AN ARTIFICIAL INTELLIGENCE-BASED SYSTEM FOR AUTOMATIC REFLECTION QUESTION GENERATION IN EDUCATIONAL SETTINGS

Ondřej Holub

Bachelor's thesis
Faculty of Information Technology
Czech Technical University in Prague
Department of Applied Mathematics
Study program: Informatics
Specialisation: Artificial Intelligence 2021
Supervisor: Rodrigo Augusto da Silva Alves, Ph.D.
Supervisor Specialist: Essi Ryymin, Ph.D.
May 15, 2025

Assignment of bachelor's thesis

Title:	An Artificial Intelligence-Based System for Automatic Reflection Question Generation in Educational Settings
Student:	Ondřej Holub
Supervisor:	Rodrigo Augusto da Silva Alves, Ph.D.
Study program:	Informatics
Branch / specialization:	Artificial Intelligence 2021
Department:	Department of Applied Mathematics
Validity:	until the end of summer semester 2025/2026

Instructions

Education is continually evolving with the integration of advanced technologies that support teaching and learning processes. One promising avenue is the automatic generation of high-quality questions for reflection and assessment from lesson content. A reflection question in education is an open-ended question that prompts students to critically analyze and connect their personal experiences with academic concepts to deepen their understanding. Traditional methods of creating reflective and assessment questions are labor-intensive and may not always capture the nuanced understanding required to assess student comprehension effectively. Recent advances in Large Language Models (LLMs) offer an opportunity to automate this process by extracting semantic insights from educational material and formulating insightful reflection questions. Although LLMs are being used to generate questions [1,2], less attention has been given to exploring modern reasoning mechanisms (e.g., chain-of-thoughts) to generate reflections, considering their inherent complexities. Thus, generation of reflection question based on artificial intelligence is still an open problem. This project aims to fill this gap by developing and evaluating an Artificial Intelligence-based system that automatically generates high-quality questions from lessons while investigating the capabilities and limitations of LLMs in the educational domain by considering the recent advances of LLMs-based reasoning.

The following tasks are proposed:

Electronically approved by Ing. Magda Friedjungová, Ph.D. on 5 February 2025 in Prague.

1) Conduct a comprehensive review of recent literature on LLMs and their applications in education, with a focus on automated question generation.

2) Develop an AI-based framework that integrates LLM-generated embeddings for the automatic creation of high-quality reflection questions.

3) Evaluate the performance of the developed system by comparing it with the state of the art using both qualitative and (if possible) quantitative methods.

4) Discuss the interpretability and practical implications of the results, focusing on the strengths and weaknesses of LLMs in the context of educational question generation.

[1] Lu, Xinyi, and Xu Wang. "Generative students: Using llm-simulated student profiles to support question item evaluation." In Proceedings of the Eleventh ACM Conference on Learning@ Scale, pp. 16-27. 2024.

[2] Hasan, A.S.M., Ehsan, M.A., Shahnoor, K.B. and Tasneem, S.S., 2024. Automatic question & answer generation using generative Large Language Model (LLM) (Doctoral dissertation, Brac University).

Electronically approved by Ing. Magda Friedjungová, Ph.D. on 5 February 2025 in Prague.

Citation of this thesis: Holub Ondřej. *An Artificial Intelligence-Based System for Automatic Reflection Question Generation in Educational Settings.* Bachelor's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2025.

I would like to thank my supervisor, Rodrigo Augusto da Silva Alves, Ph.D., for his continuous guidance and support from the beginning to the end of this thesis. This thesis would not have been possible without his help. I would also like to thank my specialist supervisor, Essi Ryymin, Ph.D., for her valuable insights into the educational aspects of the thesis. Furthermore, I extend my gratitude to AI dětem for generously providing the necessary resources for creation of the system. Finally, I would like to thank my family and friends for their support and encouragement throughout this process. I am grateful for their understanding and patience during the long hours spent working on this thesis.

Declaration

I hereby declare that the presented thesis is my own work and that I have cited all sources of information in accordance with the Guideline for adhering to ethical principles when elaborating an academic final thesis.

I acknowledge that my thesis is subject to the rights and obligations stipulated by the Act No. 121/2000 Coll., the Copyright Act, as amended. In accordance with Section 2373(2) of Act No. 89/2012 Coll., the Civil Code, as amended, I hereby grant a non-exclusive authorization (licence) to utilize this thesis, including all computer programs that are part of it or attached to it and all documentation thereof (hereinafter collectively referred to as the "Work"), to any and all persons who wish to use the Work. Such persons are entitled to use the Work in any manner that does not diminish the value of the Work and for any purpose (including use for profit). This authorisation is unlimited in time, territory and quantity.

I declare that I have used AI tools during the preparation and writing of my thesis. I have verified the generated content. I confirm that I am aware that I am fully responsible for the content of the thesis.

In Prague on May 15, 2025

Abstract

This thesis presents a system for generating reflection questions in educational settings using large language models (LLMs). The system uses the Socratic method in a multi-round dialogue between two separate LLM instances each with its own unique parameters to improve the quality of generated questions. For the LLMs the 4o-mini ChatGPT model was used for ease of testing and evaluation. The final system shows promise in generating high-quality reflective questions – the usage of Socratic dialogue improving the quality of the results. Few areas for improvement exist, mainly in the evaluation of quality of final questions and determining the right time to end the dialogue. The prompts created during the development of the system are available in the attachments.

Keywords Large language model, automatic question generation, Socratic method, reflection questions, reflection in education

Abstrakt

Tato práce se zabývá systémem pro generování reflexivních otázek ve vzdělávacím prostředí pomocí velkých jazykových modelů (LLM). Systém využívá Sokratovskou metodu v dialogu mezi dvěma samostatnými instancemi LLM, z nichž každá má své vlastní jedinečné parametry, aby zlepšil kvalitu generovaných otázek. Pro LLM byl použit model 4o-mini ChatGPT pro snadné testování a hodnocení. Konečný systém vykazuje slibné výsledky při generování vysoce kvalitních reflexivních otázek – použití sokratického dialogu zlepšuje kvalitu výsledků. Existuje několik oblastí pro zlepšení, zejména v hodnocení kvality konečných otázek a určení vhodné doby pro ukončení dialogu. Výzvy vytvořené během vývoje systému jsou k dispozici v přílohách.

Klíčová slova Velký jazykový model, automatické generování otázek, Sokratovská metoda, reflexivní otázky, reflexe ve vzdělávání

Contents

List of Figures

List of Tables

List of abbreviations

AI	Artificial Intelligence
AQG	Automatic Question Generation
BLEU	BiLingual Evaluation Understudy Score
BERT	Bidirectional Encoder Representations from Transformers
LLM	Large Language Model
LSTM	Long Short-Term Memory
MCQ	Multiple-Choice Question
METEOR	Metric for Evaluation of Translation with Explicit ORdering
ML	Machine Learning
NLP	Natural Language Processing
QA	Question Answering
QG	Question Generation
RNN	Recurrent Neural Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
seq2seq	Sequence-to-Sequence
SQuAD	Stanford Question Answering Dataset

Chapter 1
Introduction

The problem of automatic question generation (AQG) has been explored widely in the literature with a variety of different methods used for the question generation (QG) process. Many different automatic systems for question generation have been developed. These systems are of a great benefit to various areas such as the educational process, easing the workload of teachers and providing an easy way for students to test their knowledge, or in the process of interviewing, where the system can create questions for the interviewer to ask the interviewee. However, the focus of these systems was mainly on generating multiple-choice, true/false, or fill-in-the-blank questions that test particular knowledge of the student. The questions created by these systems are not specifically designed to take into account the process of reflection, which is known to be an important part of the learning process.

The focus of this thesis is to fill that gap and propose an architecture for a system that generates high-quality reflection questions about a given topic. The created system should lower the difficulty for the creation of reflection and self-assessment questions for teachers and students and ease the process of reflection for students, thus enhancing their learning experience. Additional research will be conducted in cooperation with real teachers to evaluate the quality of questions generated by the system. However, the results of those tests will not be presented in this thesis due to time constraints.

Main goals of this thesis are: First, review the literature on the topic of reflection, educational models, and existing systems for automated question generation. Second, develop an automated question generation system that generates high-quality reflection questions. Third, evaluate the system using both qualitative and (if possible) quantitative evaluation. Fourth, discuss the interpretability and practical implications of the results.

The thesis fulfills these goals as follows:

- The development of an AI-based framework for the automatic creation of high-quality reflection questions is primarily addressed in Chapter 3, detailing the system's architecture and implementation.
- To evaluate the performance of the developed system, Chapter 4 presents the experiments conducted, comparing the system with the state of the art using both qualitative and, where feasible, quantitative methods.
- A discussion concerning the interpretability and practical implications of the results, with a focus on the strengths and weaknesses of LLMs in the context of educational question generation, is provided in Chapter 5, based on the experimental findings from Chapter 4.
- A comprehensive review of recent literature on LLMs and their applications in education, with a particular focus on automated question generation, is undertaken in Chapter 2.

Chapter 2

Literature Review

2.1 Reflection

Reflection is considered to be a very important part of the learning process and its usefulness has been recognized for a long time. The use of reflection can be found in many other fields besides learning as well, such as psychology, sociology, and philosophy [1, 2].

However, the term itself lacks clarity for many people and its conceptualization is inconsistent. This is due to the delineation of reflection as a *construct* – an abstract concept which is not easily observable and may be difficult to define [3]. It is a compound process which involves complex cognitive, emotional, and social processes and can be influenced by many factors. Hence, establishing a clear definition of reflection within the framework of this thesis is essential.

2.1.1 History of Reflection

Roots of reflection can be traced back to ancient Greece and the works of Socrates. Though the term ”reflection” was not yet invented at the time, Socrates was already using a method of questioning to help his students think critically about their beliefs and assumptions. Socrates believed that this process of questioning and dialogue was essential for learning and understanding. [4]

In Plato's Apology [5], Socrates is quoted as saying, ”The unexamined life is not worth living.” and that he would rather die than to be exiled and forced to stop questioning. This method, from then on known as the Socratic method, is still used today to encourage critical thinking and reflection in students. It is discussed in more detail in section 2.2.

Although the concept of reflection has been around for a long time, its use in education was properly introduced in the 20th century with the works of John Dewey. John Dewey had a significant impact on the field and his ideas served as a foundation for many theories about reflection. According to Dewey, reflection is a distinct type of thought separate from other types of thought, such as belief or invention [6]. He stated that reflection is a deliberate and systematic thought process about an uncertain problem or situation in order to understand it better and to make informed decisions. Reflection is not just a passive process of thinking about experiences, but an active process of engaging with them [6]. He believed reflection is a way of making sense of experiences and that it is essential for personal and professional growth. It is influenced by a wider context of affective, social, and cultural factors. Dewey emphasized the importance of the learner's experiences as the base blocks for reflection and learning [7]. A specific definition of reflection by Dewey is provided in Definition 2.1.

He also proposes that the reflective way of thinking should be equipped with three qualities of attitudes: *open-mindedness* to new ideas and thoughts, be willing to accept them; *whole-heartedness* to seek out new approaches and fully engage with them; and *responsibility* to realize the consequences of our own actions [8].

► **Definition 2.1.** *Active, persistent and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it and the further conclusion to which it tends, constitutes reflective thought* [7]*.*

A few decades later, Donald Schön built on Dewey's work and introduced the concept of reflection in professional practice. Schön's work focused on the role of reflection in the professional learning process and how it can be used to improve practice. He argued that reflection is an essential part of professional practice and that it is necessary for practitioners to engage in reflective thinking in order to improve their skills and knowledge [9]. Schön observed that even though students were trained and taught to be professionals in their fields, they often struggled to apply their knowledge in real-world employment when faced with uncertain and unique situations not covered in their training. Thus, they had to rely on their own previous experiences [8]. In practice, the issue of problem setting (what is the goal) is as important as problem solving (how to achieve the goal) [6].

As a result of this, Schön introduced the concepts of *reflection-on-action* and *reflection-in-action*. Reflection-on-action is when one thinks back on an experience or practice and prepares possible improvements for future situations. Reflection-in-action, on the other hand, is used when unexpected problems arise and it is necessary to immediately modify one's actions based on current information or insights [8]. Thinking and action are linked together. Those who are skilled in reflection-in-action sometimes deal well with situations of uncertainty, instability, uniqueness, and value conflict [1]. It is however unclear whether reflection-in-action and reflection-on-action are different processes besides the time of their occurrence [6]. (See Table 2.1)

Unfortunately, due to the nature of Schön's work being often written in

Table 2.1 Types of Reflection by Schön [10]

Reflection-in-action	- Happens in midst of a current, ongoing event - Individuals have internal dialogues where they constantly identify and revisit a problem or experience as it occurs to the individual
Reflection-on-action	- Looks at a past event, experience, practice, or belief - Individuals have already constructed an idea or decision in their own minds

metaphors and analogies, it is difficult to find a specific definition of reflection in his works. The closest to a definition is the one provided in *Educating the Reflective Practitioner* [11] and shown in Definition 2.2. Simply put, it is a process through which one becomes a more skillful practitioner.

▶ **Definition 2.2.** *As I think back on my experience with [. . .], I may consolidate my understanding of the problem or invent a better or more general solution to it. If I do, my present reflection on my earlier reflection-in-action begins a dialogue of thinking and doing through which I become a more skillful [. . .].* [11, pp. 31–32]

In [1], Clarà combines the aspects where Schön and Dewey agree in their works and synthesizes a definition of reflection (see Definition 2.3), which is more in line with both their thoughts and ideas. He provides specific example of the reflection process in teaching and explains it step by step using quotes from both authors. According to Clarà, reflection is not a pedagogical ideal but an extremely common psychological phenomenon – a type of thinking which is simply part of our daily lives.

Reflection begins with an uncertain situation with specific observed events, which are then interpreted by the subject. From this situation inferences are drawn and evaluated. If the results of this inference are in line with observations, they are accepted. The inference provides coherence to the initial incongruent situation and the process of reflection is finished. If the results are not coherent, the subject must go back to the beginning and re-evaluate the situation. This process is repeated until a coherent solution is found. The process of reflection is not linear, but rather a circular process where the subject goes back and forth between different stages of the process. The process can be repeated multiple times until a satisfactory solution is found. [1]

▶ **Definition 2.3.** *Reflection is spontaneous, common, real thinking process that gives coherence to an initially incoherent and unclear situation, where situation is understood to be a convergence of events with agency (vectors), holistically experienced by the subject.* [1]

Moon's analysis of reflection in her book, *Reflection in Learning and Professional Development: Theory and Practice* [12] offers a comprehensive exploration of reflective practice, underscoring its central role in both the acquisition and transformation of knowledge. In her work, she dissects reflection as a nuanced mental activity – one that goes beyond mere contemplation to actively engage with complex, unstructured challenges. In chapter 12, Moon illustrates how reflection operates at multiple levels: it not only initiates learning but also helps articulate and refine one's understanding, paving the way for ongoing professional development.

Drawing on a wide array of academic conceptualizations, Moon also addresses the diverse representations of reflection found in educational literature. As she explains in her later article, reflection is not confined to its internal, cognitive aspect but is equally expressed through external modes such as writing, dialogue, or visual media. This duality is central to understanding reflective learning in formal educational settings, where structured reflective writing and other representations serve both as evidence of the reflective process and as mechanisms for enhancing learning outcomes. In this context, the process is often marked by clear intentions and desired impacts, including personal growth and actionable improvements. [13]

Moon outlines three main stages in which reflection contributes to learning: *in initial learning*, reflection helps learners make sense of new or unfamiliar material; *in the representation of learning*, it enables them to organize and express what they have understood; and *in upgrading learning*, reflection supports the critical evaluation and restructuring of existing knowledge, often leading to deeper insight or change in perspective. [12]

▶ **Definition 2.4.** *[...] reflection is a mental process with purpose and/or outcome in which manipulation of meaning is applied to relatively complicated or unstructured ideas in learning or to problems for which there is no obvious solution.* [12, p. 161]

In synthesizing these perspectives, it is clear that Moon's contribution lies in her balanced integration of theory and practice. She emphasizes that reflection is both an internal process and one that benefits greatly from its external expression – whether through writing, speech, or other creative forms – thus fostering a deeper understanding and enabling meaningful change in both learning and practice. This integrated view supports the idea that reflective practice is essential not only for personal insight but also for progressive professional and educational development.

For the purpose of this thesis, the definition 2.4 made by Moon will be used. Questions which prompt the reflection process, as per the definition, will be for the purposes of this thesis called *reflection questions*.

2.1.2 Reflection in Learning

Reflection has been proven to be an effective tool for enhancing learning and teaching. It can help students to better understand the material they are learning, to identify their strengths and weaknesses, and to develop critical thinking skills. Reflection can also help teachers to improve their teaching methods and to better understand the needs of their students. [10]

Reflection should not be classified as a task or a goal in itself, but rather as a process which is part of the learning experience, intertwined with teaching and learning. It should be used to enhance the learning experience and to help students develop their critical thinking skills. It should not be used just for the sake of using it, but rather as a tool to help students learn and grow. [10]

The learners themselves should know what reflection is and why they should use it [10]. However, simply stating the definition may not be enough to ensure effective application of reflective practices. For such problems, many different educational frameworks exist (such as Gibbs' reflective cycle, Kolb's experiential learning cycle, the 5Rs, etc.) which can help students and teachers to understand the process of reflection. Teachers who wish to use reflective practices in their teaching should first understand the process of reflection themselves. This is especially important for students who may not be familiar with the concept of reflection or who may not have had much experience with it in the past. [13]

Reflection can be either a guided process or a self-directed process. Thanks to this, students can use reflection questions to reflect on their own learning and understanding of the material. However they need to be taught how to do this. Writing reflective journals, for example, can be very helpful. [13]

Teachers can apply the process of reflection on themselves as well, taking on the role of the reflective practitioner, to improve the quality of student materials and avoid possible confusion in the students' understanding. Using test or quiz results, they can find out what confuses the students, their common mistakes and misconceptions. [14] Then, they can use one of the frameworks for reflection or simply ask themselves – *What is the source of the problem? How can this be improved?* By following the thinking process of the student they can arrive at an ambiguous word in a definition, a different better approach, a lack of examples, etc. The teacher reflects on his teaching and can later provide better service to his students. This way, they can iteratively improve their teaching and provide better materials for the students. [1]

2.2 Educational Models

In the area of education, various educational models and frameworks have been developed to help teachers and students learn and simulate the process of reflection. These models try to provide a structured approach to reflection and can be used to guide the reflection process. Some of the most well-known educational models include the Socratic method, Bloom's taxonomy, and Gibbs' reflective cycle. Several of these models will be discussed in this section. The models are not mutually exclusive and can be used together to create a more comprehensive approach to reflection.

2.2.1 Socratic Method

The Socratic method or Socratic questioning is a form of cooperative argumentative dialogue between individuals. It is a technique which can be used to pursue thought in diverse directions and to stimulate critical thinking. The method has many applications in many fields such as education and law [15], psychotherapy [16] and philosophy [4].

The method originates from the ancient Greek philosopher Socrates, who used it to help his students think critically about their beliefs and assumptions. It is based on the idea that asking questions can help individuals clarify their thoughts and beliefs, and that this process of questioning can lead to a deeper understanding of a topic. [4]

Over the centuries from its conception, the Socratic method has evolved into various different methods, distinct yet sometimes overlapping in their names and definitions. The term *Socratic method* is often used to refer to a variety of different techniques and approaches, which can make it difficult to pin down a specific definition. Some authors use the term to refer to any type of questioning or dialogue that encourages critical thinking, while others use it more narrowly to refer to the specific method developed by Socrates himself [17]. For the purposes of this thesis, term *Socratic method* will refer specifically to the method developed by Socrates and its modern applications in education.

In the Socratic method, as it is a dialogue, there exist two separate roles – a person who only asks questions (they will be referred to as the *teacher*) and a person who answers them (the *student*). There may be more students in the dialogue. The teacher serves as a guide, an assistant and facilitator of the process of self-discovery for the student. The questioning is systematic and disciplined, directing the course of the discussion [18]. The student is expected to think critically about their beliefs and assumptions. The teacher asks questions and examines the claims of the student to help the student clarify their thoughts and to encourage them to think more deeply about the topic at hand [4].

Some examples [18] of questions focusing on thinking used in the Socratic method include:

- **Questioning Goals and Purposes**
  – What is the purpose of this? Why are we doing this?
- **Questioning Information, Data and Experience**
  – On what evidence do you base that? Could it be false?
- **Questioning the Question**
  – What is the question really asking? Is this the most important question?
- **Questioning Inferences and Conclusions**
  – Could you explain your reasoning? Is there an alternative?
- **Questioning Concepts and Ideas**
  – Are we using the appropriate concepts?
- **Questioning Assumptions**
  – What are you assuming? Is that assumption justified? Shouldn't we rather assume the opposite?
- **Questioning Implications and Consequences**
  – What are the implications of this? What are the consequences of this?
- **Questioning Points of View**
  – What is the other point of view? How would someone from another point of view see this?

Students often believe that they are simply supposed to receive knowledge from the teacher. However, according to Socrates, this is not the case. The real learning is achieved only through self-discovery. So, the goal of the Socratic method is to help the student discover the truth on their own, rather than simply providing them with answers. [4]

As another bonus, through participation in the dialogue, the students can improve their communication skills and gain more confidence in their own abilities. They also learn to express their thoughts more clearly and effectively. This is especially important in today's world, where effective communication is essential for success in many fields. [19]

The process of dialogue between teacher and student closely resembles the process of reflection and is precisely the reason why it is still used today in the field of education. It is also the reason why it will be used in this thesis to improve the quality of the generated reflection questions. One could also consider this thesis to be a test of how well LLMs can simulate the Socratic method and reflect on their own answers.

2.2.2 Bloom's Taxonomy

Bloom's taxonomy is a framework for categorizing educational goals and objectives. It was developed by Benjamin Bloom in collaboration with a group of educational psychologists in the 1950s. The taxonomy was first published in 1956 in a book titled *Taxonomy of Educational Objectives: The Classification of Educational Goals* [20]. The taxonomy was revised in 2001 by a group of cognitive psychologists, led by Lorin Anderson. The revised version was published in *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* [21]. The revised version is often referred to as *Anderson and Krathwohl's Taxonomy* or simply *Revised Bloom's Taxonomy*.

The original taxonomy was created to provide a common language for educators to classify and assess learning objectives. It was designed to help teachers create clear and measurable learning outcomes for their students. The authors split the educational objectives into three domains: *cognitive*, *affective*, and *psychomotor*. Each domain specializes in a different aspect of learning and development. The cognitive domain focuses on mental skills and knowledge, the affective domain focuses on emotions and attitudes, and the psychomotor domain focuses on physical skills and coordination. [20]

The cognitive domain is the most well-known and is often used to classify learning objectives in education. It consists of six levels of cognitive skills, arranged in order from lower-order thinking skills to higher-order thinking skills. Mastering of the lower-order skills is necessary to progress to the higher-order skills. The six levels of the cognitive domain in order from the lower to higher are: *Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation.* The original taxonomy is shown in Table 2.2. [20]

Table 2.2 Original Bloom's Taxonomy Categories of Cognitive Domain (lower-order at the bottom, higher-order at the top)

Category	Description
Evaluation	Make judgments about information.
Synthesis	Combine parts to form a new whole.
Analysis	Break down information into parts, determine relations and principles.
Application	Use information in new situations.
Comprehension	Understand the meaning of information.
Knowledge	Recall facts and concepts without necessarily understanding them.

The revision of the taxonomy was created to address some of the limitations of the original version and to reflect changes in the field of education. It instead introduces a two-dimensional approach to the taxonomy, with the *cognitive dimension* and the *knowledge dimension*. The cognitive dimension closely resembles the original taxonomy, but with changes to the names and descriptions of the levels. Two levels were also rearranged, with *Evaluation* being moved to a higher level than *Synthesis*. The names were replaced with verbs, to emphasize the active nature of learning. The six levels of the cognitive dimension are: *Remember, Understand, Apply, Analyze, Evaluate, and Create*. The revised taxonomy with brief explanation of each level is shown in Table 2.3. [21]

The levels of the new taxonomy also are not as rigidly defined as in the original version, with more flexibility in how they can be applied. Some cognitive processes of a lower order may overlap into a higher order. For example *explaining* (part of *Understand*) is more cognitively complex than *executing* (part of *Apply*). [22]

Table 2.3 Revised Bloom's Taxonomy Categories (lower-order at the bottom, higher-order at the top)

Category	**Description**
Create	Produce new or original work.
Evaluate	Justify a stand or decision.
Analyze	Draw connections among ideas.
Apply	Use information in new situations.
Understand	Explain ideas or concepts.
Remember	Recall facts and basic concepts.

The knowledge dimension is a new addition to the taxonomy born from the old dual version of Knowledge. Knowledge level of the old taxonomy consisted of two types of knowledge: *knowing* and *recognizing knowledge*, but existed only as single level. The revised version creates the knowledge dimension, which is a separate dimension that describes the different types of knowledge that students can acquire. It consists of four categories: *Factual, Conceptual, Procedural, and Metacognitive*. The revision also emphasizes the importance of metacognition and self-regulation in the learning process, which was not explicitly addressed in the original version [22]. The knowledge dimension is shown in Table 2.4.

Table 2.4 Knowledge Dimension of Revised Bloom's Taxonomy

Category	**Description**
Factual	Knowledge of terminology and specific facts.
Conceptual	Knowledge of classifications and categories.
Procedural	Knowledge of subject-specific skills and algorithms.
Metacognitive	Knowledge of self-regulation and cognition.

Together these two dimensions form a table with the cognitive processes on one axis and the types of knowledge on the other. This table can be used to create learning objectives and assessments that are aligned with the goals of the curriculum. By allocating activities planned for the students into the table, teachers can ensure that they are addressing all levels of cognitive processes and types of knowledge. This can help to create a more balanced and comprehensive approach to teaching and learning. [22]

The process of reflection therefore aligns with the higher-order thinking skills of Bloom's Taxonomy, as it requires individuals to analyze their experiences, evaluate their actions, and create plans for future improvement. By guidance of the framework, teachers can create reflection questions that encourage students to engage in deeper thinking and self-assessment. Thus the framework can aid teachers in the creation of reflection questions for their students.

2.2.3 Gibbs' Reflective Cycle

Another widely recognized framework, specifically designed for reflection, is Gibbs' Reflective Cycle. This model provides a structured approach to reflection, guiding individuals through a series of stages to analyze their experiences and learn from them. The cycle is especially suitable for repeated experiences. [23]

The Gibbs' Reflective Cycle was developed by Graham Gibbs in 1988 and is based on the work of Kolb's Experiential Learning Cycle. It consists of six stages: *Description, Feelings, Evaluation, Analysis, Conclusion, and Action Plan*. The cycle is often represented as a circular diagram, with each stage leading to the next (as shown in Figure 2.1). The stages are not linear and can be revisited as needed. The cycle helps individuals to reflect on their experiences by asking them questions in a systematic way based on the current stage. The completion of one rotation of the cycle can improve their practice in the future. [23]

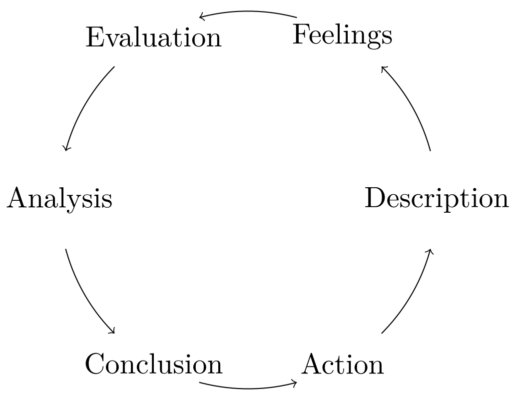

Figure 2.1 Gibbs' Reflective Cycle

The six stages of Gibbs' Reflective Cycle with sample questions [24] for each are as follows:

- **Description** of the experience
  – What happened? Describe the experience in detail, including the context and any relevant background information.
- **Feelings** and thoughts about the experience
  – What were your thoughts and feelings during the experience? How did you feel about the situation?
- **Evaluation** of the experience
  – What was good and bad about the experience? What went well and what didn't?
- **Analysis** to make sense of the situation
  – Why did things happen the way they did? What were the underlying causes of the situation?
- **Conclusion** about what you learned
  – What did you learn from the experience? How can you apply this learning in the future?
- **Action Plan** for the future
  – What will you do differently next time? How can you improve your practice based on what you have learned?

The cycle is often used in education and professional development to help individuals reflect on their experiences and improve their practice. It can be applied to a wide range of situations, including teaching, learning, and professional development. The cycle encourages growth and development by prompting individuals to think critically about their experiences and to consider how they can improve their practice in the future. It is a valuable tool for educators, students, and professionals alike. [23]

2.3 Automated Question Generation

The problem of AQG has been studied for many years and various methods have been proposed to generate questions automatically. The goal of AQG is to create questions that are relevant, meaningful, and appropriate for the intended audience. The generated questions should also be grammatically correct and easy to understand. Systems exist for different categories of questions, such as conversational, visual or standalone questions. [25]

As the focus of this thesis is on the generation of standalone questions, the models which would fall under the category of visual questions are not relevant to this thesis. A brief overview of different models used in AQG follows.

2.3.1 Rule-Based

In rule-based AQG, systems apply handcrafted predefined patterns or templates to transform declarative text into questions. These systems often use a combination of natural language processing (NLP) techniques and linguistic knowledge to identify the relevant parts of the input text and generate questions based on them. These methods typically use syntactic parses (POS tags, dependency trees) and semantic features (semantic roles, named entities) to identify answer phrases and insert appropriate wh-words. [25]

Rule-based systems often handle each interrogative form separately, guided by linguistic rules. Such methods require significant expert effort and are often domain- or language-specific but can incorporate rich grammatical knowledge to ensure grammatical correctness. Evaluation of rule-based AQG has shown that, with careful design, questions can approach neural outputs in form, though human assessments often reveal differences in fluency and answerability. [26]

These systems are often limited in their ability to generate diverse and complex questions, as they rely on a fixed set of rules and templates. Additionally, rule-based systems can be difficult to scale and maintain, as they require a significant amount of manual effort to create and update the rules and templates [27]. However, the explicit definition of the rules and templates can also be considered to be their large benefit, as they can be more easily interpreted and understood.

Dhole and Manning (2022) present *Syn-QG*, a purely rule-based system using Universal Dependencies and shallow semantic frames [28]. They define transparent syntactic transformation rules and leverage PropBank/VerbNet annotations: for each verb predicate, they generate questions by rearranging its arguments according to predetermined templates. To improve grammaticality, Syn-QG applies back-translation, translating rule-generated questions to another language and back to correct fluency. In evaluation, their pipeline yields many correct, semantically rich questions and outperforms prior rule systems, demonstrating that well-engineered transformations and semantics can generate diverse questions without neural learning.

Leite and Cardoso (2023) develop a Portuguese rule-based WH-question generator [26]. Their system first identifies which interrogative (who, what, where, etc.) applies, then transforms declarative sentences into questions using handcrafted patterns. The pipeline exploits five linguistic aspects: part-of-speech tags, semantic role labels, dependency relations, discourse connectors, and relative pronouns. For example, to ask “How did X happen?”, the system recognizes a causal conjunction or verb and applies an appropriate template. The authors conduct both automatic and human evaluation: they compare 150 rule-generated questions to gold ones using metrics (BLEU, ROUGE, METEOR) and also have people judge well-formedness and answerability. They find that the rule-based outputs achieve reasonable similarity to human questions and that human judges rate many as grammatical, though domain expertise is needed to cover specific terminology. Their work shows that rule-based pipelines can still produce high-quality questions within a targeted pedagogical domain, especially when thoroughly engineered and evaluated.

Keklik et al. (2019) present a comprehensive rule-based AQG framework that integrates dependency parsing, named-entity recognition (NER), and semantic role labeling (SRL) to generate both shallow and deep questions from single sentences [29]. Their system first preprocesses input by expanding contractions and removing idioms to improve parsing accuracy. In the deconstruction phase, it employs SpaCy for dependency and NER tagging alongside AllenNLP for SRL to extract predicate-argument structures. Three sets of templates are then applied: dependency-based, NER-based and SRL-based. Each template prescribes how to rearrange and transform sentence constituents – subject, verb, object, and adjuncts – into a grammatically correct interrogative form. They demonstrate that enabling or disabling specific templates allows selective generation of deep (why, how, purpose) versus shallow (who, what, where, how many) questions. Evaluated on SQuAD sentences, their approach achieves state-of-the-art automatic scores (e.g., METEOR 25.01, ROUGE-L 40.38) and human judges rated over 85% of generated questions as both grammatical and answerable, significantly outperforming prior template-only systems in naturalness and answerability.

2.3.2 Neural Network-Based

Neural AQG systems treat question generation as a sequence generation task. Early neural models used encoder-decoder architectures: an encoder (usually RNN/LSTM) reads the input sentence or passage (and often a target answer span), and a decoder produces the question word-by-word. Compared to rule-based methods, neural QG can generate more fluent and diverse questions without manual templates as these systems learn transformation implicitly from data. Modern approaches increasingly use deep learning enhancements: they may incorporate graph neural networks, memory modules, or multi-task learning to capture context and improve coherence across questions. In practice, many models now leverage large pretrained language models (as discussed below) but the underlying paradigm remains: encode context and answer into a latent representation, then decode into a question. [25]

2.3.2.1 Encoder-Decoder Models

A common neural framework is the answer-aware encoder-decoder. In such models, the encoder takes the passage text and a marker for the answer span. Then the decoder generates the question conditioned on that encoding.

For example, Pham et al. (2022) propose a dual-module system with a *selector* and a *generator* [30]. The selector network focuses on sentences relevant to the given answer (identifying which parts of the passage are most informative), and the generator then attends to both the selected sentences and the full context to produce a question. During training, the selector helps the encoder emphasize local answer-related content. Pham et al. show this improves quality over using a pretrained language model alone: their model outperforms strong baselines on standard QG benchmarks, demonstrating that explicit sentence selection can enhance a neural encoder-decoder pipeline.

Wang et al. (2022) introduce a *primal-dual* approach that tightly couples QG with a question-answering (QA) task [31]. Their model jointly trains a question generation component and a QA component. Concretely, the QG model encodes the answer together with the passage (concatenated or specially tagged) and decodes a question. The dual QA model then takes the generated question and passage and predicts the answer, which should match the original. This cycle ensures that the generated question indeed targets the given answer. Additionally, they include a knowledge distillation module to better generate rare or unseen words. Experiments on SQuAD and HotpotQA show this framework achieves state-of-the-art results, producing questions that not only are fluent but also better preserve the answer semantics, partly by the self-checking loop.

2.3.2.2 Deep Learning Models

Beyond vanilla sequence-to-sequence (seq2seq) models, deeper architectures exploit context structure and auxiliary tasks. Su et al. (2020) tackle multi-hop QG – generating a question whose answer requires reasoning across several sentences [32]. They propose the *MulQG* network: the encoder builds a graph from multiple evidence paragraphs and applies a Graph Convolutional Network to capture multi-hop relationships. At each "hop" the model fuses encoded context with the answer representation via a learned gating mechanism. In experiments on the HotpotQA dataset, MulQG significantly outperforms one-hop baselines – raising BLEU by over 20% in multi-hop question evaluation. This shows that graph-based deep models can aggregate dispersed information to ask higher-level inferential questions.

Li et al. (2022) address the challenge of generating an entire series of related questions for a passage ("consecutive QG") [33]. They introduce a dynamic multi-task framework with one main task – generate a QA pair – and several auxiliary tasks – generating the answer, rationale, and context history for each question. By jointly training these tasks with self-reranking losses, the model learns to produce a coherent set of QA pairs that cover the document well and remain accurate. The authors further use QA data augmentation and human evaluation, demonstrating that their approach yields more complete and informative question series compared to treating each question in isolation. This work illustrates how deep learning QG can move beyond single questions to multi-turn or dialogue-like scenarios by leveraging multi-task objectives and global reranking.

2.3.2.3 Transformer-Based Models

The current dominant paradigm uses Transformer architectures and pretrained language models. A simple but effective approach is to fine-tune a pretrained Transformer as a conditional language model. Lopez et al. (2021) show that even a unidirectional model can perform paragraph-level QG with minimal extras [34]. They prepend the context and answer to the input prompt and fine-tune GPT-2 to generate a question. Their model outperforms many complex baselines on SQuAD and produces natural-sounding questions with little engineering overhead. This highlights how transfer learning from LLMs can yield strong QG systems through straightforward fine-tuning.

Duong-Trung et al. (2024) propose *BloomLLM*: a ChatGPT-3.5-turbo model fine-tuned with Bloom's taxonomy to generate educational questions [35]. BloomLLM is trained on ∼1 000 curriculum questions spanning 29 topics labeled with competency levels. By fine-tuning for each taxonomic level, it learns to produce semantically coherent questions at all cognitive depths. The authors report that BloomLLM even outperforms ChatGPT-4 with carefully designed prompts in generating on-target questions for varied topics. This shows that supervised fine-tuning of LLMs for domain-specific question styles can yield substantial gains over vanilla models.

Mucciaccia et al. (2025) demonstrate transformer-based QG for closed-domain multiple-choice questions [36]. They use LLM prompt engineering to automate both MCQ creation and evaluation in the domain of university regulations. Their generation system consists of two modules - Preliminary Question Module and Question Review Module. The first module generates the questions and the second provides a validation of the generated questions. The evaluation system is built on another LLM, to provide a systematic way to measure the performance of theirs and other QG systems. Their experiments showed the system to be effective in the generation of valid MCQs, though they note that if the LLM used in generation does not have enough information about the target topic, it can generate incorrect questions.

2.3.3 Evaluation of AQG Systems

Evaluation of AQG systems is a critical aspect that ensures the quality and effectiveness of generated questions. However, no standardized evaluation framework exists for AQG and as such, metrics commonly used in NLP tasks are often adopted. These include BLEU, ROUGE, METEOR, and BERTScore. Besides automatic evaluation, human evaluation is also employed to assess the quality of generated questions. Automatic evaluation metrics are often used to provide a quick and objective assessment of the generated questions, while human evaluation is used to provide a more comprehensive assessment of the quality of the generated questions. [37]

2.3.3.1 Automatic Evaluation

Automatic evaluation metrics provide a quantitative measure of the quality of generated questions, allowing for scalable and reproducible assessment. These metrics are based on comparing the generated questions to reference questions, typically using overlap of n-grams (sequences of n words) or semantic similarity measures. This need for reference questions is a significant limitation, as it requires a large set of high-quality reference questions to be available for each domain or task. These metrics can be biased towards n-gram matching, rather than fluency, coherence, or content of the question. [25]

Some of the most commonly used automatic evaluation metrics for AQG include:

- **BLEU (BiLingual Evaluation Understudy)** [38]: BLEU is a precision-based metric that compares n-grams in the generated question to n-grams in reference questions. It uses a brevity penalty to penalize short sentences, but it does not consider semantic similarity or context. For a given value of n, the precision is computed as the fraction of n-grams in the generated hypothesis which match some n-gram in the reference hypothesis. The final BLEU score is the geometric mean of n-gram precisions for n from 1 to N (typically 3 or 4).

- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** [39]: METEOR utilizes both precision and recall, considering synonyms and stemming. It computes a score based on the alignment of words between the generated and reference questions, allowing for some flexibility in matching. The final score is a combination of precision, recall, and a penalty for fragmentation (i.e., how many chunks the matched words are in).

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** [40]: ROUGE is a recall-based metric that measures the overlap of n-grams between the generated and reference questions. It is often used for evaluating text summarization but can also be applied to AQG. ROUGE-N computes the recall of n-grams, while ROUGE-L considers the longest common subsequence between the generated and reference questions.

- **BERTScore** [41]: BERTScore uses contextual embeddings from BERT to compute the similarity between the generated and reference questions. It captures semantic similarity by comparing the cosine similarity of the embeddings of words in the generated and reference questions. BERTScore is more robust to lexical variations and can capture semantic meaning better than n-gram-based metrics.

2.3.3.2 Human Evaluation

Human evaluation provides a more nuanced assessment of the quality of generated questions, particularly for open-ended ones. Evaluation is typically performed by human annotators who assess the generated questions based on various criteria and scales. However, this type of evaluation is often costly, time-consuming and subjective, leading to variability in results. Human evaluation is often used in conjunction with automatic metrics to provide a more comprehensive assessment of the generated questions. [25]

The criteria differ between studies [26, 36, 42], but common ones include:

- **Format**: The format of the generated question, including whether it is a yes/no question, multiple-choice question, or open-ended question.
- **Grammar/Fluency**: The grammatical correctness and naturalness of the generated question.
- **Relevance**: The degree to which the generated question is relevant to the given context or answer.
- **Answerability**: The ability of the generated question to be answered based on the input text or topic.
- **Complexity**: The cognitive complexity of the generated question, often assessed using frameworks like Bloom's taxonomy.
- **Difficulty**: The difficulty level of the generated question.

Recent advances in AQG evaluation have seen the emergence of more sophisticated approaches, including the use of LLMs to automate the human evaluation process. These approaches aim to combine the scalability of automatic metrics with the nuanced understanding of human evaluators. Such methods leverage the capabilities of LLMs to act as human evaluators. [36]

2.3.4 Baseline Models

For the evaluation of the system, one baseline model was chosen for comparison. The *GPT-4o-mini* model, which is a smaller version of the GPT-4 model, provides state-of-the-art performance in many natural language processing tasks. The model is trained on a large corpus of text data and is capable of generating high-quality text in a variety of styles and formats. It serves as a control to assess the performance of the system without any additional collaborative learning or feedback mechanisms. The model is used in the same way as the student-teacher, generating questions based on the provided topic and concepts without any interaction with the teacher-educator. The same prompt and materials are provided to the model as in the system.

Separately, in collaboration with AI dětem [43], a study will be conducted to evaluate the performance of the system in a real-world educational setting. This study is set to involve teachers, who will create their own reflection questions based on the provided topic and concepts. The generated questions will then be compared with the questions created by the system. The results of this study will be used to assess the effectiveness of the system in generating high-quality reflection questions and its potential impact on the educational process. Due to time constraints, the results of this study will not be presented in this thesis.

Chapter 3
Methodology

3.1 System Architecture

The system proposed in this thesis is designed to generate reflection questions using a collaborative approach between two LLMs. The architecture is based on the principles of collaborative learning, specifically the Socratic method, which emphasizes dialogue and inquiry as a means of fostering understanding and reflection.

Each LLM serves a distinct role in the process: one acts as a *student-teacher* and the other as the *teacher-educator*. The student-teacher in education is someone who is learning how to become a qualified teacher and the teacher-educator is the ”teacher of teachers”. He tutors the student-teachers how to teach students.

In this system, the student-teacher is responsible for creating the reflection question. He is learning how to create a proper reflection question about the given topic. The teacher-educator is responsible for improving the quality of the created question, checking if the question is understandable and relevant. However, he does not simply state what is wrong with the question or what should be changed and how it should be done. In the eyes of the Socratic method, the teacher-educator is in the role of the questioner – he asks the student-teacher guiding open-ended questions to facilitate deeper thinking and reflection. This process encourages the student-teacher to critically analyze their own questions and refine them for better clarity and depth.

At the start of the process, the human teacher provides materials and topic of the question. The system then operates in a cyclical manner, where the student-teacher generates a question, the teacher-educator provides feedback through guiding questions, and the student-teacher revises the question based on this feedback. The teacher-educator then evaluates the revised question and provides further feedback, continuing the cycle of inquiry and refinement. This iterative process allows for multiple rounds of question generation and refinement until a satisfactory reflection question is produced.

The steps of the process are as follows:

1. A human teacher provides a topic and concepts he wishes to incorporate in the question.
2. The student-teacher formulates a reflection question based on the provided topic and concepts.
3. The teacher-educator reviews the question and asks guiding questions to enhance its quality.
4. The student-teacher revises the question based on the feedback received.
5. The teacher-educator evaluates the revised question and provides further feedback.
6. The steps 4 and 5 are repeated until an acceptable reflection question is generated.
7. The final reflection question is presented to the human teacher for approval.

The interaction between these components is illustrated in Figure 3.1.

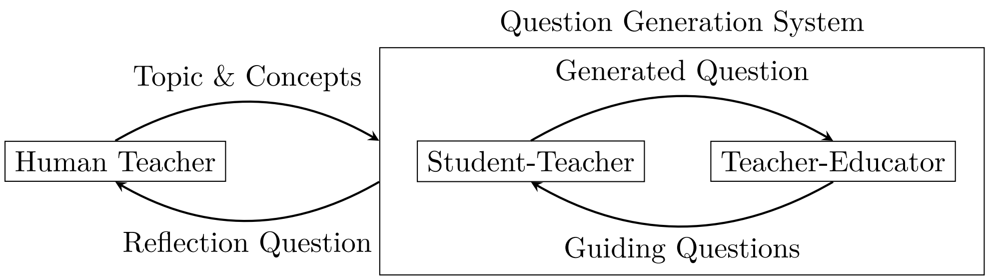

Figure 3.1 System Architecture

The system stores the entire dialogue between the student-teacher and teacher-educator, from both sides, allowing them to refer back to previous questions and answers. This helps maintain context and continuity in the conversation, enabling the LLMs to build on previous interactions and refine the questions more effectively. The dialogue history is also useful for the human teacher, who can review the interactions and understand the reasoning behind the generated questions.

For the LLMs used in the system, *GPT-4o-mini* from OpenAI was chosen. This model is a smaller version of the GPT-4 model, which is known for its ability to generate high-quality text and understand complex prompts. The choice of this model was made based on its performance in generating coherent and contextually relevant questions, as well as its ability to adapt to different styles and tones of writing. It also has a good balance between pricing and performance, making it a suitable choice for this application.

The models used in the experiments are not trained or fine-tuned on the task of question generation. Instead, they rely on the knowledge already present in the models. The prompts may contain some information about the task, but the model is not specifically trained on it. This allows for a more flexible and adaptable system that can generate questions based on a wide range of topics and contexts.

This approach was chosen because existing datasets usually used for question generation, such as *SQuAD* and *TriviaQ*, are not suitable for generation of open-ended questions. These datasets contain information about the question which should be created, from what context the question is made and also an answer to the question. The specific structure of the dataset depends on the type of question being generated. Datasets for multiple-choice questions, for example, also typically include distractors aside from the correct answer.

The dataset that possibly could have been used is the *SocratiQ* dataset, which focuses on open-ended questions and provides a more suitable context for our needs. However, the questions contained in the dataset are simply open-ended and not necessarily reflective. Filtering out the irrelevant questions and ensuring they align with the desired outcomes would take significant effort and careful consideration to ensure quality and relevance.

Thus, the decision was made to use the inherent knowledge of the LLMs to generate the questions. A downside of this approach is a more uncontrolled variety of responses, which may lead to less consistent quality in the generated questions. However, this is mitigated by the collaborative approach between the two LLMs, where the teacher-educator provides feedback and guidance to refine the questions.

For ease of testing and evaluation, the system was implemented using Python and Jupyter notebooks, which are widely used for data analysis and machine learning tasks. The system is designed to be modular and extensible, allowing for easy integration of new features and improvements in the future. The use of Jupyter notebooks also allows for easy visualization and analysis of the generated questions, making it easier to assess their quality and relevance.

3.2 Prompt Engineering

Prompt engineering is a crucial aspect of the system, as it determines how effectively the LLMs stick to the given task and generate relevant questions. Several different prompts were designed with different levels of detail and specificity and were iteratively refined based on the performance of the LLMs.

Three different types of prompts were created for the system. These prompts were used with the LLM role system, which categorizes each message in the conversation as *system*, *user*, or *assistant*. The *system* message is used to set the behavior of the assistant, while the *user* message is used to provide input to the assistant. The *assistant* message is used to provide the output of the assistant. The user and assistant messages, in this case, simulate the conversation between the student-teacher and teacher-educator. The teacher-educator and student-teacher both have their own system prompts. The teacher-educator has only a single one and the student-teacher has two similar but slightly different prompts, only one of which is used depending on the current state of the dialogue.

As mentioned previously, two instances of the dialogue are stored for each of the LLMs to facilitate the correct behavior of each model. In a normal conversation, between an LLM and a human user, the LLM sees itself as the assistant and the human as the user. To make sure that the responses are taken in the same way in the system, each instance of the dialogue has the assistant and user roles swapped. This way the LLMs are not confused by the fact that they are both in the role of the assistant. So from the teacher-educator's perspective, the student-teacher is the user and the teacher-educator is the assistant. The same applies to the student-teacher in reverse.

All the system prompts contain the specification of the role of the LLM in the conversation, to ensure that the LLM understands its role in the conversation. The topic and its important concepts are crucial parts of the prompt, as they provide the context necessary for the process. The concepts serve as the direction for the model as the overall topic can be quite a broad term and the resulting questions may then vary significantly or not even touch the desired subject. Both models also know that this is part of the Socratic dialogue, so they are aware that they are not simply generating questions or providing feedback, but rather engaging in a collaborative process of inquiry and reflection.

The student-teacher prompt has two different versions, because the student-teacher has to perform two different tasks during the process. The first task is the initial creation of the question, when there is nothing to revise or improve. The second task is the revision of the question based on the feedback provided by the teacher-educator. As such, the student-teacher prompt is divided into two: the first prompt is used for the initial question generation, while the second prompt is used for the revision of the question. The difference between these two prompts is in the goal of the task but the overall structure is the same. After the question is created, the system prompt is replaced in the dialogue history with the second prompt.

Several issues were encountered during the development of the student-teacher prompts. The most significant one was the creation of overly complex questions as the student-teacher tried to connect and mention as many concepts as possible provided in the prompt as part of the question. This often led to confusing and unintelligible questions. To mitigate this, the prompt was designed to encourage the student-teacher to focus on a single concept or a few closely related concepts, as well as asking the LLM to choose the concepts randomly from the provided list to additionally ensure a variety of different questions.

The teacher-educator prompt is designed to guide the student-teacher in the process of improving the quality of the generated question. It contains a few examples of Socratic questions, which guide the teacher-educator in formulating questions. The teacher-educator is also aware of the aspects important for the quality of the generated question (see 3.1). This allows the teacher-educator to provide targeted feedback and ask specific questions that help the student-teacher refine their question.

Table 3.1 Aspects of Reflection Questions

Aspect	Description
Clarity	Is the question clear and understandable?
Depth	Does the question encourage deep thinking and exploration of the topic?
Relevance	Does the question connect to the key concepts provided?
Engagement	Will the question engage students and provoke discussion?
Interconnections	Does the question encourage students to explore the relationships between different concepts?

However, it was found that asking too many questions at once can overwhelm the student-teacher, leading to confusion and difficulty in responding. To mitigate this, the teacher-educator is asked to provide a single question in each feedback message, allowing the student-teacher to focus on changing one aspect of the question at a time.

3.2.1 Prompt Variants

For the experiments, several variants of the prompts were created to test their influence on the quality of the generated questions. These variants were made for the entire group of prompts (both the student-teacher prompts and teacher-educator) to ensure consistent behavior.

The following parameters were considered for the prompt variants:

- **Materials:** These are additional materials provided to the LLMs, which can include lesson plans, teaching strategies, and other resources that the teacher may use in their class. These materials are not part of the topic and concepts provided by the human teacher, but rather provide extra information that can help the LLMs generate better questions. To make these files accessible to the models, they were uploaded using the OpenAI API's file handling capabilities. When the materials are present, the LLMs are instructed to utilize them to create the question. However, their usage comes with a price, multiplying the cost and time taken for each LLM call by adding a lot more tokens to the prompt.

a) **None:** No additional source materials are provided to the LLMs. The topic and concepts are the only information available.
b) **With materials:** Additional source materials are provided to the LLMs.

- **Student level:** The student level is a parameter that indicates the expected knowledge and skills of the students for whom the question is intended. This information is provided in terms of a grade level, such as 1st grade, 2nd grade, etc. The LLMs are informed about the student level and are asked to create a question that is appropriate for that level. This parameter is important because it could significantly influence the complexity and depth of the generated questions.

  a) **None:** No information about the student level is provided to the LLMs.
  b) **With level:** Information about the student level is provided to the LLMs.

- **Iterations:** The most important parameter tested was the number of iterations of the dialogue. In the system, the number of iterations is counted as the number of back-and-forth exchanges between the teacher-educator and student-teacher. The first call to the student-teacher is excluded from the count (it is the zeroth iteration). Each iteration then begins with the teacher-educator's feedback and ends with the student-teacher's revised question. The number of iterations is vital because it could significantly influence the quality of the questions. The system was tested with both fixed and dynamic iteration limits. The fixed iteration limit is a set number of iterations that the system will go through, while the dynamic iteration limit allows the teacher-educator to decide when to stop the conversation.

  a) **Fixed:** The system will go through a set number of iterations, regardless of the current state and quality of the question.
     a. **0 iterations:** This case simulates the option where no dialogue is performed and the student-teacher solely generates a question without any feedback from the teacher-educator.
     b. **5 iterations:** 5 iterations were chosen to test the system with a moderate number of iterations as a middle ground between 0 and 10 iterations.
     c. **10 iterations:** Based on a few testing samples, 10 iterations were chosen as the maximum number of iterations. It is expected that the quality of the questions will improve with more iterations, but it may also lead to diminishing returns as the dialogue continues.
  b) **Dynamic:** The teacher-educator can decide when to stop the conversation, if he deems the question to be of sufficient quality. This ability

would be very useful to ensure that the dialogue is efficient and avoids unnecessary iterations. Since the ideal question could fall just below or above the set iteration limit, specifying a fixed number of rounds may reduce overall quality.

The combinations of these parameters were used to create different variants of the prompts, which were then tested in the system. The goal was to determine which combination of parameters leads to the best quality of generated questions. The results of these tests are discussed in detail in Chapter 4.

3.3 System Evaluation

Evaluation of the system is based on the criteria defined in section 2.3.3. Automatic metrics could not be used, because, as was mentioned previously, no suitable dataset was found and no gold standard questions created by experts for the topics were available. Thus, evaluation was performed using a combination of LLM and human evaluation.

For evaluation, the *o4-mini* model from OpenAI was used. This model was selected for its strong reasoning capabilities and cost-effectiveness compared to larger models like GPT-4, making it suitable for the large number of comparisons required in the evaluation process. It is important to note that this evaluation model (*o4-mini*) is distinct from the *GPT-4o-mini* model used for the question generation system itself. This separation ensures that the evaluation is performed by a different model than the one that generated the questions, mitigating potential bias where a model might favor its own output and providing a more objective assessment of the generated questions' quality.

The criteria used were chosen based on the works found in the literature review and the specific needs of the system. Criteria such as format, grammar, and fluency were consolidated into the clarity criterion. Relevance was kept to measure how pertinent the question is to the topic. Depth was added as it is an important aspect of reflection questions – deeper questions provide better opportunities for reflection. Lastly, the criterion of overall quality was added to simply measure the question as a whole. Thus, the main criteria are as follows: clarity, relevance, depth, and overall quality.

Other criteria would surely provide useful additional information about the generated questions, but they were not included in the evaluation due to time constraints and the need to focus on the most relevant aspects. The chosen criteria were deemed sufficient to assess the quality of the generated reflection questions and ensure that they meet the desired standards. The criteria in more detail are as follows:

- **Clarity:** The question should be clearly stated and easy to understand. It should not contain any grammatical errors or typos that could confuse the reader. There should not be any ambiguous or confusing language that could lead to misinterpretation.

- **Relevance:** The question should be relevant to the topic and concepts provided by the human teacher. It should connect to the key ideas and encourage exploration of the subject matter.
- **Depth:** The question should encourage critical thinking and exploration of the topic. It should not be a simple yes/no question, but rather one that prompts further discussion and reflection.
- **Overall Quality:** The question should be of high quality, well-structured and thought-provoking. It should engage students and provoke discussion.

The aforementioned criteria were then adapted to be suitable for use in the evaluation prompt. To ensure that the LLMs understand the criteria and can evaluate the questions correctly, the criteria were rephrased and presented in a more structured format as a single question per each criterion. The questions can be seen in Table 3.2.

Table 3.2 Questions for Evaluation Criteria

Criterion	Description
Clarity	Is the question clearly stated and easy to understand?
Relevance	Is the question relevant to the topic?
Depth	Does the question encourage critical thinking and deeper exploration of the topic?
Overall Quality	Is the question of high quality overall, well-structured and thought-provoking?

Questions were compared pairwise using a difference score, which indicates how much better or worse the left question is compared to the right question on the given criterion. The scores range from -2 to 2, skipping 0 to force the evaluator to choose one question as better. The more negative the score, the better the left question is over the right question. The more positive the score, the better the right question is over the left question. In addition to providing the numerical score, the evaluator LLM was also prompted to provide a brief explanation justifying its choice for each comparison. This explanation provides qualitative insight into the reasoning behind the quantitative scores.

An additional test was performed to find the best dialogue ending iteration if the number of iterations is fixed. In this test, all the questions from a single dialogue were provided to the evaluator, and they were asked to choose the best question from the list. The order of the questions was randomized to avoid any bias in the evaluation. The evaluator was asked to choose the best question based on the same criteria as before. This test was performed to determine if there is a specific iteration at which the quality of the questions generally peaks. The results of this test can be used to inform the choice of fixed iteration limits in future experiments.

Chapter 4
Experiments

Several experiments were conducted to evaluate the performance of the proposed system. The experiments were designed to test the system's ability to generate high-quality reflection questions and to compare its performance with other state-of-the-art automated question generation systems. The evaluation was performed using LLM methods, as human evaluation is very complex and will be conducted separately in the future.

All the experiments were conducted using the OpenAI API, which provides access to their state-of-the-art language models. The API key was graciously provided by AI dětem [43], a non-profit organization that focuses on the use of AI in education. The experiments were conducted in a controlled environment, with the same input text and parameters used for each experiment to ensure consistency and reliability of the results. All the Jupyter notebooks used for generation and evaluation of the questions are available in the attachments.

4.1 Domain

The experiments were conducted on the domain of information technology, specifically focusing on the basics of the internet, including important concepts such as decentralization, routers, servers, and IP addresses. The materials used for testing were provided by a specialist in the field of education, ensuring that the content was appropriate for the target audience of 8th to 9th-grade students. The specific topics and concepts extracted and used for testing are listed in Table 4.1.

The source materials – presentations and teacher guidelines – from which the topics were taken are available in the attachments. These are also the files that were used as part of the input for the models in the specific prompt combinations.

Table 4.1 Topics and Concepts Used for Testing

Topic	Concepts
Basics of how the internet works	- Decentralization of the internet - Servers, datacenters and routers - Server vs client - Data packets - IP addresses

4.2 Data Generation

As per the parameters mentioned in the previous chapter, a variety of combinations of prompts were used to generate the questions. For each combination, a set of 10 questions was generated. With 4 combinations (materials/no materials x student level/no student level), this resulted in a total of 40 questions. This was done twice, once for fixed number of iterations (specifically for 10) and once for dynamic number of iterations. The results are stored in JSON files, specifically *fixed10iter.json* and *free10iter.json*, which allow easy access and manipulation with the dialogues. The 0 and 5 iteration variants are simply copies of those files, but with dialogue cut down to 0 or 5 iterations (*fixed5iter.json*).

As observed in Table 4.2, the teacher-educator model consistently stopped the dialogue significantly earlier than the maximum allowed limit of 10 iterations. The dialogues typically concluded after only 2-3 iterations, indicating that the model perceived the question refinement process to be complete relatively quickly. Whether this early termination by the teacher-educator model negatively impacts the final question quality compared to continuing for a fixed, potentially longer, number of iterations is a key point that will be explored in the subsequent evaluation sections.

This observation highlights a key challenge: determining when a generated reflection question is ”good enough” and how to reliably automate this assessment, potentially using another LLM. Furthermore, it was noted that in some longer dialogues (mostly exceeding 5 iterations, observed during development and in the fixed iteration experiments), the conversation veered off the primary task of refining a single reflection question. The dialogue occasionally drifted towards broader lesson planning concepts that incorporated reflection, rather than focusing solely on question generation. While these generated lesson planning ideas might hold pedagogical value, they fall outside the defined scope of the system’s objective.

Table 4.2 Number of Iterations For Each Combination (TT = Materials & Student Level, TF = Materials & No Student Level, FT = No Materials & Student Level, FF = No Materials & No Student Level)

Iterations	TT	TF	FT	FF
0	0	0	0	0
1	0	0	0	0
2	9	7	9	7
3	1	2	1	2
4	0	1	0	1
5	0	0	0	0
...	0	0	0	0
Avg	2.10	2.40	2.10	2.40

4.3 LLM Evaluation

As was mentioned in the previous chapter, the evaluation was done using the o4-mini model. In the best case, all the questions would have been used for the evaluation, however, this could not be done due to the number of comparisons that would have been necessary. The comparisons were made for each criterion separately, to ensure the model was fully focused on its task and not biased towards a specific criterion. This in turn resulted in a very large number of comparisons to be done. For example, for the experiment between dynamic and fixed iterations, with 10 questions per combination for four criteria in 3 files (fixed 10, fixed 5 and dynamic) with approximate cost of 0.002131 USD per LLM call (tested experimentally), this resulted in 19,200 comparisons with a cost of ~40.9 USD, which was more than the total budget on the API key. Thus, for relevant tests, only the first 5 questions were used – as questions were generated independently, the questions did not need to be selected randomly.

4.3.1 Baseline Comparison

To compare the core benefit of the Socratic dialogue itself, the dynamic dialogue approach was used and compared against the baseline of generating a question with no dialogue (0 iterations). This evaluation is crucial to understand the effect of the iterative refinement process over a simpler approach, based on Neural Network AQG methods [32, 34, 35] (see Section 2.3.2). The specifics of the dynamic iteration approach and its comparison to fixed iteration strategies are detailed in the following section (Section 4.3.2).

Comparing the results of the no dialogue and the dynamic iterations, the dialogue variant was generally considered slightly better across most criteria and combinations. The dynamic dialogue approach particularly dominated in terms of overall quality. The only notable exception where the no dialogue approach was favored was for the relevance criterion in the combination using no student level information and no additional materials (see Figure 4.1).

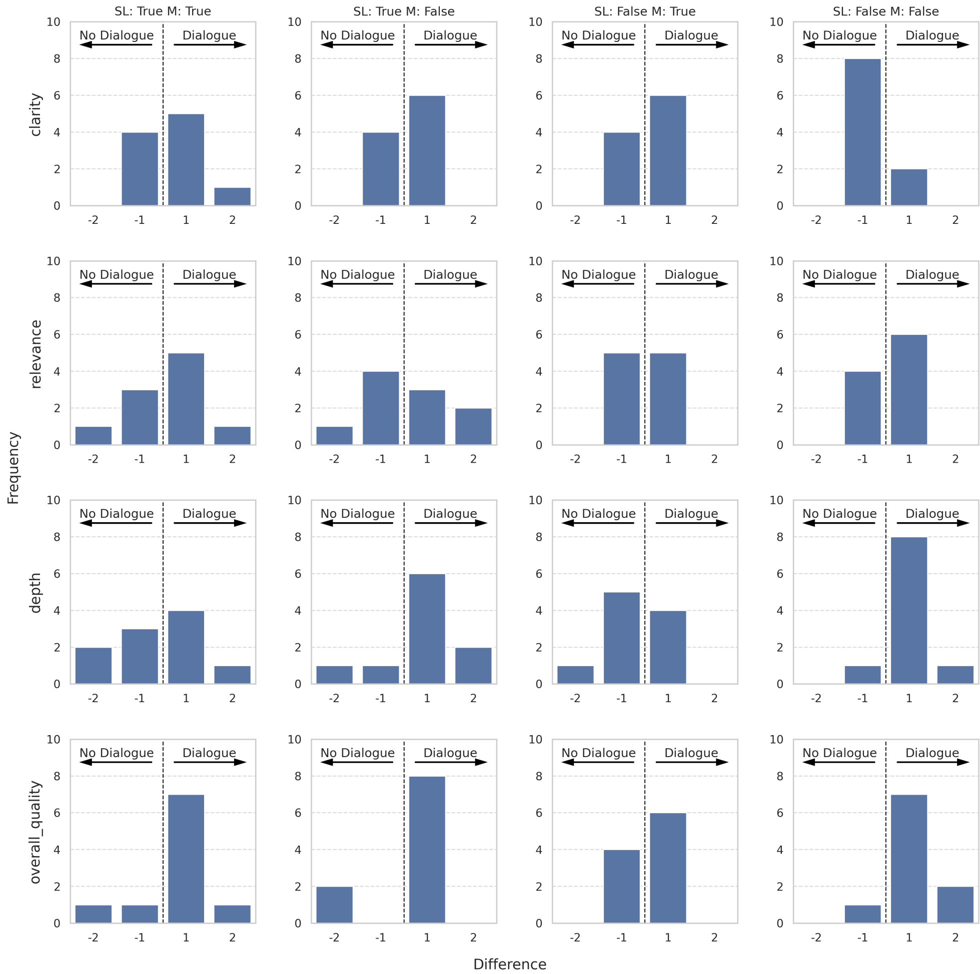

Figure 4.1 Comparison between No Dialogue (0 Iterations) and Dynamic Iterations (SL = Student Level, M = Materials)

These results show that using the Socratic dialogue generally helps improve the quality of the reflection questions compared to generating them without any dialogue. While the no-dialogue approach sometimes produced more relevant questions in specific situations, the dynamic dialogue proved to be better overall across different quality aspects.

4.3.2 Fixed VS Dynamic Iterations

In this section, a comparison is made between the results of the dynamic number of iterations and fixed numbers of iterations (specifically 5 and 10). It was necessary to establish which of the three approaches yields the best results. Given the large number of comparisons, for this experiment only the first 5 questions from each combination were used.

When comparing the dynamic and fixed 10 iterations, the results were very clear. The dynamic approach was considered to be overwhelmingly better in all four criteria for all combinations with most comparisons saying that the dynamic approach was significantly better on the scale. Only one exception arose (see Figure 4.2), where the fixed 10 iterations were considered to be deeper for combinations which excluded the usage of extra materials. This shows that for cases with no additional materials, the more iterations the dialogue has, the better it is in terms of depth.

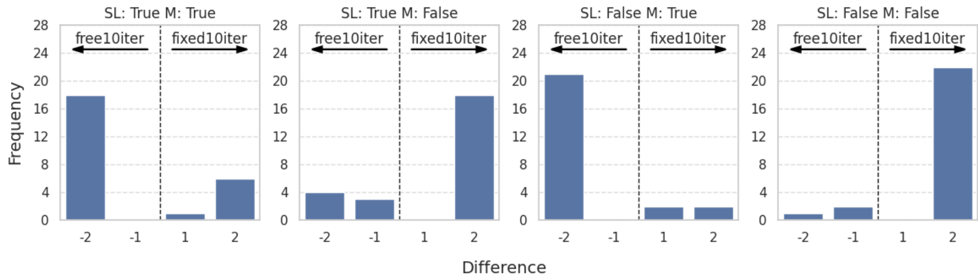

Figure 4.2 Comparison between Dynamic and Fixed 10 Iterations by Depth (SL = Student Level, M = Materials)

Results between the dynamic and fixed 5 iterations were more varied. The dynamic approach took the lead for the materials and no student level combination for all criteria, as well as being considered to be better in terms of relevance for all combinations. The pattern seen in depth criterion occurred only partly, with fixed iterations dominating only in the case of no materials and no student level.

Overall, the dynamic approach was considered better, though there was not as much of a large difference as in the case of the fixed 10 iterations. This can be attributed to the smaller distance between the number of iterations, with 5 iterations being only 2-3 iterations from the average number of iterations for the dynamic approach.

Based on the results of the previous comparisons, it was expected that the fixed 5 iterations would be considered to be better than the fixed 10 iterations. This was indeed the case, with the fixed 5 iterations being graded as better, though with a lower margin, in all criteria for almost all combinations (see Figure 4.3). The exception being, again, the depth criterion for the no materials combinations.

In total, the results suggest that the dynamic iteration approach, where the dialogue stops based on the teacher-educator's assessment, generally achieves comparable or superior question quality compared to running a fixed number of iterations (5 or 10).

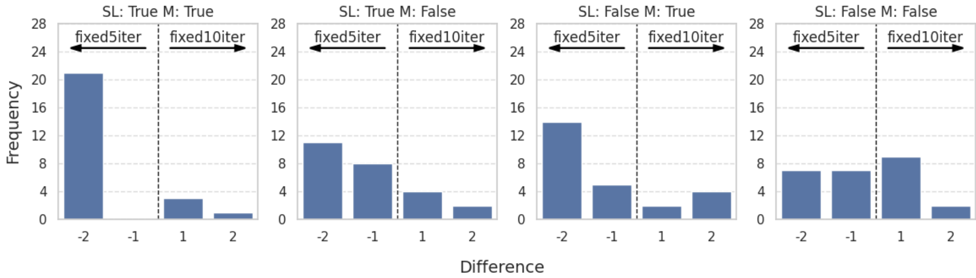

Figure 4.3 Comparison between Fixed 5 and Fixed 10 Iterations Overall Quality (SL = Student Level, M = Materials)

4.3.3 Best Iteration

In terms of finding the best iteration at which the question quality peaks, the results vary for each criterion. For clarity, questions created during the initial iterations (0-2) were considered to be the best. This result is not surprising, as it can be expected that the further the dialogue goes, the more complicated the question becomes as the student-teacher tries to improve the question. Similarly for relevance, the initial iterations (0-3) tended to yield questions that were considered to be the most relevant.

In the area of depth, the results were more varied (see Figure 4.4). The best iteration was mostly uniformly distributed between all 10 iterations. Though the pattern previously seen around the no materials variant, where the more iterations the dialogue has, the better it is considered in terms of depth, was also present here. The reason for this could possibly be that the materials provide a lot of information to work with from the start, which leads to a more complex question being generated in the initial iterations. The subsequent iterations then do not manage to make the question deeper.

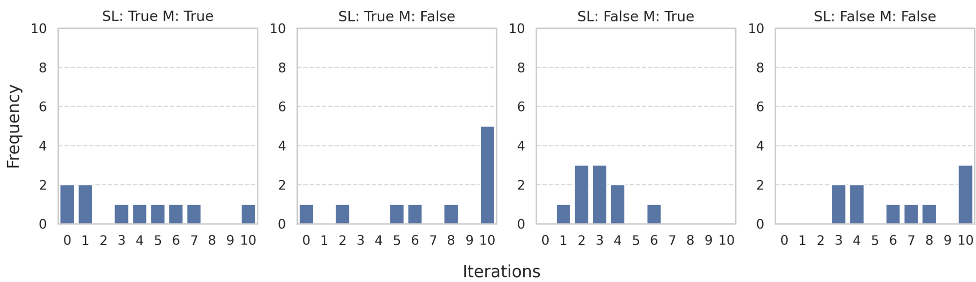

Figure 4.4 Best Iteration for Depth (SL = Student Level, M = Materials)

Overall, there is not a single specific best iteration in the results. The best iterations are mostly considered to be the lower iterations (0-4), with the exception of no student level and no materials, which is spread over a wider array of iterations (1-7). As such, more data is necessary to determine the best number of fixed iterations.

4.3.4 Best Parameter Combination

To determine the optimal parameter combination (Materials/No Materials x Student Level/No Student Level) for each of the three iteration strategies, a series of pairwise comparisons were conducted. For each strategy, the four distinct parameter combinations were compared against each other. After removing redundant comparisons, such as comparing a combination with itself or repeating comparisons (e.g., A vs B and B vs A), the number of unique pairwise comparisons for each iteration strategy was reduced to six.

For the dynamic iteration approach, the combination of using materials and student level information was generally favored for clarity (as seen in Figure 4.5). This suggests that providing the model with comprehensive context helps in generating clearly worded questions. However, for relevance, depth, and overall quality, the combination of student level without using materials was considered best. Notably, the absence of extra materials seemed to benefit relevance and depth in some cases, with student level providing a boost to relevance when materials were not used.

When examining the fixed 5 iterations, the results were more varied and less conclusive. The combination of materials and no student level consistently performed worst, particularly for relevance and depth, where it was rated poorly by a large margin. For relevance, the combinations that didn't use materials were preferred. In terms of depth, no materials and no student level information were rated best.

Overall, there was no single dominant parameter combination for fixed 5 iterations in terms of clarity or overall quality, though the combination using materials and no student level information was clearly the least effective across multiple criteria. This suggests that with a moderately fixed number of iterations, the interplay of materials and student level information is complex and does not yield a universally superior setup.

For the fixed 10 iterations, clarity was marginally best with the combination using no materials and no student level information, while other combinations performed similarly. A strong and consistent pattern emerged for relevance, depth, and overall quality: combinations that did not use additional materials were rated significantly better. This indicates that with a higher number of fixed iterations, the presence of extensive source materials might not be as beneficial, or could even be counterproductive if the dialogue does not effectively synthesize or focus the information over the extended interaction. The model might perform better by relying on its internal knowledge when forced through a longer, fixed refinement process without specific material grounding. Figure 4.6 highlights the preference for no-material combinations in terms of relevance for the fixed 10 iteration setup.

These findings across the different iteration strategies highlight that there

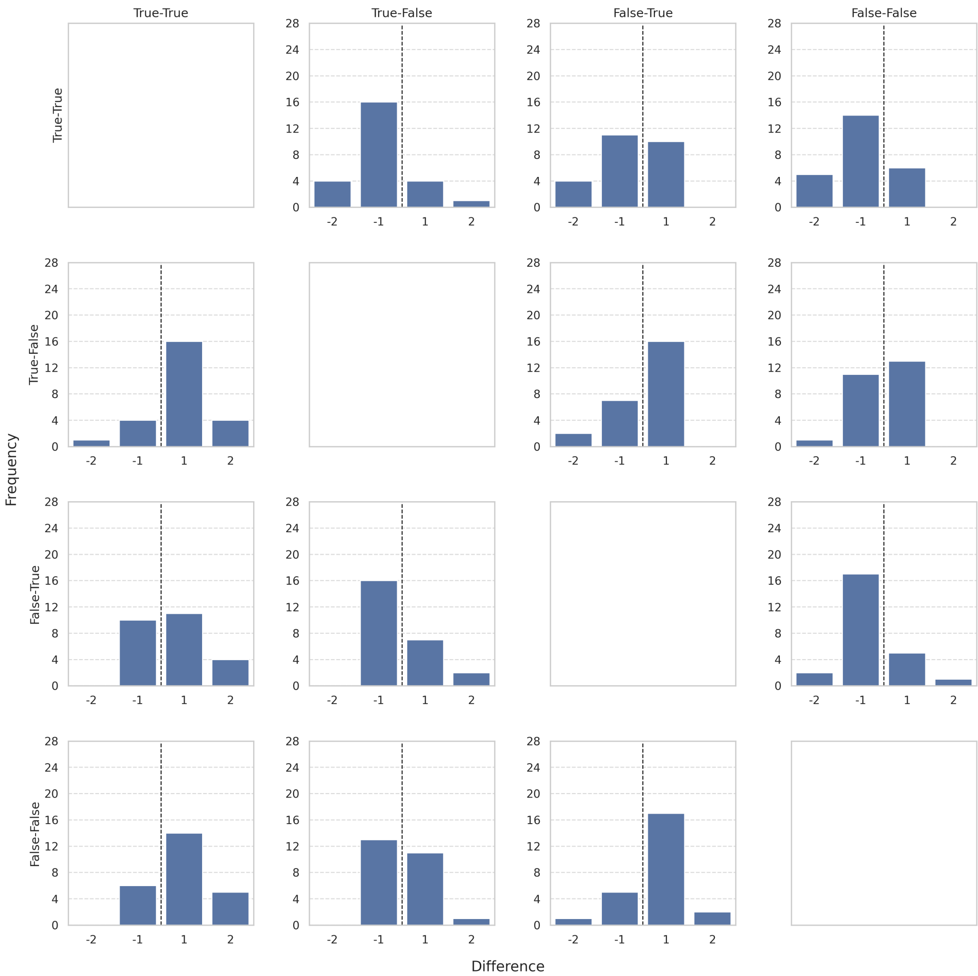

Figure 4.5 Best Parameter Combination for Dynamic Iterations - Clarity (True-True = Student Level & Materials, True-False = Student Level & No Materials, etc.)

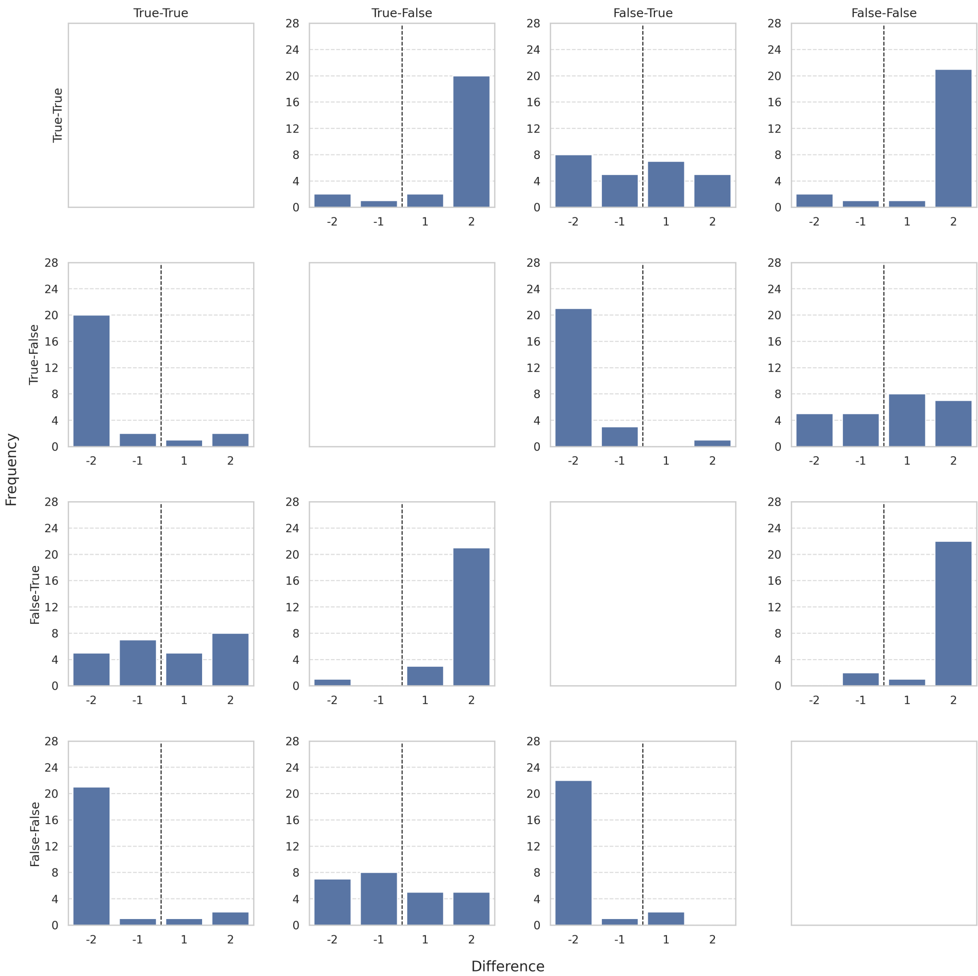

Figure 4.6 Best Parameter Combination for Fixed 10 Iterations - Relevance (True-True = Student Level & Materials, True-False = Student Level & No Materials, etc.)

is no one-size-fits-all parameter combination. The optimal setup appears to depend on the iteration strategy itself. For dynamic iterations, richer contextual information was beneficial for question clarity, though for other aspects like relevance, depth, and overall quality, omitting materials while retaining student level information yielded better results. Conversely, longer fixed iterations generally performed better with less specific input material, suggesting a greater reliance on the iterative refinement process itself when the dialogue length is predetermined.

4.3.5 Examples

All the generated questions are available in the attachments. The following examples showcase few of the generated questions, illustrating the differences in quality and relevance. The responses from both roles have a desired format, which was not always followed. Namely, student-teacher responses have to start with ”The Student’s response:” while the teacher-educator’s responses have to begin with ”The Teacher’s feedback:”. As these terms do not have a significant impact on the content, they are excluded from the shown examples.

In the first example (see Example 4.1) is shown a question considered to be one of the best by the LLM. This question was generated using the dynamic iteration approach with student level information but without any additional materials. The first section contains the question, while the second section explains the reasoning behind the changes made in the question based on the teacher-educator’s feedback.

Example 4.1 Sample LLM-generated Reflection Question and Explanation (attempt 10 in free10iter.json)

```
In what specific ways does decentralization impact the
process of data transmission over the internet from servers
to clients, and how do data packets facilitate that process?

In this revision, I replaced "influence" with "impact" and
explicitly invited students to think about "specific ways"
decentralization affects data transmission. This
clarification helps to narrow their focus on tangible
aspects of how decentralization operates in relation to the
flow of data. By highlighting the role of data packets in
facilitating this process, I encourage students to better
understand the intricate details of internet functionality.
Overall, this question promotes a more targeted exploration
of the topic while remaining accessible for 8th and 9th
graders.
```

As was mentioned in the section of data generation (Section 4.2), the longer dialogues had a tendency to go off-topic. Even though the prompts were improved several times to avoid this and the improvement was seen, it still was not enough. The LLM sometimes ignored the question revision aspect of the response and instead focused on the lesson planning aspect (see Example 4.2). However, this was not only an issue of the student-teacher, but can also be attributed to the teacher-educator's question which guided the response (see Example 4.3). Care has to be taken to ensure that the teacher-educator's question does not lead the conversation away from the main topic of generating a reflection question. In this example, the student-teacher was asked to think about how to facilitate a discussion, thus guiding the conversation away from the question generation process. Besides being a clear sign of how the teacher-educator can influence the conversation, this shows that the teacher-educator prompt has space for improvement.

Example 4.2 Response Devolving Off-Topic (iteration 7 in attempt 1 in fixed10iter.json)

```
To facilitate the discussion effectively and ensure every
student's voice is heard, I plan to implement the following
strategies:

1. Think-Pair-Share: I will first ask students to reflect
individually on the questions before discussing their
thoughts with a partner. This gives quieter students a
chance to articulate their ideas in a smaller setting
before sharing with the larger group.

2. Structured Turn-Taking: I will establish a system where
each student has an opportunity to speak, perhaps by
passing a visible object (like a talking stick) to ensure
that everyone gets a turn and encourages those who might
hesitate to share.

3. Building on Ideas: I will prompt students to respond to
each other's thoughts by asking follow-up questions, such
as, "Who agrees with this point, and can you add to it"? or
"Does anyone have a different perspective to share"?

By employing these strategies, I aim to create an inclusive
and balanced conversation, promoting a supportive
environment where all students feel empowered to express
their views and engage with their peers' ideas.
```

Example 4.3 Teacher-Educator Question Leading Off-Topic (iteration 7 in attempt 1 in fixed10iter.json)

```
How do you plan to facilitate the discussion to ensure
every student's voice is heard? Might including strategies
for promoting quieter students to share their thoughts help
create a more balanced conversation? Additionally, how
could you encourage them to build on each other's ideas
during the discussion?
```

The last two examples (see Examples 4.4 and 4.5) show how much the question can change during the dialogue (the explanations are excluded from the examples for their length). The first example is a question generated with 0 iterations (the first question generated by the model), while the second example is the same question after 10 iterations. The second question is visibly much more complex and detailed, even containing multiple sub-questions with a structured approach to the topic. In the best iteration test it was considered to be the best in terms of depth, though for other criteria the early iterations led.

Example 4.4 Reflection Question made 0 iterations (iteration 0 in attempt 15 in fixed10iter.json)

```
How does the decentralization of the internet impact the
way data packets travel from a server to a client device?

[...]
```

Example 4.5 Reflection Question changed in 10 iterations (iteration 10 in attempt 15 in fixed10iter.json)

```
What are some different opinions on how servers, routers,
and clients in a decentralized internet affect the speed
and reliability of data packets? First, identify problems
related to data congestion and redundancy. Next, suggest
solutions to these problems, and discuss the challenges or
limitations of implementing your solutions. Finally, how
might these solutions impact users' experiences with the
internet? Please provide real-world examples to support
your points.

[...]
```

Chapter 5

Conclusion

Main goals of this thesis were: First, review the literature on the topic of reflection, educational models, and existing systems for automated question generation. Second, develop an automated question generation system that generates high-quality reflection questions. Third, evaluate the system using both qualitative and (if possible) quantitative evaluation. Fourth, discuss the interpretability and practical implications of the results.

In this thesis, the use of LLMs in the field of automatic question generation has been explored. Specifically, the creation of reflection questions, that is, questions that prompt the process of reflection in the learner. Various methods of question generation were examined, from rule-based systems to more advanced methods using LLMs. However, most of the existing systems focused on generating specific types of questions, such as multiple-choice or true/false questions, rather than reflection questions. Thus, the need for a system that can generate high-quality reflection questions was identified.

Definitions of reflection and reflection questions were provided, which are essential for understanding the context and purpose of the generated questions. Various frameworks for prompting the process of reflection were also discussed, including the use of the Socratic method. The Socratic method is a well-known approach to teaching and learning that encourages critical thinking and self-reflection through dialogue. This method was chosen as the basis for the question generation system, as it aligns well with the goal of generating reflection questions.

A question generation system was developed, utilizing the method of Socratic questioning to improve the quality of the generated questions. The system works as a dialogue between two roles, each an LLM: the teacher-educator and the student-teacher. The student-teacher is the one who generates the questions, while the teacher-educator guides the process of improvement of each question. The system was designed to be interactive, allowing for a dialogue between the two roles. OpenAI models were used for the LLMs in the system.

The system was evaluated using primarily qualitative criteria assessed through quantitative scoring via LLM comparison, as the standard quantitative metrics normally used in QG were unsuitable for the current task. The results of the evaluation showed that the system was able to generate high-quality questions that were relevant to the input text and showed potential to prompt reflection in the learner. The Socratic dialogue generally proved to be an effective tool for improving the results compared to generating questions without dialogue. A study with real teachers will also be performed separately in the future.

The system demonstrated its capabilities, but certain challenges and areas for refinement were identified. A key aspect is determining the optimal length and conditions for the Socratic dialogue. While the dynamic approach, where the teacher-educator model determined the stopping point (typically after 2-3 iterations), generally yielded comparable or even superior question quality compared to fixed, longer iteration counts, the ideal number of iterations can vary based on specific quality criteria and input parameters. For instance, the ”depth” of questions sometimes benefited from more iterations, particularly when no external materials were provided. Conversely, it was observed that very long fixed dialogues (e.g., 10 iterations) could occasionally drift from the primary task of question refinement. These findings reveal complex relationships existing between iteration strategy and input parameters.

Future work can focus on refining the dialogue process, particularly in guiding the dialogue length more effectively, and improving the evaluation of the quality of the generated questions. Overall, the system has shown promise in the field of automatic question generation and has the potential to be a valuable tool for educators and learners alike.

Bibliography

1. CLARÀ, Marc. What is reflection? Looking for clarity in an ambiguous notion. *Journal of teacher education*. 2015, vol. 66, no. 3, pp. 261–271. Available from DOI: 10.1177/0022487114552028.
2. DAMEN, Inge C. M. Van Seggelen –; HEZEWIJK, René Van; HELSDINGEN, Anne S.; WOPEREIS, Iwan G. J. H. Reflection: A Socratic approach. *Theory & Psychology*. 2017, vol. 27, no. 6, pp. 793–814. Available from DOI: 10.1177/0959354317736388. PMID: 29249867.
3. BURKE, Hannah K; BUNDY, Anita C; LANE, Shelly J. If reasoning, reflection, and evidence-based practice are essential to practice, we must define them. *The Open Journal of Occupational Therapy*. 2023, vol. 11, no. 1, pp. 1–6. Available from DOI: 10.15453/2168-6408.2044.
4. SMITH, Nathan. *Introduction to Philosophy*. Socrates as a Paradigmatic Historical Philosopher [online]. Houston, Texas: OpenStax, 2022 [visited on 2025-03-30]. Available from: https://openstax.org/books/introduction-philosophy/pages/1-introduction.
5. PLATO. *Apology* [online]. Trans. by JOWETT, Benjamin. Project Gutenberg, 1999 [visited on 2025-03-30]. Available from: https://www.gutenberg.org/ebooks/1656.
6. NGUYEN, Quoc Dinh; FERNANDEZ, Nicolas; KARSENTI, Thierry; CHARLIN, Bernard. What is reflection? A conceptual analysis of major definitions and a proposal of a five-component model. *Medical education*. 2014, vol. 48, no. 12, pp. 1176–1189. Available from DOI: 10.1111/medu.12583.
7. DEWEY, John. *How we think*. D.C. Heath & Company, 1910. ISBN 978-1610279116.
8. SUAIB, Nur Rizky Alfiany. Reflective Practice in Teacher Education: A Look into Past and Present Theories, and Some Pertinent Issues. *Tell : Teaching of English Language and Literature Journal*. 2022, vol. 10, no. 1, pp. 51–62. Available from DOI: 10.30651/tell.v10i1.12137.

9. SCHÖN, Donald A. *The Reflective Practitioner: How Professionals Think in Action*. 1983. Available from DOI: 10.4324/9781315237473. PMID: 1234567890.
10. FIOCK, Holly. Reflection Defined: What Is Reflection? *The Students' Guide to Learning Design and Research* [online]. 2020 [visited on 2025-04-08]. Available from: https://edtechbooks.org/studentguide/reflection.
11. SCHÖN, Donald A. *Educating the Reflective Practitioner: Toward a New Design for Teaching and Learning in the Professions*. Jossey-Bass, 1987. ISBN 1-55542-025-7.
12. MOON, Jennifer A. *Reflection in learning and professional development: Theory and practice*. London: Routledge, 1999. ISBN 978-0-7494-3452-6.
13. MOON, Jenny. Getting the measure of reflection: considering matters of definition and depth. *Journal of Radiotherapy in Practice*. 2007, vol. 6, no. 4, pp. 191–200. Available from DOI: 10.1017/S1460396907006188.
14. SUPHASRI, Ponsawan; CHINOKUL, Sumalee. Reflective practice in teacher education: Issues, challenges, and considerations. *Pasaa*. 2021, vol. 62, no. 1, pp. 236–264. Available from DOI: 10.58837/CHULA.PASAA.62.1.9.
15. SORVATZIOTI, Demetra Fr. The Socratic method of teaching in a multidisciplinary educational setting. *International Journal of Arts & Sciences*. 2012, vol. 5, no. 5, p. 61. ISSN 1944-6934.
16. OVERHOLSER, James C; BEALE, Eleanor. The art and science behind socratic questioning and guided discovery: a research review. *Psychotherapy Research*. 2023, vol. 33, no. 7, p. 946. ISSN 1050-3307. Available from DOI: 10.1080/10503307.2023.2183154.
17. CAREY, Timothy; MULLAN, Richard. What is Socratic questioning? *Psychotherapy: Theory, Research, Practice, Training*. 2004, vol. 41, no. 3, pp. 217–226. Available from DOI: 10.1037/0033-3204.41.3.217.
18. PAUL, Richard; ELDER, Linda. *The art of Socratic questioning* [online]. Dillon Beach, CA: The Foundation for Critical Thinking, 2006 [visited on 2025-03-30]. ISBN 0-944-583-31-8. Available from: https://archive.org/details/the-art-of-socratic-questioning/mode/2up.
19. DALIM, Siti Fairuz; ISHAK, Aina Sakinah; HAMZAH, Lina Mursyidah. Promoting students' critical thinking through Socratic method: The views and challenges. *Asian Journal of University Education* [online]. 2022, vol. 18, no. 4, pp. 1034–1047 [visited on 2025-04-12]. Available from: https://ajue.uitm.edu.my/wp-content/uploads/2022/10/15-Done_F-SITI-FAIRUZ.pdf.

20. BLOOM, Benjamin S; ENGLEHART, Max D; FURST, Edward J; HILL, Walker H; KRATHWOHL, David R. *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: David McKay Company, 1956.

21. ANDERSON, Lorin W; KRATHWOHL, David R; BLOOM, Benjamin S. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* New York: Longman, 2001.

22. KRATHWOHL, David R. A revision of Bloom's taxonomy: An overview. *Theory into practice.* 2002, vol. 41, no. 4, pp. 212–218. Available from DOI: `10.1207/s15430421tip4104_2`.

23. GIBBS, Graham. *Learning by doing: A guide to teaching and learning methods.* Oxford: Oxford Polytechnic, 1988.

24. MCCABE, Gavin; THEJLL-MADSEN, Tobias. *Reflectors Toolkit* [online]. The University of Edinburgh, 2025 [visited on 2025-04-08]. Available from: `https://reflection.ed.ac.uk/reflectors-toolkit/reflecting-on-experience/gibbs-reflective-cycle`.

25. MULLA, Nikahat; GHARPURE, Prachi. Automatic question generation: a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence.* 2023, vol. 12, no. 1, pp. 1–32. ISSN 2192-6360. Available from DOI: `10.1007/s13748-023-00295-9`.

26. LEITE, Bruno; CARDOSO, Humberto. Do Rules Still Rule? Comprehensive Evaluation of a Rule-Based Question Generation System. In: *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU).* 2023, vol. 2, pp. 27–38.

27. HEILMAN, Michael; SMITH, Noah A. Good Question! Statistical Ranking for Question Generation. In: KAPLAN, Ron; BURSTEIN, Jill; HARPER, Mary; PENN, Gerald (eds.). *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* [online]. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 609–617 [visited on 2025-05-01]. Available from: `https://aclanthology.org/N10-1086/`.

28. DHOLE, Kaustubh D.; MANNING, Christopher D. Syn-QG: Syntactic and Shallow Semantic Rules for Question Generation [online]. 2022 [visited on 2025-05-01]. Available from arXiv: `2004.08694 [cs.CL]`.

29. KEKLIK, Onur; TUGLULAR, Tugkan; TEKIR, Selma. Rule-Based Automatic Question Generation Using Semantic Role Labeling. *IEICE Transactions on Information and Systems.* 2019, vol. E102.D, no. 7, pp. 1362–1373. Available from DOI: `10.1587/transinf.2018EDP7199`.

30. QUOC-HUNG, Pham; NGUYEN, Minh-Tien; TRAN-TIEN, Manh; LE, Hung; PHAN, Xuan-Hieu. Learning to Generate Questions by Enhancing Text Generation with Sentence Selection [online]. 2022 [visited on 2025-05-01]. Available from arXiv: 2212.12192 [cs.CL].

31. WANG, Qifan; YANG, Li; QUAN, Xiaojun; FENG, Fuli; LIU, Dongfang; XU, Zenglin; WANG, inong; MA, Hao. Learning to Generate Question by Asking Question: A Primal-Dual Approach with Uncommon Word Generation. In: GOLDBERG, Yoav; KOZAREVA, Zornitsa; ZHANG, Yue (eds.). *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 46–61. Available from DOI: 10.18653/v1/2022.emnlp-main.4.

32. SU, Dan; XU, Yan; DAI, Wenliang; JI, Ziwei; YU, Tiezheng; FUNG, Pascale. Multi-hop Question Generation with Graph Convolutional Network. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020. Available from DOI: 10.18653/v1/2020.findings-emnlp.416.

33. LI, Yunji; LI, Sujian; SHI, Xing. *Findings of EMNLP 2022*. Consecutive Question Generation via Dynamic Multitask Learning [online]. 2022. [visited on 2025-05-01]. Available from DOI: 10.48550/arXiv.2211.08850.

34. LOPEZ, Luis Enrico; CRUZ, Diane Kathryn; CRUZ, Jan Christian Blaise; CHENG, Charibeth. *Simplifying Paragraph-level Question Generation via Transformer Language Models* [online]. 2021. [visited on 2025-05-01]. Available from arXiv: 2005.01107 [cs.CL].

35. DUONG-TRUNG, Nghia; WANG, Xia; KRAVCIK, Milos. BloomLLM: Large Language Models Based Question Generation Combining Supervised Fine-Tuning and Bloom's Taxonomy. In: *Technology Enhanced Learning for Inclusive and Equitable Quality Education (EC-TEL 2024)*. 2024, pp. 93–98. ISBN 978-3-031-72311-7. Available from DOI: 10.1007/978-3-031-72312-4_11.

36. MUCCIACCIA, Sérgio Silva; MEIRELES PAIXÃO, Thiago; WALL MUTZ, Filipe; SANTOS BADUE, Claudine; FERREIRA DE SOUZA, Alberto; OLIVEIRA-SANTOS, Thiago. Automatic Multiple-Choice Question Generation and Evaluation Systems Based on LLM: A Study Case With University Resolutions. In: RAMBOW, Owen; WANNER, Leo; APIDIANAKI, Marianna; AL-KHALIFA, Hend; EUGENIO, Barbara Di; SCHOCKAERT, Steven (eds.). *Proceedings of the 31st International Conference on Computational Linguistics* [online]. Abu Dhabi, UAE: Association for Computational Linguistics, 2025, pp. 2246–2260 [visited on 2025-05-01]. Available from: https://aclanthology.org/2025.coling-main.154/.

37. LU, Chao-Yi; LU, Sin-En. A Survey of Approaches to Automatic Question Generation:from 2019 to Early 2021. In: LEE, Lung-Hao; CHANG, Chia-Hui; CHEN, Kuan-Yu (eds.). *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)* [online]. Taoyuan, Taiwan: The Association for Computational Linguistics and Chinese Language Processing (ACLCLP), 2021, pp. 151–162 [visited on 2025-05-01]. Available from: `https://aclanthology.org/2021.rocling-1.21/`.

38. PAPINENI, Kishore; ROUKOS, Salim; WARD, Todd; ZHU, Wei-Jing. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 311–318. ACL '02. Available from DOI: `10.3115/1073083.1073135`.

39. LAVIE, Alon; AGARWAL, Abhaya. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: *Proceedings of the Second Workshop on Statistical Machine Translation*. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 228–231. StatMT '07.

40. LIN, Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Text Summarization Branches Out* [online]. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81 [visited on 2025-05-01]. Available from: `https://aclanthology.org/W04-1013/`.

41. ZHANG, Tianyi; KISHORE, Varsha; WU, Felix; WEINBERGER, Kilian Q.; ARTZI, Yoav. *BERTScore: Evaluating Text Generation with BERT* [online]. 2020. [visited on 2025-05-02]. Available from DOI: `10.48550/arXiv.1904.09675`.

42. SCARIA, Nicy; DHARANI CHENNA, Suma; SUBRAMANI, Deepak. Automated Educational Question Generation at Different Bloom's Skill Levels Using Large Language Models: Strategies and Evaluation. In: *Artificial Intelligence in Education*. Springer Nature Switzerland, 2024, pp. 165–179. ISBN 9783031642999. ISSN 1611-3349. Available from DOI: `10.1007/978-3-031-64299-9_12`.

43. *AI dětem - Lídr vzdělávání v oblasti umělé inteligence* [online]. AI dětem z.s. [visited on 2025-04-08]. Available from: `https://www.aidetem.cz/`.

Attachment contents

```
/
├── README.md
├── requirements.txt ........................ Jupyter notebook dependencies
├── openai.txt ......................... file containing the OpenAI API key
├── Data .......................... implementation of the generation system
│   ├── free10iter ........................... dynamic number of iterations
│   │   ├── Prompts ............................. prompts for combinations
│   │   └── free10iter.json ........................... generated dialogues
│   └── fixed10iter. ............................ fixed number of iterations
│       ├── Prompts ............................. prompts for combinations
│       └── fixed10iter.json .......................... generated dialogues
├── Evals ..................................... results of the evaluation
│   ├── BestCombination ................ best combination of the parameters
│   ├── BestIter ........................... best number of fixed iterations
│   ├── FreeVSFixed ....... comparison between dynamic and fixed iterations
│   └── NoneVSFree .. comparison between no dialogue and dynamic iterations
├── Materials ............. source materials used for the question generation
└── Thesis ............................... source code of the thesis in LaTeX
```