.6: Average score of participants with and without expertise

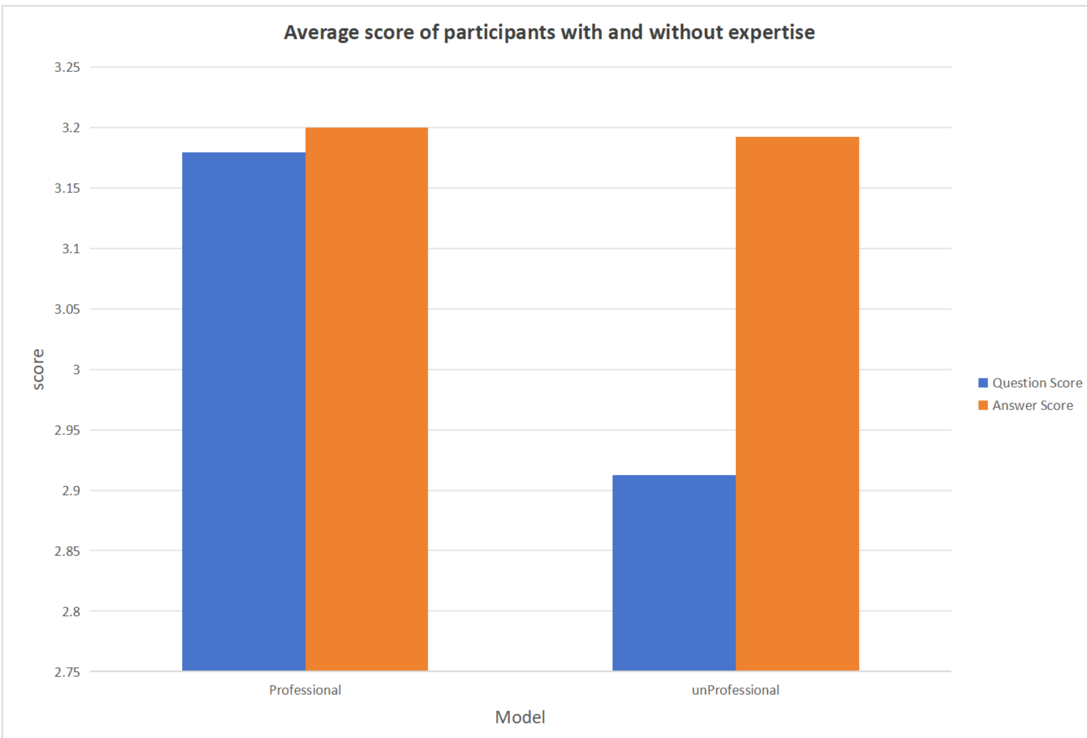

Figure 5.5: Average score of participants with and without expertise

Five of the 10 participants had specialized knowledge and the rest did not. All the participants studied by reading book ”Computer Networking” before they were graded. The average scores of the questions generated by Full Model (LLaMA3-8B) and Full Model (LLaMA3-3B) are shown in Table 5.6. And The Figure 5.5 visualizing these results is also included for better understanding.

The comparison between professional and non-professional participants reveals significant differences in how the generated questions and answers were evaluated. The professional group tends to rate both questions and answers more positively compared to the non-professional group, which is reflected in the higher Question Score and Answer Score for the professionals.

Professional participants, who have expertise in the subject matter, generally provided higher ratings for both the quality of the questions and the answers. This indicates that they likely have a better understanding of the content being assessed and can more accurately gauge the relevance and depth of the generated questions and answers.

The non-professional participants rated the questions lower than the professionals, especially in terms of the quality and relevance of the questions. However, their evaluation of the answers was closer to that of the professional group, which could indicate that even non-experts were able to understand the answers provided by the model, though they may not have been as equipped to assess the questions themselves.

Possible Reasons for the Difference:

- **Understanding of Subject Matter:** The primary reason for the score difference between the professional and non-professional groups likely stems from the participants' familiarity with the content. Professional participants are more adept at identifying high-quality, well-structured questions that align with the learning objectives of the subject matter, whereas non-professionals may find it more challenging to evaluate the relevance and depth of the generated questions, especially when they are complex or domain-specific.

- **Question Complexity:** The questions generated for this experiment were likely more complex for participants without professional knowledge to evaluate. For example, non-professionals may struggle to assess whether a question effectively captures core concepts or if it sufficiently challenges learners. Professionals, on the other hand, can better judge whether a question encourages deeper understanding or critical thinking.The questions generated for this experiment were likely more complex for participants without professional knowledge to evaluate. For example, non-professionals may struggle to assess whether a question effectively captures core concepts or if it sufficiently challenges learners. Professionals, on the other hand, can better judge whether a question encourages deeper understanding or critical thinking.

- **Clarity of Answers:** While the answers provided by the model were rated similarly by both groups, professionals might have been better able to identify subtle details or flaws in the answers. However, since the answers are directly related to the text, even non-professional participants could likely discern whether the answers were correct or meaningful, leading to a higher score for the answers from both groups.

- **Expertise in Evaluation Criteria:** Professionals likely had a more refined approach to assessing questions and answers based on educational standards or learning objectives, while non-professionals might have based their evaluations more on personal understanding or general knowledge. This discrepancy in evaluation criteria can explain the higher scores given by professionals for both questions and answers.

The data underscores the impact that subject matter expertise has on the evaluation of automatically generated educational content. While both professional and non-professional participants agreed on the quality of answers, professionals were able to better evaluate the questions, leading to higher scores in that category. This highlights the importance of tailoring question generation models to both experts and non-experts to ensure that questions meet the needs of a broad audience, especially in educational settings where learners at various expertise levels interact with the content.

5.3 Experimental Results(Automated Evaluation)

Model	Question Score	Answer Score
Model With No Agent	2.59	2.50
No Evaluation & Refinement agent	3.08	2.79
Full Model (LLaMA3-8B)	4.13	3.79

Table 5.7: Average score for each Model(Chat-GPT4o)

In this experiment, We use the book Computer Networking as input text. LLaMA(3-8B)-based models(Model With No Agent, model with no evaluation and refinement agent and full model) were used to generate 90 question-answer pairs (10 questions per chapter, across 9 chapters), totaling 270 question-answer pairs.

We use the paid version of ChatGPT4o, which can handle up to 8K tokens in its context window to score. After entering the text of the entire book and the grading criteria (as shown in Table5.1 and Table5.2), each of the three models generated 90 questions (out of a total of 270 questions). Finally, the average score of each model is obtained, as shown in Figure5.7.And The figure5.6 visualizing these results is also included for better understanding.

Full model show strong performance, likely due to the integration of multiple agents for text segmentation, summarization, question generation, evaluation, and refinement. These agents work together to ensure that questions and answers are relevant, accurate, and clear.

In comparison to other systems that have been discussed in the literature, this study demonstrates the effectiveness of using multiple agents within a LLaMA-based framework for educational content generation. While prior models [7] (like GPT-based models) often use only simple question generation techniques without evaluation or refinement, our system benefits from a more structured approach, including text segmentation, summarization, and refinement.

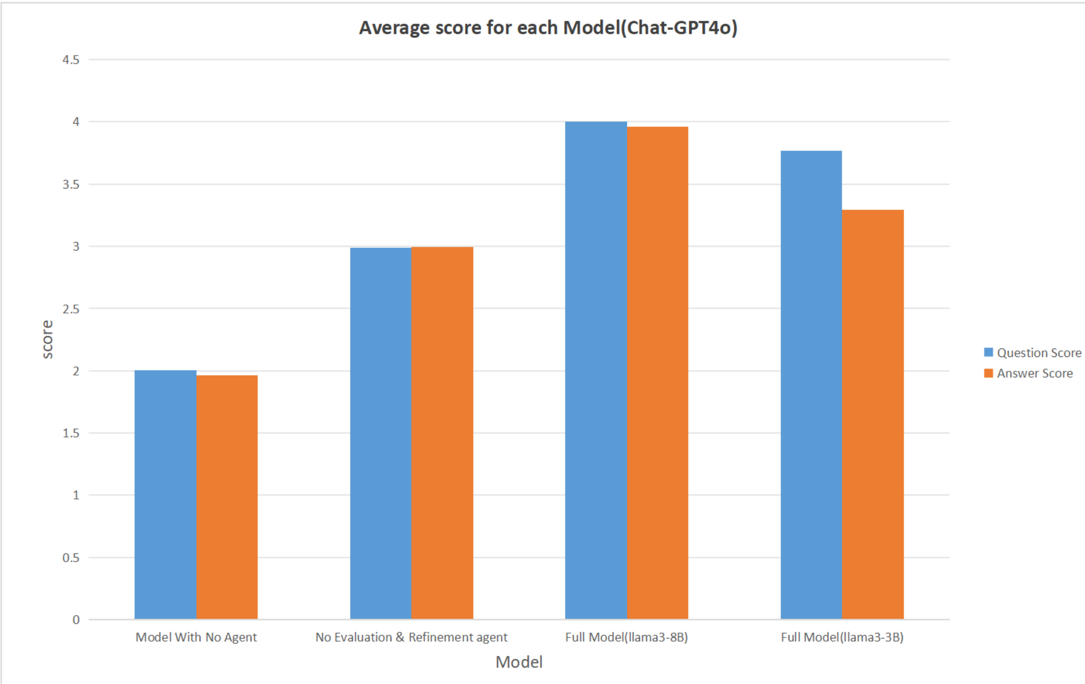

Figure 5.6: Average score for each Model(Chat-GPT4o)

From the results, it is evident that the Full Model performs significantly better than the No Agent and No Evaluation and Refinement models. The integration of Text Summarization, Evaluation, and Refinement agents significantly improves both the quality of the questions and the accuracy of the answers.

5.4 Comparison with Existing AQG Models

My model differs from similar models like T5, GPT-3, BERT, Seq2Seq models, and others primarily in its modular structure, which integrates distinct agents for material analysis, question generation, evaluation, refinement, and answer generation. This modular approach provides more precise control over each stage of the question generation process, allowing for better alignment with educational objectives. For instance, unlike T5 or GPT-3, which generate questions in a more general context, my model segments the input text based on learning objectives, ensuring that each generated question is more relevant and focused on specific educational goals. Additionally, my model incorporates an evaluation and refinement step that improves the clarity, difficulty, and relevance of the generated questions and answers. This iterative process ensures higher quality and educational value, whereas traditional models like

Seq2Seq may not have a similar built-in mechanism to assess and refine their output. While GPT-3-based MCQ models excel in generating plausible distractors, my model's approach goes further by refining not only the questions but also the answers, ensuring that they align with the content and provide educational value. Although my model does not explicitly use techniques like RAG or CBR, its segmentation and iterative feedback process function similarly by improving the relevance and factual accuracy of the generated questions. This structure enhances the overall effectiveness of question generation in educational contexts, though the integration of external knowledge sources, like RAG, could potentially be beneficial for more knowledge-intensive domains.

Chapter 6

Conclusion

In this study, we developed and evaluated an automatic question generation (AQG) system designed to enhance educational content through the use of large language models, specifically LLaMA. By leveraging a multi-agent approach, including components for material analysis, question generation, evaluation, refinement, and answer generation, we demonstrated how AI can assist in generating high-quality, relevant, and accurate questions and answers from educational texts.

6.1 Key Findings

- **Effectiveness of Full Model:** The results clearly show that the Full Model (LLaMA 3-8B), which incorporates all four agents (Material Analysis and Segmentation Agent, Question Generation Agent, Question Evaluation and Refinement Agent, and Answer Generation Agent), outperforms the other configurations in both question quality and answer relevance. It achieved the highest average scores for both question and answer evaluations, emphasizing the importance of having a comprehensive system that includes not only generation but also evaluation and refinement steps.

- **Impact of Evaluation and Refinement Agents:** The inclusion of the Question Evaluation and Refinement Agents led to significant improvements in both question and answer scores. This highlights the crucial role of iterative refinement processes in ensuring the generation of high-quality educational content. The No Evaluation and Refinement Agent Model performed considerably worse, demonstrating the limitations of relying solely on a generation-focused approach without human-like oversight or feedback mechanisms.

- **Effect of Learning Objectives:** The analysis showed that the complexity of learning objectives significantly influenced the quality of generated questions. Simpler objectives, such as defining basic concepts, tended to result in clearer and more relevant questions, leading to higher scores. In contrast, more complex objectives, such as those requiring the application of multiple concepts or the synthesis of ideas from different sections, generated questions that were harder to evaluate and refine, resulting in lower scores for both questions and answers.

- **Professional vs. Non-professional** The comparison of evaluations by professional and non-professional assessors revealed that professional assessors consistently provided higher ratings for both question clarity and answer relevance. This suggests that expertise in the subject matter plays a key role in assessing the quality of educational content, and further highlights the need for fine-tuning models to align with the expectations and needs of the target audience.

6.2 Contributions of this Thesis

This research makes several key contributions to the field of automatic question generation:

- **Comprehensive Multi-Agent System:** We presented a novel multi-agent system for AQG, which not only generates questions but also evaluates, refines, and generates corresponding answers. This end-to-end approach is essential for generating high-quality educational content.

- **Integration with LLaMA Models:** By integrating the LLaMA models, particularly the 3-8B variant, into the AQG process, this study demonstrates the potential of using state-of-the-art large language models for educational content generation. The results confirm that LLaMA can effectively handle complex text inputs and generate meaningful, contextually appropriate questions and answers.

- **Evaluation Framework:** The study introduced a structured evaluation framework, with both human assessors and the system itself playing crucial roles in determining the effectiveness of the generated questions. This ensures that the generated content is both relevant and of high quality.

Chapter 7

Future Work

While this study demonstrates the potential of using multi-agent AQG systems with LLaMA models, several areas remain ripe for further exploration and enhancement. In future work, we aim to address existing challenges, improve the system's capabilities, and explore new avenues for AI-driven educational content generation. Below are some key directions for future research:

- **Improved Text Segmentation and Contextualization**

  Although the study shows promising results with segmented text input, further advancements in automatic text segmentation are needed to better handle large and complex educational materials. More sophisticated segmentation algorithms could help divide the content more effectively, ensuring each segment maintains coherence and aligns with specific learning objectives. Additionally, the system could be improved to contextualize the segments better across chapters and subchapters, which would enhance the relevance and continuity of the questions generated.

- **Fine-Tuning for Domain-Specific Content**

  While the current system relies on general models like LLaMA, domain-specific fine-tuning would increase the relevance and precision of generated questions and answers. Fine-tuning the model on large, domain-specific educational datasets can ensure that the model produces more accurate and contextually appropriate content, especially for technical or niche subjects. For instance, datasets specific to computer networking, cybersecurity, or machine learning could be used to refine the system for better subject-specific performance.

- **Adaptive Learning Capabilities**

  One of the most promising directions for future research is the integration of adaptive learning techniques. By incorporating learner data—such as past responses, learning progress, and feedback—the system could dynamically adjust the complexity of generated questions and answers based on the learner's current knowledge level. This adaptive approach could create a more personalized learning experience, helping students tackle questions that align with their abilities and prior knowledge. The system could become more interactive by incorporating real-time adjustments based on student performance and feedback.

- **Incorporation of Real-Time Feedback**

  Incorporating real-time feedback loops could significantly enhance the model's ability to improve over time. This could involve integrating user feedback (such as student or instructor reviews) into the question generation process, allowing the system to self-correct and evolve. For example, students could provide feedback on the clarity or difficulty of questions, and the model could adjust subsequent questions to ensure better engagement and learning outcomes. Real-time adaptation to user needs could be crucial for more effective teaching tools.

- **Cross-Model and Cross-Domain Comparison**

  The current comparison of LLaMA-based models could be expanded to include a broader range of models, such as GPT-4, T5, or other state-of-the-art architectures. Cross-domain comparisons across different subjects (e.g., literature, mathematics, history) would allow for a more comprehensive understanding of the model's strengths and weaknesses. Evaluating the same AQG setup across diverse datasets would help identify potential areas of improvement and provide more robust benchmarks for future models.

- **Human-AI Collaboration in Education**

  Another key area for future work is the exploration of human-AI collaboration in educational settings. Rather than having the AI generate content in isolation, educators could be involved in the process of refining, adjusting, or selecting the generated content based on their pedagogical goals. AI could act as a co-creator with educators, enabling them to more efficiently generate and customize questions for specific educational objectives. Exploring how teachers and AI can interact in the question-generation process will be critical for ensuring the system aligns with real-world teaching needs.

- **Scalability and Deployment**

  Scaling the system to handle larger datasets and more diverse educational texts remains an essential next step. A more robust architecture that can efficiently process and generate questions from substantial volumes of educational content would be valuable. Additionally, ensuring that the system can be deployed in real-world educational settings, including integration with learning management systems (LMS) and other digital education platforms, is crucial. Research in this area should focus on ensuring the system is easily scalable and deployable in classrooms, online learning environments, and other educational platforms.

- **Ethical Considerations and Bias Mitigation**

  As AI models increasingly influence educational content, it is important to address ethical considerations such as bias mitigation in question generation. The system's outputs should be free from bias and should represent diverse perspectives, ensuring equitable learning opportunities for all students. Developing methodologies to detect and eliminate biases in the generated questions and answers is crucial to ensure fairness and inclusivity in AI-driven education systems. This is especially important in content related to sensitive topics or diverse student populations.

- **Evaluation Methodology Improvements**

  Lastly, improving the evaluation methodology for generated questions and answers is essential for advancing AQG systems. While human assessors provide valuable feedback, more automated evaluation techniques, such as incorporating rubric-based scoring or peer-review systems, could increase the efficiency of the evaluation process. Additionally, expanding the evaluation criteria to include metrics such as question difficulty, diversity, and cognitive load would allow for a more comprehensive assessment of the model's output.

- **Exploring Multi-Modal Educational Content Generation**

  Future research could also explore multi-modal learning environments, where the AQG system generates not only text-based questions but also questions based on visual, audio, and interactive content. For example, educational videos or infographics could be used as input, with the system generating questions that test comprehension of both the textual and visual components. This would enable more engaging and interactive learning experiences that cater to different learning styles and needs.

- **Generation of Multiple-Choice Questions (MCQs)**

  Multiple-choice questions (MCQs) are essential in large class sizes, providing an efficient way to assess student learning. However, generating appropriate distractors (incorrect answers) remains a significant challenge, as mentioned earlier in the thesis. The task is to create distractors that are both conceptually relevant and cognitively demanding. Future work will focus on addressing this challenge by improving the quality and relevance of generated distractors. Key directions include enhancing the model's ability to understand context more effectively, utilizing external knowledge bases to automatically generate plausible distractors, and developing new evaluation metrics specifically to assess distractor quality. By addressing these issues, we aim to make the system better suited for large-scale assessments and improve the quality and effectiveness of automatic question generation.

In summary, the future development of AQG systems presents numerous opportunities for improvement and innovation. By addressing the challenges of text segmentation, domain-specific fine-tuning, real-time adaptation, and integrating human feedback, these systems can evolve to create more effective and personalized educational content. The continued advancement of these technologies holds the potential to revolutionize education, offering scalable, adaptive, and accessible learning experiences for students worldwide.

References

[1] Diwan, C., Srinivasa, S., Suri, G., Agarwal, S. & Ram, P. AI-based learning content generation and learning pathway augmentation to increase learner engagement. *Computers And Education: Artificial Intelligence*. **4** pp. 100110 (2023)

[2] Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A. & Others A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. *Proceedings Of The 26th Australasian Computing Education Conference*. pp. 114-123 (2024)

[3] Lu, O., Huang, A., Tsai, D. & Yang, S. Expert-authored and machine-generated short-answer questions for assessing students learning performance. *Educational Technology & Society*. **24**, 159-173 (2021)

[4] Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. (2018)

[5] Qu, F., Jia, X. & Wu, Y. Asking questions like educational experts: Automatically generating question-answer pairs on real-world examination data. *ArXiv Preprint ArXiv:2109.05179*. (2021)

[6] Liu, B., Zhao, M., Niu, D., Lai, K., He, Y., Wei, H. & Xu, Y. Learning to generate questions by learningwhat not to generate. *The World Wide Web Conference*. pp. 1106-1118 (2019)

[7] Pan, L., Lei, W., Chua, T. & Kan, M. Recent advances in neural question generation. *ArXiv Preprint ArXiv:1905.08949*. (2019)

[8] Biancini, G., Ferrato, A. & Limongelli, C. Multiple-choice question generation using large language models: Methodology and educator insights. *Adjunct Proceedings Of The 32nd ACM Conference On User Modeling, Adaptation And Personalization*. pp. 584-590 (2024)

[9] Scaria, N., Dharani Chenna, S. & Subramani, D. Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation. *International Conference On Artificial Intelligence In Education*. pp. 165-179 (2024)

[10] Goyal, R., Kumar, P. & Singh, V. Automated question and answer generation from texts using text-to-text transformers. *Arabian Journal For Science And Engineering*. **49**, 3027-3041 (2024)

[11] Kıyak, Y., Coşkun, Ö., Budakoğlu, I. & Uluoğlu, C. ChatGPT for generating multiple-choice questions: evidence on the use of artificial intelligence in automatic item generation for a rational pharmacotherapy exam. *European Journal Of Clinical Pharmacology*. **80**, 729-735 (2024)

[12] Kumar, A., Nayak, A., K, M., Chaitanya & Ghosh, K. A novel framework for the generation of multiple choice question stems using semantic and machine-learning techniques. *International Journal Of Artificial Intelligence In Education*. **34**, 332-375 (2024)

[13] Wang, X., Aitchison, L. & Rudolph, M. LoRA ensembles for large language model fine-tuning. *ArXiv Preprint ArXiv:2310.00035*. (2023)

[14] Abdullin, Y., Molla-Aliod, D., Ofoghi, B., Yearwood, J. & Li, Q. Synthetic dialogue dataset generation using llm agents. *ArXiv Preprint ArXiv:2401.17461*. (2024)

[15] Wiratunga, N., Abeyratne, R., Jayawardena, L., Martin, K., Massie, S., Nkisi-Orji, I., Weerasinghe, R., Liret, A. & Fleisch, B. CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. *International Conference On Case-Based Reasoning*. pp. 445-460 (2024)

[16] Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J. & Kim, H. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education And Information Technologies*. **29**, 11483-11515 (2024)

[17] Yuan, X., Wang, T., Wang, Y., Fine, E., Abdelghani, R., Lucas, P., Sauzéon, H. & Oudeyer, P. Selecting better samples from pre-trained LLMs: A case study on question generation. *ArXiv Preprint ArXiv:2209.11000*. (2022)

[18] Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J. & Kim, H. Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education And Information Technologies*. **29**, 11483-11515 (2024)

[19] Duan, N., Tang, D., Chen, P. & Zhou, M. Question Generation for Question Answering. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pp. 866-874 (2017)

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems*. **30** (2017)

[21] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A. & Others The llama 3 herd of models. *ArXiv Preprint ArXiv:2407.21783*. (2024)

[22] Kurose, J. & Ross, K. Computer Networking: A Top-down Approach. (Pearson Education,2017), Global ed.