Engineering Degree Project

Improving the Reliability of Educational AI Chatbots Using Retrieval-Augmented Generation

Author: Khaled Matar, Yousef Mohammad
Supervisor: Ahmed Taiye
Lnu Supervisor: Marcelo Milrad
Semester: Spring 2024
Subject: Computer Science

Abstract

This thesis explores the integration of Retrieval-Augmented Generation (RAG) technology into AI-driven chatbots, aimed at enhancing the quality of educational tools by addressing the issue of "hallucinations" common in Large Language Models (LLMs). The project, named AWAITS (Academic Writing AI Tutoring System), leverages RAG to provide precise, reliable, and contextually relevant educational support to students, particularly in academic writing and critical thinking. By utilizing external databases for real-time information retrieval, AWAITS demonstrates a significant reduction in misinformation, thereby boosting the educational efficacy of chatbots. The research systematically analyzes the improvements in AI-driven educational interactions through comparative studies between RAG-equipped systems and traditional LLMs without RAG capabilities. Findings underscore the potential of RAG systems to transform educational tools by providing substantiated and authoritative content, thereby fostering a more engaging and informative learning environment. This study contributes to the on-going research efforts in this field, providing the benefits of integrating RAG into educational AI applications but also sets a precedent for future research and development in enhancing AI reliability and functionality of AI applications.

Keywords: Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), AI in education, chatbots, educational technology, hallucination problem, academic writing support, critical thinking enhancement, AI reliability, contextual relevance.

Acknowledgments

We, Yousef Mohammad and Khaled Matar, would like to express our sincere gratitude to Dr. Ahmed Taiye Mohammed for his guidance and support throughout the development of our thesis. Dr. Mohammed's expertise and insightful feedback have been invaluable in shaping our research and addressing complex challenges. We are also thankful to Professor Marcelo Milrad, our university supervisor, for his knowledgeable advice and constructive feedback, which have significantly contributed to our analytical approaches and methodologies.

Furthermore, we extend our appreciation to our families and friends for their understanding and encouragement during our academic journey. Their support has provided us with the motivation to overcome obstacles and persevere in achieving our goals.

Additionally, we acknowledge our peers and faculty members from our department, whose valuable insights and suggestions have helped refine our thesis. Their collaborative spirit and scholarly input have enriched our research process. Finally, we express our gratitude to all those who have supported us, directly or indirectly, throughout this endeavor. Their contributions have been instrumental in shaping this academic achievement.

Contents

1 Introduction **1**
1.1 Background 1
1.2 Related Work 2
1.3 Problem Formulation 2
1.4 Motivation 3
1.5 Milestones 4
1.6 Scope/Limitation 4
1.7 Target group 5
1.8 Outline 5

2 Theoretical Perspectives **6**
2.1 Models 6
2.2 Large Language Models (LLM) 6
2.3 Retrieval-Augmented Generation (RAG) 7
2.3.1 Retriever 7
2.3.2 Generator 8

3 Methodological Approach **9**
3.1 Context and Settings 9
3.2 Methods 9
3.3 Reliability and Validity 10
3.3.1 Reliability 10
3.3.2 Validity 10
3.4 Ethical considerations 11
3.5 Team Work 11

4 Implementation **12**
4.1 Testing 16
4.2 Evaluation 16

5 Experimental Setup **17**

6 Results and Analysis **18**

7 Discussion **21**
7.1 RQ 1 21
7.2 RQ 2 21
7.3 Comparative Analysis and Improvements 22

8 Conclusion **25**
8.1 Future work 25

References **27**

A Appendix A **A**

B Appendix B **C**

C Appendix C **D**

D Appendix D E

E Appendix E G

F Appendix F I

G Appendix G I

1 Introduction

"*After the internet and mobile internet triggered the Third Industrial Revolution, artificial intelligence technologies, driven by big data, are fuelling a Fourth Industrial Revolution.*" — Dr. Qiang Yang, Hong Kong University of Science and Technology (qtd. in Chao 2023) [1] "*The great opportunity offered by ICTs comes with a huge intellectual responsibility to understand them and take advantage of them in the right way.*"— Luciano Floridi (2014), The Fourth Revolution: How the Infosphere is Reshaping Human Reality [2]

1.1 Background

Recognizing the challenges students often encounter in navigating extensive course materials or other academic resources to locate specific information is crucial to their academic success [3]. The sheer volume of content available can be overwhelming and time-consuming for students seeking precise information for their assignments or inquiries. This challenge is particularly pronounced among first-year university students, especially those grappling with the nuances of academic writing, which can be even more daunting for non-native English speakers [4]. Consequently, their writing and critical thinking skills may be adversely affected, particularly in subjects requiring effective articulation or communication of their understanding of the subject matter or discourse on a given topic [4]. This issue extends across students in both technical fields like sciences and non-technical fields like humanities and social sciences. It suggests that students may need to navigate between multiple languages, including second or even third-party languages. [5]

The advent of Artificial Intelligence (AI) represents a significant turning point, altering the dynamics of writing practices, the content produced, and the frameworks within which writing and education occur [6]. Large Language Models (LLMs), characterized by their extensive databases amassed during pre-training and the capability of Deep Neural Networks (DNNs) to manage numerous parameters, signify a pivotal moment in the history of writing [6]. This current AI era is comparable in significance to past revolutionary milestones in the realm of writing [7]. New tools and technologies that rely on AI can help create personalized learning and can make education more interactive and effective [7]. However, using AI to help students improve their critical thinking and writing skills comes with its own set of challenges, including making sure it is used in a way that is ethically fair and does not cheat the system [8]. To address some of these challenges, a proof of concept was developed by the Knowledge Environment Digital Transformations (KE-DT) in collaboration with researchers, students, and teachers at Linnaeus University (LNU), known as CHAT4-ISP (the research paper on this can be viewed under Appendix G). CHAT4-ISP is a chatbot designed to provide well-structured feedback and suggestions on submitted text articles, focusing on the coherence and grammatical structure of exercises or assignments [9]. Like other chatbots built on LLMs, the CHAT4-ISP system also suffers from a common issue known as hallucinations, which occurs when the system generates inaccurate responses that do not align with the input, intended context, or subject matter [10].

In addition to addressing the challenge of hallucinations associated with some of the outcomes produced by CHAT4-ISP, it is crucial to overcome its susceptibility to overfitting, wherein the model may memorize specific training data patterns, resulting in suboptimal performance on unseen data [11]. To achieve this, an improved method is required to mitigate hallucination effects and guide the system in extracting relevant information from designated sources [12, 13, 14]. This enhancement is crucial as it enables LLMs to better capture contextual dependencies and long-range relationships within text data, thereby enhancing their performance across various Natural Language Processing (NLP) tasks [15, 16].

1.2 Related Work

The concept of information retrieval has long been established in the field of computer science. As early as 2011, Ricardo Baeza-Yates co-authored a seminal work, "*Modern Information Retrieval*" [17], which detailed methods for extracting information from external sources. This foundational text has significantly influenced subsequent developments in information retrieval systems, setting a standard for techniques that are crucial in the architecture of contemporary AI-driven conversational systems.

A significant challenge in the development of conversational AI systems is mitigating the issue of "*hallucination*", where chatbots generate plausible but factually incorrect responses (fabricating non-existent facts). Recent advancements in dialogue models have significantly enhanced their conversational abilities; however, these models still frequently suffer from inaccuracies and knowledge hallucinations, which can mislead users or degrade the user experience [18].

In addressing these concerns, Shuster et al. explore the use of neural-retrieval-in-the-loop architectures for knowledge-grounded dialogue. Their work is particularly relevant for enhancing the factual accuracy of responses generated by state-of-the-art chatbots. They employ RAG techniques, which integrate retrievers, rankers, and encoder-decoders to enhance the chatbot's knowledge base in real-time during dialogue. This integration enables the chatbot to access accurate information pertinent to the conversation, thereby reducing the incidence of hallucinations [19].

Their findings demonstrate that RAG systems not only retain the conversational fluency of large language models but also significantly reduce knowledge hallucination by leveraging dynamically retrieved documents to inform responses. Through both automatic and human evaluations, the models utilizing RAG architecture showed a marked improvement in generating factually correct responses, reducing hallucinations substantially compared to traditional models that do not use retrieval mechanisms [19].

This approach aligns with our current efforts into enhancing chatbot reliability by reducing hallucinations, as it provides a promising method to ensure that responses not only appear coherent but are also underpinned by verifiable facts.

Several studies explore the use of AI for writing support in higher education [9, 20]. Bond and colleagues highlight the need for ethical and rigorous research in AI for education [20], which aligns with our focus on responsible development and use of AWAITS.

In terms of ethical considerations and practical implications, Taiye et al [9]. expands and present CHAT4-ISP AI, a Generative AI tool for academic writing, demonstrating the potential of this technology in our domain. With regard to the project, it will build on this prior research.

1.3 Problem Formulation

As outlined in the background, LLMs are prone to hallucinations which often lead to confusion or misinformation for users, rendering generated output not entirely accurate. This phenomenon is not considered a "bug" in LLMs but rather a consequence of their operating nature [15, 20]. Upon reviewing existing literature, one potential solution to address hallucination issues is to increase the size of context windows. An example of this approach can be observed in the operational architecture of Google Gemini 1.5, which utilizes a million token context window [15, 21]. In comparison, the context window for GPT4 is 128k tokens [11, 15], this indicates that models with larger context windows are capable of processing vast amounts of information in a single iteration, eliminating the need to sift through documents to retrieve relevant information.

While larger context windows can enhance a model's accuracy by providing more immediate data. Large context windows bring challenges related to increased costs, computational demands, and slower response times. These windows also require significant hardware resources, affecting its scalability [21].

In contrast, RAG addresses these issues differently. RAG models augment their responses by pulling or retrieving information from external databases as needed during the generation process [12, 13, 14, 16], serving as a "second memory" where it can access relevant sentences from the content. This capability not only enhances the accuracy and relevance of responses [11] but also allows for detailed references such as specific page numbers and book sections [16]. This makes RAG particularly useful in educational settings, where guiding students to the exact source material can significantly aid learning.

Regarding the use of AI in education (AIEd), the work carried out in this thesis has been done in close collaboration with KE-DT and other LNU's stakeholders from the social sciences and humanities. Our main goal has been to explore and improve the impact of a chatbot that is built on top of CHAT4-ISP equipped with RAG capabilities. This new improved version of the system will be called AWAITS (Academic Writing AI Tutoring System)

To address the hallucination issue mentioned earlier, the AWAITS system will rely on the use of RAG technologies. By using external databases such as documents, books, or databases, RAG systems can significantly enhance the accuracy and dependability of the generated content [16]. This integration of external data not only boosts precision but also allows for the inclusion of specialized domain knowledge and ensures continuous updates to the system's knowledge base. Consequently, RAG systems emerge as a promising solution, particularly for tasks demanding advanced expertise [16]. To achieve this, in our work we have developed an enhanced version of the CHAT4-ISP chatbot into a version that utilizes a RAG system.

To achieve this, the frontend and backend components of the project have underwent an upgrade to improve the functionality. Specifically, the backend has been modified to incorporate support for RAG using the built-in features of OpenAI's API, enhancing the chatbot's responses with more precise, and referenced information.

Furthermore, RAG ensures that the advice and feedback provided to students are grounded in factual and authoritative sources, minimizing the risk of inaccuracies or misinformation. By referencing specific pages or sections from course materials, the chatbot can guide students directly to the relevant content, fostering a deeper understanding of the subject matter and promoting effective learning [16].

Based on the problems dicussed and presented in the previous sections, we have formulated the following research questions as described below:

1. *Does the implementation or use of RAG help reduce hallucination problems?*
2. *Does integrating RAG in a chatbot affect the quality of advice?*

1.4 Motivation

This study is motivated by the recognition of the challenges students face in navigating extensive course materials to find specific information, particularly affecting first-year university students and non-native English speakers [22]. The emergence of AI offers promising opportunities for personalized learning [3], but integrating AI into education presents challenges such as ensuring fairness and avoiding cheating [8]. The CHAT4-ISP chatbot, developed at LNU [9], suffers from hallucination issues common in chatbots built on LLMs [23]. Overcoming these challenges requires addressing hallucination effects and susceptibility to overfitting, ultimately aiming to enhance AI-powered educational tools [24] effectiveness in supporting students' critical thinking and writing skills.

1.5 Milestones

Project Milestones	
Date	Milestone/Task
2024-04-04	Draft Project Plan Peer Review
2024-04-10	Report work, Abstract, Introduction, Problem Formulation
2024-04-10	Initialised Work Environment (clone the repo, install dependencies, read the code)
2024-04-11	Completed Project Plan Submission
2024-04-24	Completed Half of the Report
2024-04-26	Seminar
2024-04-30	Implement RAG on AWAITS
2024-05-07	Thorough Testing of the Application / Bug fixes
2024-05-20	Finish sections on the Report, Discussion, Method, Team work
2024-05-23	Report Submission

Table 1.1: *Timeline of project milestones for the AWAITS development.*

1.6 Scope/Limitation

The work performed as part of this thesis has used the RAG feature from OpenAI's Assistant API. Due to time constraints, other models, or RAG frameworks like LangChain have not been explored. The project scope does not include a comparison of all available AI-powered writing assistance tools, but rather focuses on the specific comparison between RAG-based and non-RAG-based systems.

Numerous methods exist for assessing the accuracy of RAG responses. These include metrics such as Precision@k, Recall@k, F1@k, Mean Reciprocal Rank (MRR), Average Precision (AP), and Discounted Cumulative Gain (DCG@k) [25]. While these metrics are crucial for a deeper understanding of RAG performance, due to time constraints, this thesis will not go into a deeper detailed discussions of these topics.

This project employs AI technologies, such as ChatGPT [11] and Claude 3 [26], to enhance the composition and structure of sentences and sections. This is done by providing initial drafts written by the authors and then asking for suggestions and modifications to improve clarity, grammatical accuracy, and coherence. This method is particularly beneficial for non-native English speakers and individuals less experienced in technical writing, a theme that is further explored in an upcoming section.

Nonetheless, it is essential to recognize the limitations of AI assistance and human oversight remains crucial to preserve the authenticity of original ideas and ensure contextual accuracy. This theme of "*human in loop*" will also be further explored.

1.7 Target group

The primary target audience for the AWAITS system is students and non-native English speakers as they struggle the most within writing, the project aims to investigate the effectiveness of RAG systems among this target group [27]. By focusing on students, the AWAITS project aims to provide insights and tools that can directly benefit students in their academic journeys [28]. The project's findings and the AWAITS chatbot itself may also be of interest to educational institutions, instructors, and policymakers who are interested in exploring the potential of AI-powered tools to enhance learning outcomes [28]. These stakeholders may find the insights generated by the AWAITS project useful in informing their decisions about the integration of AI technologies in educational settings [28]. Additionally, the AWAITS project may hold relevance for the broader research community studying the intersection of AI and education [29]. The project's contributions to the understanding of RAG systems and their impact can inform future research and development in this field, ultimately benefiting a wider range of stakeholders beyond the immediate target audience of students.

1.8 Outline

The content of this thesis is organized into sections. In *Section* 1, Introduction, it provides background information on AI's role in technological advancements and its significance in education and writing. This section also explores the motivation and necessity for advanced tools like RAG systems in enhancing critical thinking and writing skills [27]. The emphasis is placed on problem formulation, which explicitly outlines research questions and objectives, focusing on comparing chatbot systems with and without RAG capabilities, while also establishing the study's scope and limitations. In the Related Work section, existing studies on AI in education are synthesized, and relevant theories supporting AI integration in educational environments are explored [9, 19, 20]. *Section* 2 is for exploring different important theoretical persepectives, such as models, LLMs and RAG. *Section* 3, Methodological Aprroach, details research design, testing procedures, and teamwork distribution. *Section* 4, Implementation, introduces technical specifications, development processes, and tools and technologies utilized. *Section* 5, Experimental Setup, Results and Analysis, elaborates on the configuration of chatbot systems and the testing environment provides interpretations of results. *Section* 7, Discussion, provides a discussion that connects our results with the research questions of this study, while *Section* 8, Conclusion, discusses our findings, analyzes societal relevance, and identifies areas for potential improvement.

Building on the foundational understanding presented in the introduction, the subsequent section explores the theoretical frameworks that underpin the research methodologies and models.

2 Theoretical Perspectives

The theoretical perspectives section introduces concepts that inform this study and the reader, by providing a foundational understanding of the computational methods used. This overview sets the stage for further exploration of methodologies and applications in subsequent sections, focusing on their relevance and impact without going deeply into technical details.

2.1 Models

In the field of machine learning and artificial intelligence, a model refers to a mathematical or computational representation of a system or process. Models are designed to capture patterns and relationships within data, allowing them to make predictions, classifications, or generate outputs based on new inputs [30].

Models are typically trained on large datasets, using various algorithms and techniques, such as neural networks, decision trees, or support vector machines. The training process involves adjusting the model's parameters to minimize the inconsistency between its predictions and the actual, desired outputs. This often involves techniques from iterative and incremental development in software engineering, which have been adapted to optimize learning processes in AI systems [31].

Neural networks, in particular, are a class of models inspired by the structure and function of biological neural networks in the human brain. They are composed of interconnected nodes (neurons) that process and transmit information, enabling them to learn and make predictions from data. The architecture and functioning of neural networks have been significantly influenced by advancements in understanding how individual units in a network contribute to its overall performance [30, 32].

Furthermore, the application of neural networks in educational settings has demonstrated their capability to enhance personalized learning experiences, showcasing their adaptability across various domains [29]. Neural network models not only mimic cognitive functions but also extend to applications where they facilitate real-time data processing and decision-making [33, 34].

2.2 Large Language Models (LLM)

Large Language Models are a specialized type of AI model that are based on deep neural networks and trained on vast amounts of text data. They are designed to perform a wide range of natural language processing tasks, such as text generation, translation, and question answering [10].

Examples of LLMs include GPT-4 (Generative Pre-trained Transformer 4) developed by OpenAI [11], LLaMA developed by Meta [10, 33] and Claude 3 developed by Anthropic [26]

LLMs are characterized by their ability to capture and model the statistical patterns and relationships present in the training data, allowing them to generate human-like text and understand complex language structures. The "*large*" in LLMs refers to the massive number of parameters (often in the billions) that these models possess in its neural network, enabling them to encode and process an extensive amount of linguistic information [10].

One of the key advantages of LLMs is their capability to transfer learning across various tasks and domains. By pre-training on a diverse corpus of text data, LLMs can acquire general language understanding and knowledge, which can then be fine-tuned or adapted to specific tasks or domains with relatively small amounts of task-specific data [10].

It is important to note that LLMs lack the ability to access real time data or external data that it is not trained on, Figure 2.1 shows a simple example of a response from LLM that requires real time data, which it is unable to answer.

2.3 Retrieval-Augmented Generation (RAG)

RAG is a technique introduced in LLMs to enhance the model's ability to generate accurate and relevant responses by incorporating external knowledge sources during the generation process [16]. RAG systems combine the strengths of pre-trained language models with information retrieval systems, allowing them to access and utilize relevant information from external sources (e.g., databases, documents, or the internet) when generating text [16]. See 2.2

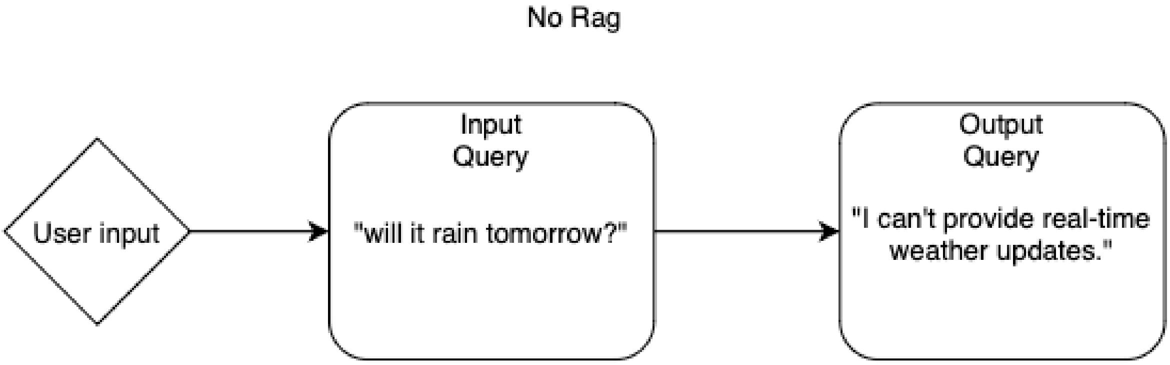


Figure 2.1: *Flowchart depicting a simple user query about weather prediction and the system's response, highlighting a direct decision process without additional actions or data retrieval.*

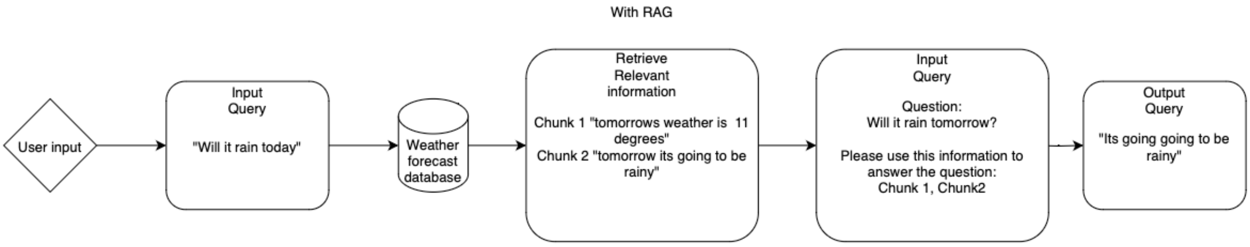


Figure 2.2: *Similar to Figure 2.1, this diagram illustrates the same query regarding weather prediction but employs a RAG approach to fetch data from an external database.*

The RAG approach typically involves two main components [13, 14, 16]:

2.3.1 Retriever

This component is responsible for retrieving relevant information from external sources based on the input query or context. It typically uses information retrieval techniques, such as term-based or semantic search, to identify and retrieve the most relevant documents or passages [13, 16]. This can be seen in figure 2.2 under "Retrieve Relevant Information" component.

2.3.2 Generator

The generator component is the language model itself, which is trained to generate text based on the input context and the retrieved information from the retriever component. The generator can incorporate the relevant information into its output, providing more accurate and knowledge-grounded responses [14, 16]. This can be seen in figure 2.2 under "Input Query" component

RAG systems aim to reduce the issue of hallucinations, where language models generate output that is inconsistent with factual information or lacks grounding in external knowledge. By leveraging external sources, RAG systems can generate responses that are more factually accurate and better aligned with the available knowledge [16].

In the context of the AWAITS project (ie CHAT4-ISP equipped with RAG as described previously), integrating RAG capabilities into the chatbot is expected to enhance the system's ability to provide more accurate and relevant advice and feedback to students, by incorporating information from external sources such as course materials, reference books, or other relevant documents [7, 12, 13, 14]. A more detailed explanation about the workflow of AWAITS using RAG systems will be presented in coming section.

With an overview of the theoretical underpinnings that guided our work, this section transitions into the methodological approaches that will empirically test these theories within the study framework.

3 Methodological Approach

This section describes the method used in the research, and how the results were checked for accuracy. The ethical guidelines followed are also discussed. This research evaluates the effectiveness of RAG technology in chatbots through a controlled experimental approach [35].

Controlled experiments are a fundamental research method used to investigate cause-and-effect relationships between variables. In these experiments, researchers manipulate an independent variable while holding all other factors constant, allowing them to observe its impact on a dependent variable [35, 36].

3.1 Context and Settings

In the context of evaluating RAG technology in chatbots, a controlled experiment enables a direct comparison between systems with and without RAG. By posing identical questions to both versions of the chatbot under controlled conditions, researchers can isolate the impact of RAG technology on the chatbot's performance. This method allows for inferences about the relationship between RAG implementation and any observed improvements in the chatbot's responses, while also providing results that can be analyzed.

Both versions, CHAT4-ISP and AWAITS—
one equipped with RAG AWAITS and the other CHAT4-ISP without—has been assessed and evaluated in a controlled question-and-answer format. In this phase, identical questions have been posed to both versions of the chatbot to compare their responses, specifically focusing on their ability to reference relevant information accurately.

The rationale for this approach is to establish the technical functionality of RAG.

3.2 Methods

The experimental component of the study utilizes a controlled environment where the primary variable—RAG technology presence—is manipulated, also the type of questions posed to both systems is also a variable that will be changes. The questions posed to test the systems were design to be in this three different formats that are described below:

Targeted Chapter Queries: These questions were designed to test the system's ability to retrieve and summarize specific information from designated chapters within the provided books. For example: "*What does chapter 5 on Introduction to Machine Learning with Python cover?*". These questions effectively test the system's precision in retrieving specific information, demonstrating its ability to navigate and extract relevant content from structured sources.

General Knowledge Questions: These broader questions assessed the system's ability to synthesize information across multiple sources and chapters. For instance: "*Compare and contrast the advantages and disadvantages of renewable energy sources.*". These questions evaluate the system's ability to provide broad overviews and demonstrate comprehensive understanding of the material as a whole, testing its capacity to summarize large volumes of information and present key concepts cohesively.

Out-of-Scope Questions: To test the system's ability to recognize its knowledge boundaries, questions were included that were not covered in the provided materials, such as: "*Explain the concept of quantum computing.*". Including these questions evaluates the system's ability to recognize and communicate its knowledge limitations, an essential trait for maintaining user trust and preventing misinformation.

This diverse set of questions allows for an assessment of the system's performance, showcasing its strengths in information retrieval, as well as its ability to handle queries beyond its knowledge base.

In the context of this study, hallucination is defined as the generation of information by the AI system that is either:

1. Factually incorrect when compared to the provided source materials.
2. No references to the provided source materials.

By using these criteria, we can effectively identify hallucinations, or the generation of incorrect or misleading information, which has been recognized as a significant challenge impacting the reliability and educational efficacy of the chatbot [15]. Ensuring that the system's responses are both accurate and properly sourced. This approach is crucial for evaluating the effectiveness of the RAG technology in reducing hallucinations and improving the overall reliability of AI-generated responses in educational contexts.

3.3 Reliability and Validity

To ensure the reliability and validity of the evaluation, two techniques and measures were employed throughout the testing process of the AWAITS chatbot.

Firstly, a Q&A format expirement, where a set of questions will be posed to both systems, and evaluated manually by checking the responses. Due to time constraints, the study focused on exploring six questions to test the system. These questions are related to the three selected books used for the testing. The second method will be a consistency test, where the same questions will be asked multiple times to see if the responses are consistent with themselves.

3.3.1 Reliability

The reliability of the AWAITS chatbot's responses was measured by creating multiple test runs with identical inputs (questions) and reference materials. This approach allowed for the verification of the consistency of the chatbot's outputs when presented with the same set of conditions across different runs [37]. By repeating the tests several times, the potential impact of any random variations or fluctuations in the model's performance was aimed to be reduced. Furthermore, to reduce the inherent randomness of the language model's generation process, low-temperature sampling techniques were utilized, specifically top-k sampling [38]. This technique restricts the model's output distribution to the k most probable tokens, thereby minimizing the likelihood of generating random or irrelevant responses.

3.3.2 Validity

Manual inspection and double-checking of the information provided in the chatbot's outputs against the actual content present in the reference materials accessible to the RAG system were conducted. This step is important to confirm that the chatbot's responses are accurate and appropriately sourced from the provided reference documents.

3.4 Ethical considerations

AI-powered education tools require careful consideration of ethical issues such as accuracy, plagiarism, privacy, societal fairness and bias, [8, 39]. So it is crucial to implement techniques and test AWAITS to prevent unfair biases. Accuracy and plagiarism are essential, as the system should promote original thought and expression. Privacy and data protection are crucial, as AWAITS must ensure secure handling of student data and comply with relevant regulations [39]. Currently, no data collection is available on the testing of the RAG system.

3.5 Team Work

Student Name	Responsibilities
Khaled	Problem Formulation, Milestones, Target Group, Motivation, Theoretical perspectives, Experimental Setup, Result and Analysis, Discussion
Yousef	Background, Related Work, Scope/Limitation, Outline, Methodolological Approach, Implementation, Conclusion

Table 3.2: *Distribution of Responsibilities*

Both students will be actively writing and correcting all sections, but the responsibility of the sections will be divided equally between the two students.

One of the authors of this paper is also involved in developing the CHAT4-ISP system and has co-authored the related paper, "*Generative AI-Enhanced Academic Writing: A Stakeholder-Centric Approach for the Design and Development of CHAT4ISP-AI*" [9].

4 Implementation

The implementation was designed with a core principle in mind: the human element must not be overshadowed by automation. This approach is crucial in maintaining oversight and ethical integrity in AI applications. The figure 4.3 illustrates three distinct scenarios of AI collaboration, each representing different degrees of human involvement—from direct human control to full automation. These scenarios help us explore the balance between human intuition and AI efficiency, ensuring that our technology augments rather than replaces human decision-making.

The figure 4.3 has been insipired by the article "*Who Determines What Is Relevant? Humans or AI? Why Not Both?* [40]. Incorporating a "*human in the loop*" approach ensures that automated judgments are balanced with human oversight, enhancing both the accuracy and trustworthiness of decisions. This method leverages the strengths of both humans and AI, allowing for refined judgments in complex scenarios where machine-only approaches might fall short, thus fostering a more reliable and nuanced understanding of relevance. More about this is explained in the upcoming section.

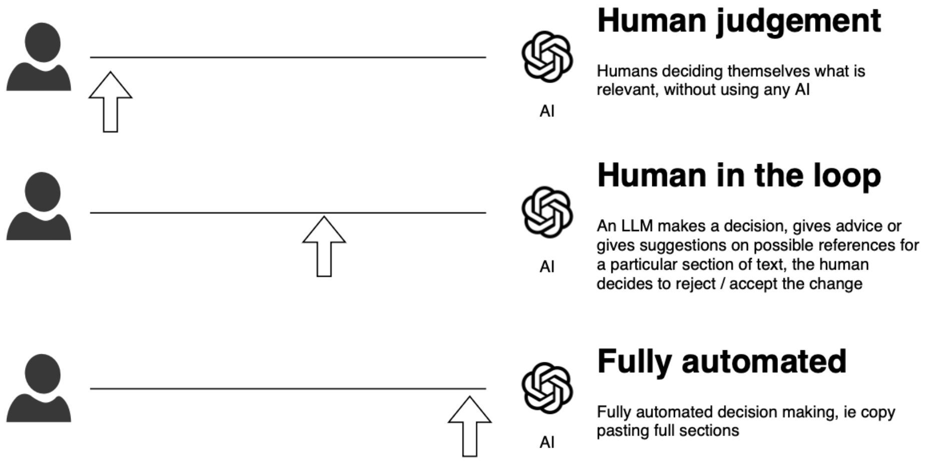

Human judgement
Humans deciding themselves what is relevant, without using any AI
AI
Human in the loop
An LLM makes a decision, gives advice or gives suggestions on possible references for a particular section of text, the human decides to reject / accept the change
AI
Fully automated
Fully automated decision making, ie copy pasting full sections
AI

Figure 4.3: *Collaboration perspecive: Possibilities for collaborative Human and AI to make decisions* [40]

Human judgement: In the Human Judgment scenario, decision-making processes are entirely in the hands of human operators. This traditional approach relies on human expertise, intuition, and direct interaction, where all decisions and critical thinking are manually conducted without the support of AI tools. Such a scenario is prevalent in contexts where precision and personal experience are paramount [40].

Human in the loop: The Human in the Loop scenario describes a collaborative environment where AI systems rapidly generate preliminary content, but human oversight remains crucial. Humans review, refine, and approve AI-generated outputs, ensuring that the final decisions reflect both AI efficiency and human judgment. This balanced approach is ideal for maintaining control over AI's capabilities while leveraging its efficiency for repetitive or data-intensive tasks [40].

Fully automated: In the Fully Automated scenario, AI systems operate independently, making decisions without human input. Leveraging advanced algorithms, these systems analyze vast datasets to produce outcomes at speeds and accuracy unattainable by human capabilities alone. While promising, this scenario raises questions about the need for ethical frameworks and controls to prevent misuse and ensure AI decisions align with human values [40].

The aim is to prioritize the "*Human in the Loop*" approach. This model fosters an optimal balance where AI's rapid data processing capabilities are complemented by human oversight, ensuring both the accuracy and ethical integrity of decisions. By maintaining this balance, we leverage the strengths of AI while preserving the insights and judgments only humans can provide. This strategy not only mitigates risks associated with full automation but also aligns with our commitment to ethical AI implementation, ensuring that our technological advancements enhance rather than replace human expertise.

Moving from the theoretical frameworks of AI collaboration to practical applications, the next figure and discussion illustrate how these concepts are realized in our current systems. Figure 4.4, specifically, showcases the architecture of CHAT4-ISP, which operates without the RAG capabilities, compared to the AWAITS system in figure 4.5 that illustrates the integration of RAG . This comparison highlights the technological advancements and the implementation of our preferred "*Human in the Loop*" model in real-world settings.

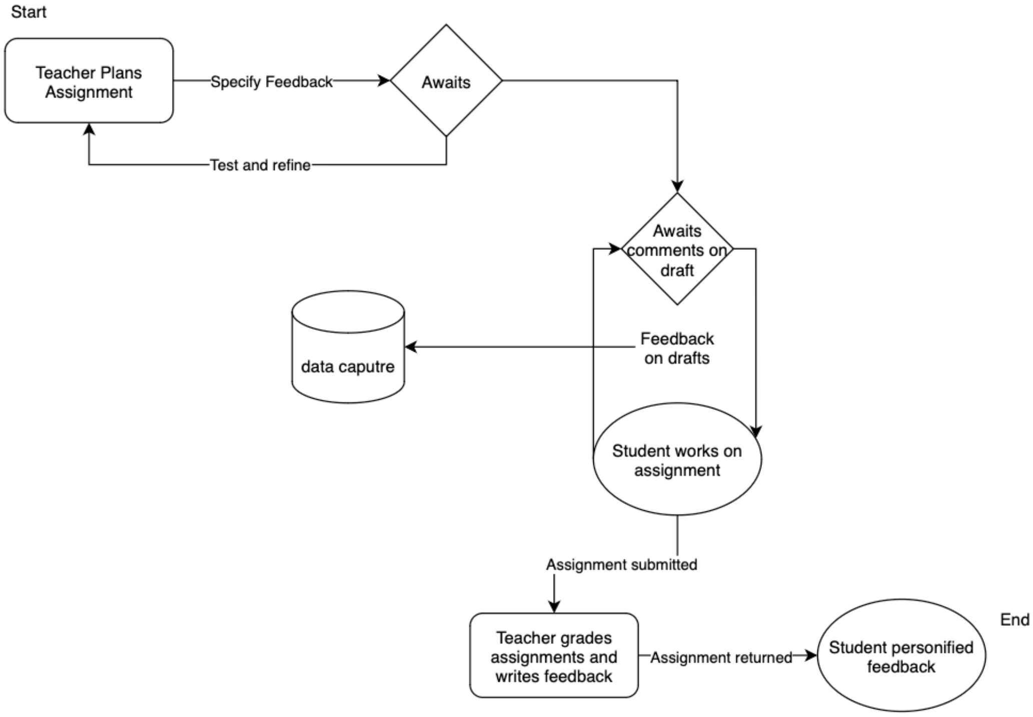


Figure 4.4: *illustrates an educational workflow of CHAT4-ISP with no RAG capabilities.*

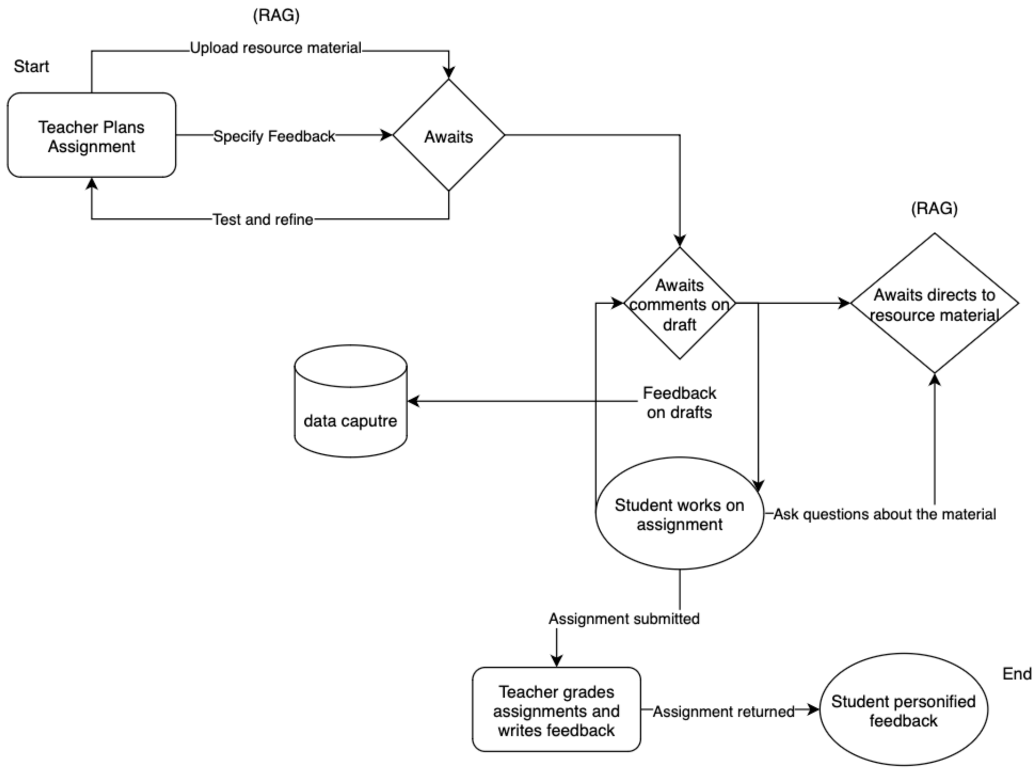


Figure 4.5: *Illustrates the educational information workflow using RAG capabilities.*

Figure 4.5 describes the overall workflow architecture for AWAITS. It starts with the teacher uploading resource material and specifying a pre prompt to guide the model. The assignment is then made available to students, who can work on it while having the ability to ask questions and receive guidance directly related to the provided resources. As students engage with the assignment, they can submit drafts and receive feedback, creating an iterative process that enhances learning. This feedback loop is critical and is directly supported by resource materials, which are accessed as needed. Once the final assignment is submitted, the teacher grades it and provides written feedback, concluding the workflow. The architecture emphasizes continuous interaction and adjustment, facilitated by resource access and feedback, aiming to optimize educational outcomes.

Figure 4.6 showcases the architectural design behind AWAITS. The Frontend allows users to interact through a chatbot (CHAT4-ISP) and the feedback bubbles (AWAITS UI), and supports user authentication via Firebase, and admin actions such as changing prompts or managing files. The Backend handles chat completions and advanced feedback by utilizing various AI models (OpenAI, Claude3 etc), that interact with a database (RAG DB) for retrieving data. Configuration of the GPT models is also handled in the backend, supporting different languages and functionalities.

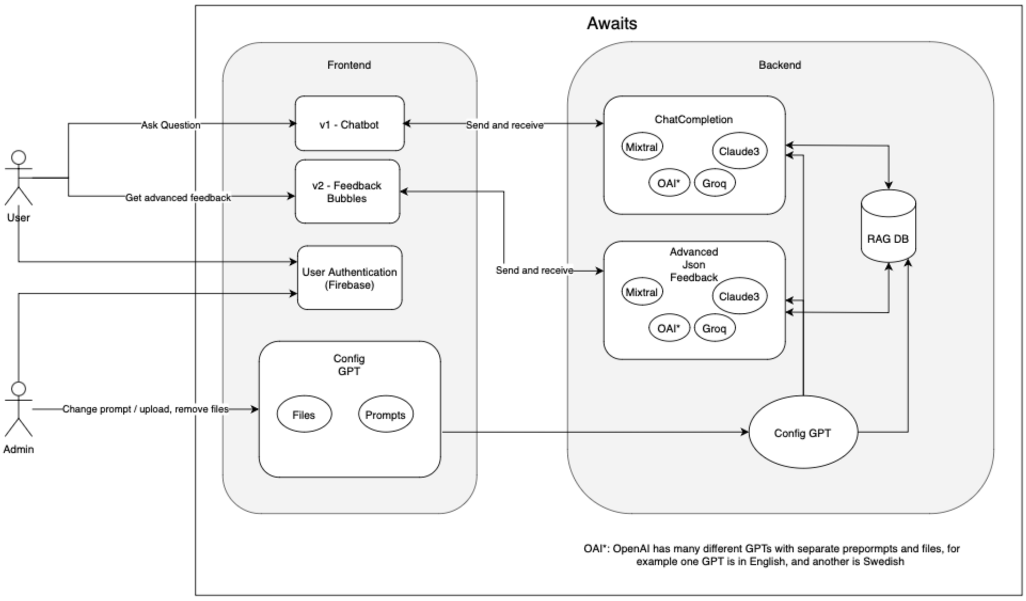
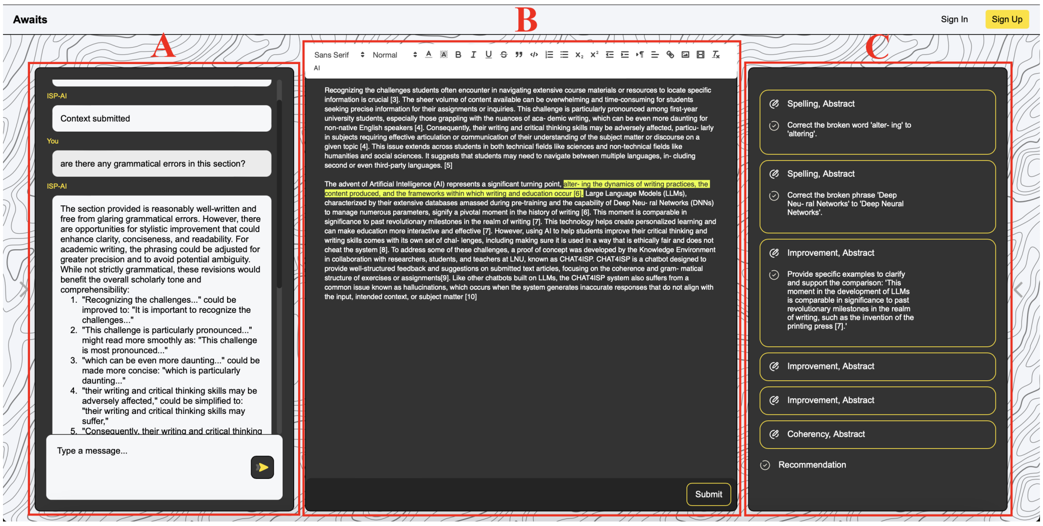


Figure 4.6: *AWAITS architecture*

Figure 4.7 Below is the AWAITS UI interface.



Context submitted

are there any grammatical errors in this section?

The section provided is reasonably well-written and free from glaring grammatical errors. However, there are opportunities for stylistic improvement that could enhance clarity, conciseness, and readability. For academic writing, the phrasing could be adjusted for greater precision and to avoid potential ambiguity. While not strictly grammatical, these revisions would benefit the overall scholarly tone and comprehensibility:

1. "Recognizing the challenges..." could be improved to: "It is important to recognize the challenges..."
2. "This challenge is particularly pronounced..." might read more smoothly as: "This challenge is most pronounced..."
3. "which can be even more daunting..." could be made more concise: "which is particularly daunting..."
4. "their writing and critical thinking skills may be adversely affected," could be simplified to: "their writing and critical thinking skills may suffer,"
5. "Consequently, their writing and critical thinking

Recognizing the challenges students often encounter in navigating extensive course materials or resources to locate specific information is crucial [3]. The sheer volume of content available can be overwhelming and time-consuming for students seeking precise information for their assignments or inquiries. This challenge is particularly pronounced among first-year university students, especially those grappling with the nuances of aca- demic writing, which can be even more daunting for non-native English speakers [4]. Consequently, their writing and critical thinking skills may be adversely affected, particu- larly in subjects requiring effective articulation or communication of their understanding of the subject matter or discourse on a given topic [4]. This issue extends across students in both technical fields like sciences and non-technical fields like humanities and social sciences. It suggests that students may need to navigate between multiple languages, in- cluding second or even third-party languages. [5]

The advent of Artificial Intelligence (AI) represents a significant turning point, alter- ing the dynamics of writing practices, the content produced, and the frameworks within which writing and education occur [6]. Large Language Models (LLMs), characterized by their extensive databases amassed during pre-training and the capability of Deep Neu- ral Networks (DNNs) to manage numerous parameters, signify a pivotal moment in the history of writing [6]. This moment is comparable in significance to past revolutionary milestones in the realm of writing [7]. This technology helps create personalized learning and can make education more interactive and effective [7]. However, using AI to help students improve their critical thinking and writing skills comes with its own set of chal- lenges, including making sure it is used in a way that is ethically fair and does not cheat the system [8]. To address some of these challenges, a proof of concept was developed by the Knowledge Environment in collaboration with researchers, students, and teachers at LNU, known as CHAT4ISP. CHAT4ISP is a chatbot designed to provide well-structured feedback and suggestions on submitted text articles, focusing on the coherence and gram- matical structure of exercises or assignments[9]. Like other chatbots built on LLMs, the CHAT4ISP system also suffers from a common issue known as hallucinations, which occurs when the system generates inaccurate responses that do not align with the input, intended context, or subject matter [10]

Spelling, Abstract

✓ Correct the broken word 'alter- ing' to 'altering'.

Spelling, Abstract

✓ Correct the broken phrase 'Deep Neu- ral Networks' to 'Deep Neural Networks'.

Improvement, Abstract

✓ Provide specific examples to clarify and support the comparison: 'This moment in the development of LLMs is comparable in significance to past revolutionary milestones in the realm of writing, such as the invention of the printing press [7].'

Improvement, Abstract

Improvement, Abstract

Coherency, Abstract

✓ Recommendation

Figure 4.7: *AWAIST UI*

Note that the CHAT4-ISP UI can be found under Appendix F

Figure 4.7 showcases the frontend design of our website application, it consists of three main components denoted as **A**, **B** and **C**, a chatbot on the left, space to paste/write in the middle, and the feedback bubbles on the right.

Component A is a chatbot similar to the CHAT4-ISP chatbot, shown in Appendix F, but in this case it uses the RAG implementation. **Component B** is a simple text editor. **Component C** holds our feedback bubbles that gives advice on the text written (or pasted in) **component B**. Note that these components are connected with eachother, for example when talking with the chatbot (**Component A**), it will have the context of the provided course material, and the text in **component B**, the feedback bubbles in **component C** and generated when the text from the text editor is submitted.

Additional screenshots of various route UIs, such as the Config page and the Login page can be found under Appendix C. There is also a video demo that can be found here (link not added yet)

Following our implementation, the thesis proceeds to evaluate the setup through the experimental testing, the results and analysis of which are essential for validating the hypotheses.

4.1 Testing

In this iteration of the iterative and incremental development method, the testing of the AWAITS chatbot was constrained by time and deadlines.

4.2 Evaluation

To address and verify the resolution of hallucination issues in the system, a comparative QA test was conducted where the outputs were compared side by side in a table. The same question was presented to both the CHAT4-ISP system and the AWAITS system equipped with RAG. The outputs from both systems were then compared to evaluate their performance and accuracy.

5 Experimental Setup

The objective of overall experiment is evaluate the effectiveness of RAG on retrieving and referencing external material to enhance reliability and response accuracy. The focus is particularly on assessing whether the integration of RAG can reduce the occurrence of hallucinations

To evaluate this, the experiment will be conducted using the chatbot denoted as **component A** from Figure 4.7. This dual-setup allows for a direct comparison to assess the effectiveness of RAG in minimizing hallucinations.

The comparison will involve a series of controlled tests where both versions of the chatbot respond to identical queries designed to challenge their informational retrieval and processing capabilities.

For the experiment, the chatbot will be equipped with three distinct books, each serving as an integral part of the chatbot's external knowledge base. These books have been selected to cover a wide array of topics. These books were not chosen for any specific reason, other than them being different from each other.

Here are the books included in the chatbot's database:

1. Andreas, *Introduction To Machine Learning With Python*. Oreilly Associates Inc, 2016.
2. A. V. Da Rosa, *Fundamentals of renewable energy processes*. Academic, 2012.
3. Andrew Emory Dessler, *Introduction to modern climate change*. Cambridge Etc.: Cambridge University Press, 2018.

The chatbot is equipped with three books covering diverse subjects—machine learning, renewable energy, and climate change. This selection provides a broad base of information, ideal for rigorous testing across various academic disciplines.

Two types of queries were designed for the experiment:

1. **Targeted Chapter Queries:** These queries are specifically crafted to evaluate the chatbot's ability to pinpoint and summarize content accurately from designated chapters within the books. This type of query tests the chatbot's precision and depth of understanding in retrieving specific information.
2. **General Knowledge Questions:** This encompasses broader questions about the themes and topics covered in the books. This set assesses the chatbot's ability to generate coherent and contextually appropriate responses based on a wider scope of knowledge.

In addition to evaluating the accuracy of the RAG system, the consistency of the chatbot's responses will also be assessed. By repeatedly asking the same question, the aim is to observe whether the responses remain consistent or if there are significant variations. This test will help determine the reliability of the chatbot when faced with repeated queries under similar conditions. The consistency of responses is crucial for ensuring that the system can be trusted to provide stable and dependable information in an educational setting.

6 Results and Analysis

The results of these tests will be presented in the following section, detailing both the accuracy and consistency of the chatbot in handling the designated queries. This analysis will provide valuable insights into the effectiveness of the RAG system in enhancing the educational utility of chatbots.

The table presented bellow provides a comparison of responses generated by a chatbot under two conditions: one version without RAG capabilities and another version equipped with RAG. This side-by-side layout in the table allows for a clear visual assessment of the chatbot's performance in terms of accessing and referencing information from the provided books during the Q and A experiment setup.

Each row in the table corresponds to a different query, illustrating how each version of the chatbot responded. Responses from the chatbot with RAG are highlighted in green if the system successfully retrieved and accurately cited information from the external resources, demonstrating the effectiveness of the RAG system in enhancing the chatbot's ability to provide detailed, relevant, and contextually appropriate answers.

On the other hand, responses that failed to access the required information or provided general explanations without specific references are highlighted in red.

All the posed questions can be found both below and in the table for easy access.

1. What does chapter 5 on Introduction to Machine Learning with Python cover?
2. What is artificial intelligence?
3. Explain the concept of quantum computing.
4. Compare and contrast the advantages and disadvantages of renewable energy sources.
5. Can you give me the most important chapters on Modern Climate Change?
6. Analyze the impact of climate change on biodiversity.

Table 6.3: *View the full table under Appendix D*

Q and A experiment setup		
Question	**No RAG**	**With RAG**
What does chapter 5 on Introduction to Machine Learning with Python cover?	I don't have access to the contents of specific books, including "Introduction to Machine Learning with Python."	Chapter 5 of "Introduction to Machine Learning with Python" covers Model Evaluation and Improvement, including cross-validation, benefits of cross-validation, various strategies like stratified k-fold cross-validation, grid search techniques, the dangers of overfitting parameters and the validation set, and evaluation metrics and scoring, especially in the context of model selection.
What is artificial intelligence?	[...]	[...] [0:1 Introduction to Machine Learning with Python.]
Explain the concept of quantum computing.	[...]	The provided documents did not contain information on the concept of quantum computing; thus, I will give you a general explanation: [...]
Compare and contrast the advantages and disadvantages of renewable energy sources.	[...]	[...] [0:Aldo V. da Rosa - Fundementals of Renewable Energy Processes-Elsevier Academic Press (2005)] [1:Andrew Dessler - Introduction to Modern Climate Change-Cambridge university press (2011)]
Can you give me the most important chapters on Modern Climate Change?	[. . .]	The most important chapters of "Introduction to Modern Climate Change" [...] particularly chapters 1 to 3 that establish the introduction to the problem, and chapters 8, 11, 12, and 14 that deal with the future of climate, the fundamentals of climate change policy, mitigation policies, and long-term policy to address climate change.
Analyze the impact of climate change on biodiversity.	[...]	[...] [0:0 Introduction to Modern Climate Change].

It was found that five out of the six questions were successfully answered using information retrieved from the designated reference materials, with specific chapter references included in the responses. This indicates that the system effectively accesses and utilizes the external database to provide accurate and relevant information.

When the question "*Explain the concept of quantum computing*" was asked, the system correctly identified that this topic was not covered in the provided materials. Consequently, the response given was a general explanation, acknowledging the absence of specific information on quantum computing in the selected texts. This demonstrates the system's ability to recognize the limits of its knowledge base and adapt its responses accordingly.

It is also important to note that the responses given by the AI without RAG does not directly give hallucinated responses since they can be considered correct answers. Nonetheless, the use of RAG allows for more detailed and accurate information sourcing, enhancing the reliability and relevance of the responses, thus effectively reducing the occurrence of hallucinations.

A test to evaluate the consistency of the chatbot's responses will be presented in the next section. This involves asking the same question five times under identical conditions to determine if the chatbot can provide consistent answers across multiple inputs. Consistency is a critical metric in this context as it directly influences the reliability and stability of the chatbot's performance, which are essential for its practical use in educational settings.

The table below demonstrates the consistency test mentioned above. It can be observed that the model does not consistently use the dataset to reference material, as demonstrated by the response to the second question in the experiment. To address this inconsistency, the code could be modified to ensure that the model always employs the retrieval function when generating responses. This adjustment would guarantee that each answer is backed by data from the external sources, enhancing the reliability and factual grounding of the chatbot's responses.

Table 6.4: *Responses to "Analyze the impact of climate change on biodiversity." View the full table under Appendix E*

Responses
[0:2 Introduction to Modern Climate Change]; [...] [12:12 Introduction to Modern Climate Change].
The impact of climate change on biodiversity is significant, causing increased extinction risks, altered species interactions and ecosystems, loss of particular ecosystem types such as alpine and coastal mangroves, and disruptions to both land and marine life due to factors like sea level rise and ocean acidification.
[...] [0:2 Introduction to Modern Climate Change].
[...] [0:2 Introduction to Modern Climate Change-Cambridge University Press (2011)]; [...] [12:12 Introduction to Modern Climate Change-Cambridge University Press (2011)].
[...] [0:2 Introduction to Modern Climate Change].

The model's tendency to reference specific sections like chapters 0 and 12 could be because these chapters contain the relevant information needed to answer the questions that are posed. It suggests that the retrieval algorithm is effectively identifying and utilizing the most important sections of the text that align with the input questions.

In the next section, the research questions will be discussed in further detail.

7 Discussion

The findings of this research demonstrate the effectiveness of RAG in reducing hallucinations by leveraging external knowledge sources during the response generation process. The AWAITS chatbot, developed as part of this study, successfully incorporated RAG capabilities, enabling it to retrieve and utilize relevant information from the provided reference materials.

Through experimental evaluations and comparative analysis, the RAG-equipped AWAITS system demonstrated a significant improvement in providing accurate and context-specific responses compared to the non-RAG version of the chatbot (CHAT4-ISP). This outcome aligns with previous studies that have highlighted the potential of RAG systems to enhance factual accuracy and reduce inaccurate information.

The integration of RAG technology addresses a critical challenge faced by educational institutions and students: the need for reliable and authoritative sources of information to support learning, particularly in areas that require advanced expertise or specialized knowledge. By grounding the chatbot's responses in external databases and academic resources, the AWAITS system can serve as a valuable tool for enhancing critical thinking and writing skills among students.

Regarding our research questions.

7.1 RQ 1

Does the implementation or use of RAG help reduce hallucination problems?

The findings of this study affirms the use of RAG systems in resolving the long-standing challenge of hallucinations in AI-driven chatbots. By seamlessly integrating external, authoritative data sources into the response generation process, RAG technology anchors the chatbot's outputs in factual, verifiable content. This innovative approach represents a paradigm shift in chatbot architecture, effectively mitigating the risk of generating inaccurate, incoherent, or fabricated information – a persistent issue that has undermined the reliability of traditional chatbots.

Through experimental evaluations, the AWAITS system, powered by state-of-the-art RAG capabilities, consistently demonstrated its prowess in delivering responses substantiated by credible references. The incidence of hallucinations, which has plagued conventional chatbots, was significantly reduced, ushering in a new era of trustworthy and dependable AI-driven conversational agents.

7.2 RQ 2

Does integrating RAG in a chatbot affect the quality of advice?

The integration of RAG in chatbots has the potential to profoundly elevate the quality of advice and guidance provided to students, catalyzing the development of critical thinking and writing skills, essential competencies for academic and professional success.

By harnessing the power of RAG, the AWAITS system can dynamically access and synthesize relevant information from a vast array of external sources, including course materials, reference books, and domain-specific databases. This capability enables the chatbot to offer detailed, evidence-based explanations and recommendations, tailored to the specific inquiries and contexts of individual students.

The provision of such personalized, resource-backed guidance fosters a highly engaging and immersive learning environment. Students are actively encouraged to dig deeper into the material from the provided references, critically evaluate diverse perspectives, and develop thoughtful, well-substantiated arguments, skills that are essential to effective academic writing.

Moreover, by allowing students to directly provide course materials or lecture presentations, AWAITS can seamlessly integrate these resources into its responses, offering advice that is directly aligned with the subject matter at hand. This level of context-specificity not only promotes a deeper understanding of the material but also equips students with the tools to navigate complex academic landscapes with confidence and proficiency.

7.3 Comparative Analysis and Improvements

Table 7.5 below presents a detailed comparative analysis of three educational AI chatbots: AWAITS, CHAT4-ISP, and ChatGPT. This comparison is intended to elucidate the distinct features and capabilities of each system, emphasizing their strengths and limitations in various key aspects that are crucial for educational applications. The criteria for comparison include RAG abilities, educational focus, flexibility in model replacement, natural language understanding, customization options, support for multiple languages, and privacy and data collection practices. Each chatbot's features are assessed to determine their suitability for enhancing educational experiences, tailoring responses to specific educational needs, and adhering to privacy standards such as the GDPR.

Table 7.5: Comparative Analysis of Educational AI Chatbots

Features	**AWAITS**	**CHAT4-ISP**	**ChatGPT**
RAG abilities / Data Retrieval	Included without user configuration, set by authorative bodies (teachers)	None	General, users can choose to use this option but requires additional steps
Educational Focus	Strong, specifically designed for educational purposes	Moderate, general assistance with some educational features	General, not specifically educational but adaptable
Model replacement	The model can be switched to alternatives and not limited to OpenAI's ChatGPT	The model can be switched to alternatives and not limited to OpenAI's ChatGPT	Only ChatGPT
Natural Language Understanding	Advanced, with context-aware responses and tailored preprompts	Basic preprompting	Requires the user to construct preprompts
Customization Options	real-time modifications to preprompts and knowledge base by authorative body (teachers)	Minimal customization	Minimal customization
Support for Multiple Languages	Yes, multiple languages supported such as English and Swedish	No, primarily English	Yes, supports multiple languages out the box
Privacy and Data collection	Safe data collection, only sotred locally; GDPR complient	Safe data collection, only sotred locally; GDPR complient	Automatic data collection for model training; violates GDPR

While previous studies have explored the integration of RAG systems in conversational AI [14, 19], the AWAITS project represents a advancement in the application of this technology specifically within educational contexts. By leveraging RAG to address the persistent issue of hallucinations in chatbots, this study has demonstrated the potential to enhance the reliability and trustworthiness of AI-powered educational tools.

Existing chatbots, such as the CHAT4-ISP system [9], suffer from the inherent limitation of generating responses based solely on their pre-trained language models. Consequently, the risk of hallucinations, or the production of factually incorrect or incoherent information, remains a significant challenge. The AWAITS project addresses this limitation by seamlessly integrating RAG capabilities, enabling the chatbot to dynamically retrieve and incorporate relevant information from external, authoritative sources during the response generation process.

This approach represents a significant departure from traditional chatbot architectures and offers several key improvements. Firstly, by grounding responses in verifiable data sources, AWAITS mitigates the risk of hallucinations, providing students with accurate and reliable information. This is a critical advancement, as misinformation can be detrimental to the learning process, particularly in academic settings where precise and well-substantiated knowledge is essential.

Secondly, the AWAITS project demonstrates the potential of RAG systems to enhance the depth and context-specificity of the advice provided to students. By drawing upon external resources such as course materials, reference books, or domain-specific databases, the chatbot can offer detailed, resource-backed explanations tailored to the student's inquiries. This level of personalized guidance not only fosters a more engaging learning experience but also cultivates essential skills such as independent research, critical thinking, and effective academic writing.

Furthermore, the AWAITS project underscores the versatility of RAG systems in adapting to diverse educational domains. By incorporating a wide range of external data sources, from scientific literature to humanities texts, the chatbot can provide reliable support across various academic disciplines. This flexibility addresses the limitations of traditional chatbots, which may struggle to maintain accuracy and relevance when operating outside their narrow domains of expertise.

8 Conclusion

This thesis demonstrates the integration of RAG, improving the accuracy and relevance of responses generated by LLMs. By employing RAG, the issue of hallucinations that are commonly associated with LLMs [15, 19] has been effectively reduced, resulting in an enhancement to the educational tools' reliability and trustworthiness. The findings highlight RAG's capability to access, reference, and utilize external knowledge databases, ensuring that the content is not only precise but also backed by authoritative sources. While the occurrence of hallucinations was not immediately apparent, equipping LLMs with advanced tools and capabilities will effectively reduce the incidence of hallucinations.

As educational institutions continue to integrate technology into their curriculum, tools like AWAITS that support accurate and reliable information retrieval become all the more essential. This study showcases the potential of AI to complement traditional educational methods, demonstrating its ability to enhance the learning experience through technology [14, 29].

The methodologies applied in the AWAITS project have broader applicability than just education. For instance, similar RAG systems could be utilized in other fields requiring information accuracy and contextual relevance, for example in legal research, medical inquiry, and content creation [28, 41]. The ability to integrate relevant information from data sources can significantly benefit professions and industries where decision making relies heavily on the quality and reliability of the information.

In summary, this thesis confirms the effectiveness of RAG in enhancing the reliability and accuracy of responses by LLMs, specifically addressing common issues such as hallucinations. Through the AWAITS project, it was demonstrated that RAG's ability to integrate and utilize data from external sources can substantially improve the quality of information provided by educational chatbots. This makes RAG-equipped tools like AWAITS crucial for educational settings that increasingly rely on technology to supplement traditional teaching methods [14, 28].

The versatility of RAG extends beyond education, with potential applications in fields such as legal research, medical inquiry, and content creation, where accurate and contextually relevant information is vital [31, 32]. However, the project also highlighted areas for improvement, including expanding the data sources for RAG and utilizing all available features to enhance functionality and accuracy

8.1 Future work

Future efforts should concentrate on refining the integration of RAG technology, particularly by expanding the diversity and scope of external data sources to improve responsiveness and precision. Additionally, ongoing development should aim to optimize the retrieval processes to minimize latency and enhance user experience, especially in real-time educational applications. By addressing these areas, RAG can become even more impactful across a variety of professional fields, not only enhancing educational tools but also advancing decision-making processes where accurate and timely information is crucial.

Future research could also explore expanding the databases linked to the chatbot to include a broader range of subjects and more diverse types of information. This expansion would likely enhance the chatbot's utility, making it an even more versatile and powerful tool for students and educators alike. Additionally, future developments could focus on optimizing the retrieval processes to minimize response times and integrating more advanced natural language processing techniques to improve the chatbot's understanding of complex queries. In conclusion, the integration of RAG into the AWAITS chatbot has not only enhanced its functionality but has also set a new standard for the application of AI in education. It highlights the potential of AI to support traditional educational methods and opens up new avenues for research and application in other sectors. As we continue to refine and expand this technology, its impact on both educational settings and other information-intensive fields is likely to be profound.

More specifically for the AWAITS project, future efforts should focus on several key improvements, such as

Explore Additional Libraries:
Investigating other libraries like LangChain could reveal additional features and tools that enhance the functionality of the AWAITS project. Exploring these alternatives may provide new avenues for improving response accuracy and system flexibility.

Thoroughly Review of API Documentation:
A detailed review of the API documentation is essential to fully utilize all features and capabilities of the RAG system. A thorough understanding can uncover new optimization opportunities that enhance the system's performance and functionality.

Run the Model Locally:
Considering running the RAG model locally to speed up response times and enhance data security. Local execution could significantly reduce latency and increase the system's efficiency.

Visualize Time and Cost Differences:
It would be beneficial to plot and analyze the differences in time, cost and token usage between RAG and non-RAG systems. These visualizations will provide clear, comparative insights into the efficiency and economic impact of employing RAG technology, aiding in more informed decision-making.

References

[1] Q. Yang and C. Wang, “The fourth revolution,” France, 2018, accessed: 2024-05-26. [Online]. Available: https://policycommons.net/artifacts/8203878/the-fourth-revolution/9114768/

[2] L. Floridi, *The fourth revolution: How the infosphere is reshaping human reality*. Oxford University Press, 2014.

[3] M. P. Pratama, R. Sampelolo, and H. Lura, “Revolutionizing Education: Harnessing the Power of Artificial Intelligence for Personalized Learning,” *KLASIKAL: Journal of Education, Language Teaching and Science*, Aug 2023. [Online]. Available: https://www.journalfkipuniversitasbosowa.org/index.php/klasikal/article/view/877

[4] D. Bartholomae, “Inventing the University,” *Journal of Basic Writing*, vol. 5, no. 1, pp. 4–23, 1986.

[5] S. Bailey, *Academic Writing: A Handbook for International Students*, 4th ed. Routledge, 2014.

[6] L. A. Fernsten and M. Reda, “Helping students meet the challenges of academic writing,” *Teaching in Higher Education*, Apr 2011. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/13562517.2010.507306?casa_token=NL8KTASnP0EAAAAA%3AdUEuK3C9QxNQiwsNhVf4X-yivrANbSCdzUhgv9F4faLEX6dkIHL-s0AYahayhAlXMa1fQhcR

[7] A. Bedington, E. F. Halcomb, H. A. McKee, T. Sargent, and A. Smith, “Writing with generative AI and human-machine teaming: Insights and recommendations from faculty and students,” *Computers and Composition*, Mar. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S8755461524000094

[8] C. Ka and Y. Chan, “Is AI Changing the Rules of Academic Misconduct? An In-depth Look at Students’ Perceptions of “AI-giarism”,” 2023. [Online]. Available: https://arxiv.org/abs/2306.03358

[9] M. Taiye, C. High, J. Velander, K. Matar, R. Okmanis, and M. Milrad, “Generative ai-enhanced academic writing: A stakeholder-centric approach for the design and development of chat4isp-ai,” 2024. [Online]. Available: https://doi.org/10.1145/3605098.3636055

[10] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, “A survey on Large Language Model (LLM) security and privacy: The Good, The Bad, and The Ugly,” *High-Confidence Computing*, Mar. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S266729522400014X

[11] OpenAI, “GPT-4 Technical Report,” Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[12] Z. Jiang *et al.*, “Active Retrieval Augmented Generation,” 2023. [Online]. Available: https://arxiv.org/pdf/2305.06983.pdf

[13] R. Zhao *et al.*, “Retrieving Multimodal Information for Augmented Generation: A Survey,” Oct. 2023, accessed: Oct. 29, 2023. [Online]. Available: https://arxiv.org/pdf/2303.10868.pdf

[14] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering," *Transactions of the Association for Computational Linguistics*, 2023. [Online]. Available: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00530/114590/Improving-the-Domain-Adaptation-of-Retrieval

[15] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, and L. Yuan, "LLM LIES: HALLUCINATIONS ARE NOT BUGS, BUT FEATURES AS ADVERSARIAL EXAMPLES," 2024. [Online]. Available: https://arxiv.org/pdf/2310.01469

[16] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," 2023. [Online]. Available: https://arxiv.org/pdf/2312.10997.pdf

[17] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval: The Concepts and Technology Behind Search.* New York: Addison Wesley, 2011.

[18] K. C. Siontis, Z. I. Attia, S. J. Asirvatham, and P. A. Friedman, "ChatGPT hallucinating: can it get any more humanlike?" *European Heart Journal*, vol. 45, no. 5, pp. 321–323, 12 2023. [Online]. Available: https://doi.org/10.1093/eurheartj/ehad766

[19] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval Augmentation Reduces Hallucination in Conversation," May 2024. [Online]. Available: https://arxiv.org/pdf/2104.07567

[20] M. Bond *et al.*, "A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour," *International Journal of Educational Technology in Higher Education*, Jan. 2024. [Online]. Available: https://link.springer.com/article/10.1186/s41239-023-00436-z

[21] Google, "Our next-generation model: Gemini 1.5," Feb 2024. [Online]. Available: https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#context-window

[22] N. Cennetkuşu, "Journal of Language and Linguistic Studies," *Journal of Language and Linguistic Studies*, 2017. [Online]. Available: https://dergipark.org.tr/en/download/article-file/440773

[23] Cecylia Borek, "Comparative Evaluation of LLM-Based Approaches to Chatbot Creation: Implementing a Death Doula Chatbot," May 2024. [Online]. Available: https://trepo.tuni.fi/bitstream/handle/10024/154995/BorekCecylia.pdf?sequence=2

[24] I. V. D. W. Supervisor and A. Liesenfeld, "Bachelor Thesis Evaluating hallucinations and repair in open-domain dialogue systems," May 2024. [Online]. Available: https://theses.ubn.ru.nl/server/api/core/bitstreams/914ee283-a9a7-434d-8bfd-0a42c5db05b3/content

[25] Pinecone, "Rag evaluation: Don't let customers tell you first," 2024, accessed: May 19, 2024. [Online]. Available: {https://www.pinecone.io/learn/series/vector-databases-in-production-for-busy-engineers/rag-evaluation/},

[26] “The claude 3 model family: Opus, sonnet, haiku anthropic,” 2023, accessed: date-of-access. [Online]. Available: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

[27] C. Jeong, “A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture,” *Advances in Artificial Intelligence and Machine Learning*, 2023. [Online]. Available: https://arxiv.org/abs/2309.01105

[28] G. Biswas, T. Katzlberger, J. Bransford, and D. Schwartz, “Extending Intelligent Learning Environments with Teachable Agents to Enhance Learning,” May 2024. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=45c96840ce57cc1046bfa4cb2f285b94d63c637a

[29] X. Chen, H. Xie, and G.-J. Hwang, “A multi-perspective study on Artificial Intelligence in Education: grants, conferences, journals, software tools, institutions, and researchers,” *Computers and Education: Artificial Intelligence*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X20300059

[30] D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba, “Understanding the role of individual units in a deep neural network,” *Proceedings of the National Academy of Sciences*, Sep. 2020. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1907375117

[31] C. Larman and V. R. Basili, “Iterative and Incremental Development: A Brief History,” *IEEE Computer*, Jun 2003. [Online]. Available: {https://scrummaster.dk/lib/AgileLeanLibrary/People/CraigLarman/history-of-iterative-larman-and-basili-ieee-computer.pdf}.[Accessed:May19,2024]
.

[32] H. Wang *et al.*, “UniMS-RAG: A Unified Multi-source Retrieval-Augmented Generation for Personalized Dialogue Systems,” 2024, [Accessed: May 19, 2024]. [Online]. Available: https://scrummaster.dk/lib/AgileLeanLibrary/People/CraigLarman/history-of-iterative-larman-and-basili-ieee-computer.pdf

[33] Andreas, *Introduction To Machine Learning With Python*. Oreilly Associates Inc, 2016.

[34] A. V. D. Rosa, *Fundamentals of renewable energy processes*. Academic, 2012.

[35] A. Jedlitschka and D. Pfahl, “Reporting guidelines for controlled experiments in software engineering,” in *2005 International Symposium on Empirical Software Engineering, 2005.*, 2005, pp. 10 pp.–.

[36] A. Blandford, A. Cox, and P. Cairns, “Controlled experiments,” 2008, accessed: date-of-access. [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/14774/1/14774.pdf

[37] I. Silderhuis, “Validity and reliability of the user satisfaction with information chatbots scale (usic),” Master’s thesis, Faculty of Behavioural, Management and Social Sciences, Human Factors and Engineering Psychology, September 2020, master’s thesis. [Online]. Available: https://essay.utwente.nl/83495/

[38] Y. Yu *et al.*, “Large language model as attributed training data generator: A tale of diversity and bias,” Available online, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/ae9500c4f5607caf2eff033c67daa9d7-Paper-Datasets_and_Benchmarks.pdf

[39] A. Nassar and M. Kamal, “Ethical dilemmas in AI-powered decision-making: A deep dive into big data-driven ethical considerations,” *International Journal of Responsible Artificial Intelligence*, Aug 2021. [Online]. Available: https://neuralslate.com/index.php/Journal-of-Responsible-AI/article/view/43/28

[40] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, and H. Wachsmuth, “Who determines what is relevant? humans or ai? why not both?” *Commun. ACM*, vol. 67, no. 4, p. 31–34, mar 2024. [Online]. Available: https://doi.org/10.1145/3624730

[41] P. Feldman, J. Foulds, and S. Pan, “RAGged Edges: The Double-Edged Sword of Retrieval-Augmented Chatbots *,” May 2024. [Online]. Available: https://arxiv.org/pdf/2403.01193

A Appendix A

Creating the assistant

```
# create Assistant
'''
This is the assistant that will be used to answer questions about the b

the assistant is only create once and can be used multiple times. with
'''
assistant = client.beta.assistants.create(
    name="Awais experimental assistant",
    instructions="You can ask me anything about writing a good paper",
    #
    # retrieval tool is used to implement a RAG-system
    #
    tools=[{"type": "retrieval"}],
    model="gpt-4-1106-preview",
    file_ids=[]
)
```

Adding the vector store (files) to the assistant to use as retrieval

```
assistant = client.beta.assistants.update(
  assistant_id="asst_7g5xvjjEot9flZbcywjorIc6",
  tool_resources={"file_search": {"vector_store_ids": ["vs_MnG0twx7JLZS
)
```

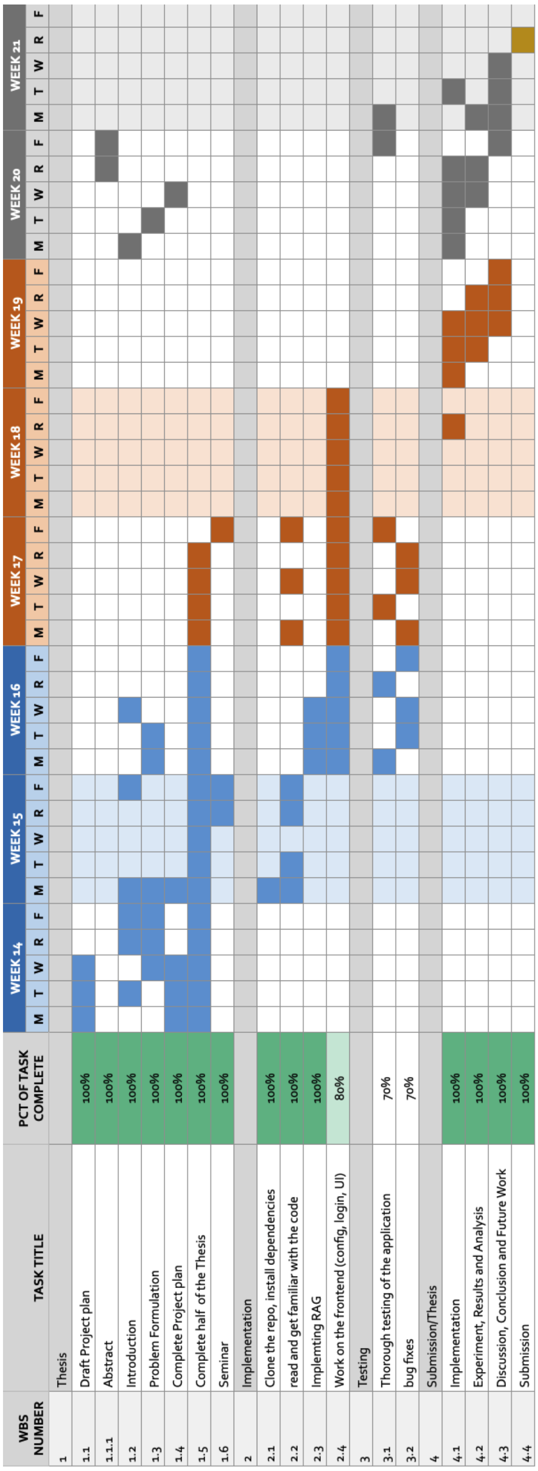
Our api route to fetch the response

```
def get_analyst_response(
    prompt: str, run_instruction: str = ""
) -> str:
    """Get chat response from OpenAI API using the Assistant api

    Args:
        chat_history (List[Dict[str, str]]): List of chat history

    Returns:
        str: Response from OpenAI API
    """
    message_content = prompt

    # TODO: since we are now using threads, we only need to send the us
    # and not the whole history to the backend, either leave it or refa
    # to only send the user message
    message = client.beta.threads.messages.create(
        thread_id=thread.id, role="user", content=message_content
    )

    run = client.beta.threads.runs.create(
        thread_id=thread.id,
        assistant_id=assistant_id,
        instructions=run_instruction,
    )

    # TODO: check if this will this be bottle neck
    while run.status != "completed":
        run = client.beta.threads.runs.retrieve(thread_id=thread.id, ru
        sleep(0.5)

    messages = client.beta.threads.messages.list(thread_id=thread.id)
    message = messages.data[0]
    return handle_response_format(message)
```

Handle the response and format, this is copied from OpenAI documentation

```
def handle_response_format(message):
    # from openai docs
    # retrieve and format mesasge

    # Extract the message content
    message_content = message.content[0].text
    annotations = message_content.annotations
    citations = []

    # Iterate over the annotations and add footnotes
    for index, annotation in enumerate(annotations):
        # Replace the text with a footnote
        message_content.value = message_content.value.replace(
            annotation.text, f" [{index}]"
        )

        # Gather citations based on annotation attributes
        if file_citation := getattr(annotation, "file_citation", None):
            cited_file = client.files.retrieve(file_citation.file_id)
            citations.append(
                f"[{index}] {file_citation.quote} from {cited_file.file
            )
        elif file_path := getattr(annotation, "file_path", None):
            cited_file = client.files.retrieve(file_path.file_id)
            citations.append(
                f"[{index}] Click <here> to download {cited_file.filena
            )
            # Note: File download functionality not implemented above f

    # Add footnotes to the end of the message before displaying to user
    message_content.value += "\n" + "\n".join(citations)

    return message_content.value
```

B Appendix B



WBS NUMBER	TASK TITLE	PCT OF TASK COMPLETE
1	Thesis	
1.1	Draft Project plan	100%
1.1.1	Abstract	100%
1.2	Introduction	100%
1.3	Problem Formulation	100%
1.4	Complete Project plan	100%
1.5	Complete half of the Thesis	100%
1.6	Seminar	100%
2	Implementation	
2.1	Clone the repo, install dependencies	100%
2.2	read and get familiar with the code	100%
2.3	Implemting RAG	100%
2.4	Work on the frontend (config, login, UI)	80%
3	Testing	
3.1	Thorough testing of the application	70%
3.2	bug fixes	70%
4	Submission/Thesis	
4.1	Implementation	100%
4.2	Experiment, Results and Analysis	100%
4.3	Discussion, Conclusion and Future Work	100%
4.4	Submission	100%

Figure 2.8: *Gantt Chart illustrating the timeline and task progression for the thesis project. The chart details weekly activities from project initiation to final submission, highlighting key phases such as planning, implementation, testing, and completion across weeks 14 to 21.*

C Appendix C

A config page was also built that an assigned administrator can modify the Assistant's pre-prompt as desired, providing teachers with an extra layer of control over the type of responses generated.

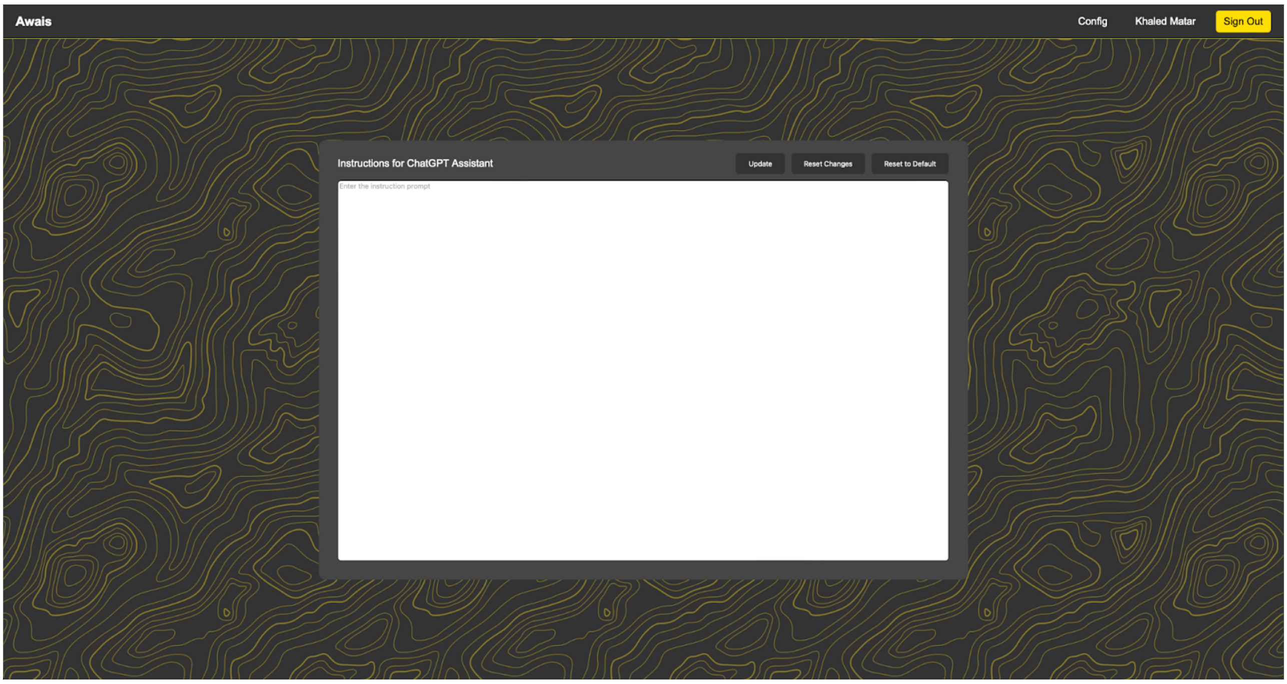


Figure 3.9: */config route UI in dark theme*

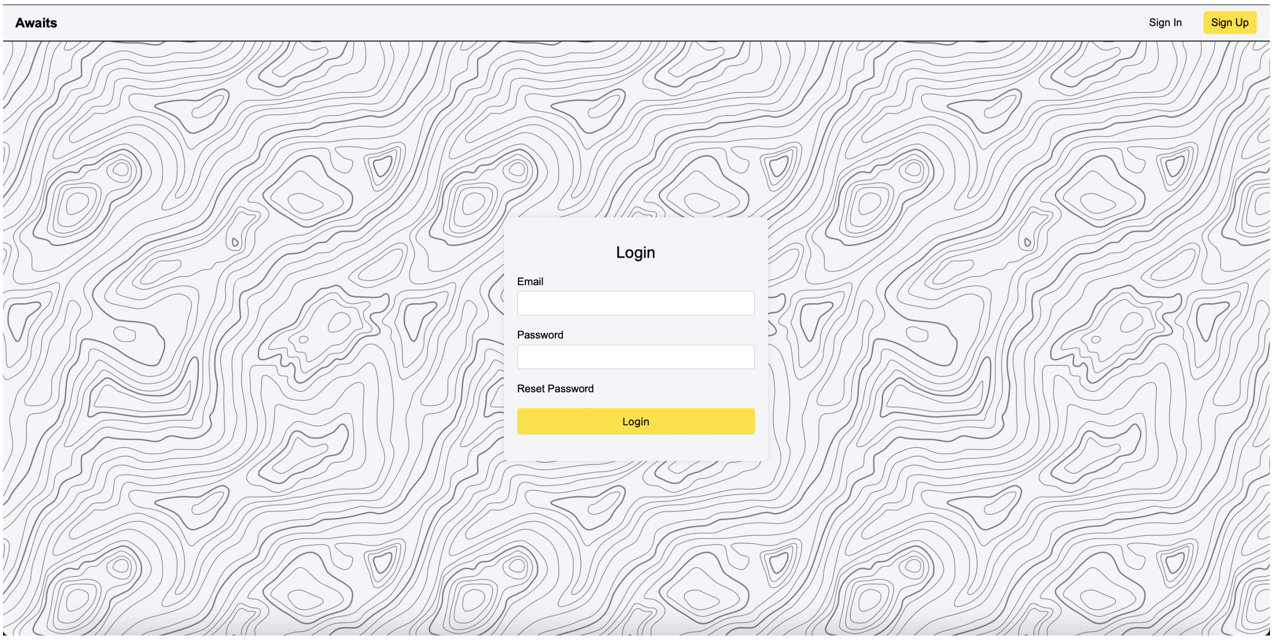


Figure 3.10: */login route UI*

D Appendix D

Table 4.6: *View the full table under Appendix D*

Q and A experiment setup		
Question	**No RAG**	**With RAG**
What does chapter 5 on Introduction to Machine Learning with Python cover?	I don't have access to the contents of specific books, including "Introduction to Machine Learning with Python."	Chapter 5 of "Introduction to Machine Learning with Python" covers Model Evaluation and Improvement, including cross-validation, benefits of cross-validation, various strategies like stratified k-fold cross-validation, grid search techniques, the dangers of overfitting parameters and the validation set, and evaluation metrics and scoring, especially in the context of model selection.
What is artificial intelligence?	Artificial Intelligence (AI) is the branch of computer science that focuses on creating machines capable of performing tasks that normally require human intelligence, such as learning, decision-making, and problem-solving.	Artificial intelligence (AI) is a research field at the intersection of statistics, artificial intelligence, and computer science, also known as predictive analytics or statistical learning; it involves the development of algorithms that can learn from and make predictions or decisions based on data [0:1 Introduction to Machine Learning with Python.]
Explain the concept of quantum computing.	Quantum computing is a type of computing that utilizes quantum-mechanical phenomena, such as superposition and entanglement, to perform operations on data at speeds significantly faster than traditional computers for certain types of problems.	The provided documents did not contain information on the concept of quantum computing; thus, I will give you a general explanation: Quantum computing is a form of computation that utilizes quantum-mechanical phenomena, such as superposition and entanglement, to perform operations on data, with quantum bits, or qubits, being the basic units that can represent multiple states simultaneously, significantly increasing computational power for some tasks compared to classical computers.

Continued on next page

Q and A experiment setup(continued)		
Question	**No RAG**	**With RAG**
Compare and contrast the advantages and disadvantages of renewable energy sources.	Renewable energy sources offer environmental sustainability and reduced greenhouse gas emissions but face challenges such as intermittency, higher initial costs, and the need for large physical footprints and advanced energy storage solutions.	Advantages of renewable energy include emissions of no CO2 for wind power, operation without leaving dangerous residues, lower decommissioning costs, dual-purpose lands (agricultural use possible), and renewable sources being non-depleting; disadvantages encompass the intermittent nature of wind and solar, lower energy production, large land area requirements, higher equipment mass, difficulty in storing some fuel types, and the need for energy storage or backup power to ensure supply reliability [0:Aldo V. da Rosa - Fundementals of Renewable Energy Processes-Elsevier Academic Press (2005)] [1:Andrew Dessler - Introduction to Modern Climate Change-Cambridge university press (2011)]
Can you give me the most important chapters on Modern Climate Change?	Sure, here's a list of commonly regarded important chapters to focus on when studying Modern Climate Change: The Science of Climate Change - This chapter usually covers the fundamental scientific principles of climate science, including the greenhouse effect, climate feedbacks, and the role of various gases and aerosols. [. . .]	The most important chapters of "Introduction to Modern Climate Change" seem to be the foundational chapters that lay out the scientific evidence of climate change, the underlying physical principles, the future projections of climate, and the policy and economic considerations related to addressing climate change, particularly chapters 1 to 3 that establish the introduction to the problem, and chapters 8, 11, 12, and 14 that deal with the future of climate, the fundamentals of climate change policy, mitigation policies, and long-term policy to address climate change.
Analyze the impact of climate change on biodiversity.	Climate change significantly impacts biodiversity by altering habitats, disrupting species' distribution and life cycles, and increasing the frequency of extreme weather events, leading to heightened risks of extinction for vulnerable species and the degradation of ecosystems.	Climate change impacts biodiversity by driving species like lizards towards extinction, disrupting ecosystems, causing loss of certain ecosystem types like alpine, mangrove, and coral reef systems, and affecting water availability, which in turn can lead to decreases in water resources and increases in disease risk due to altered distributions of disease vectors like mosquitoes [0:0 Introduction to Modern Climate Change].

E Appendix E

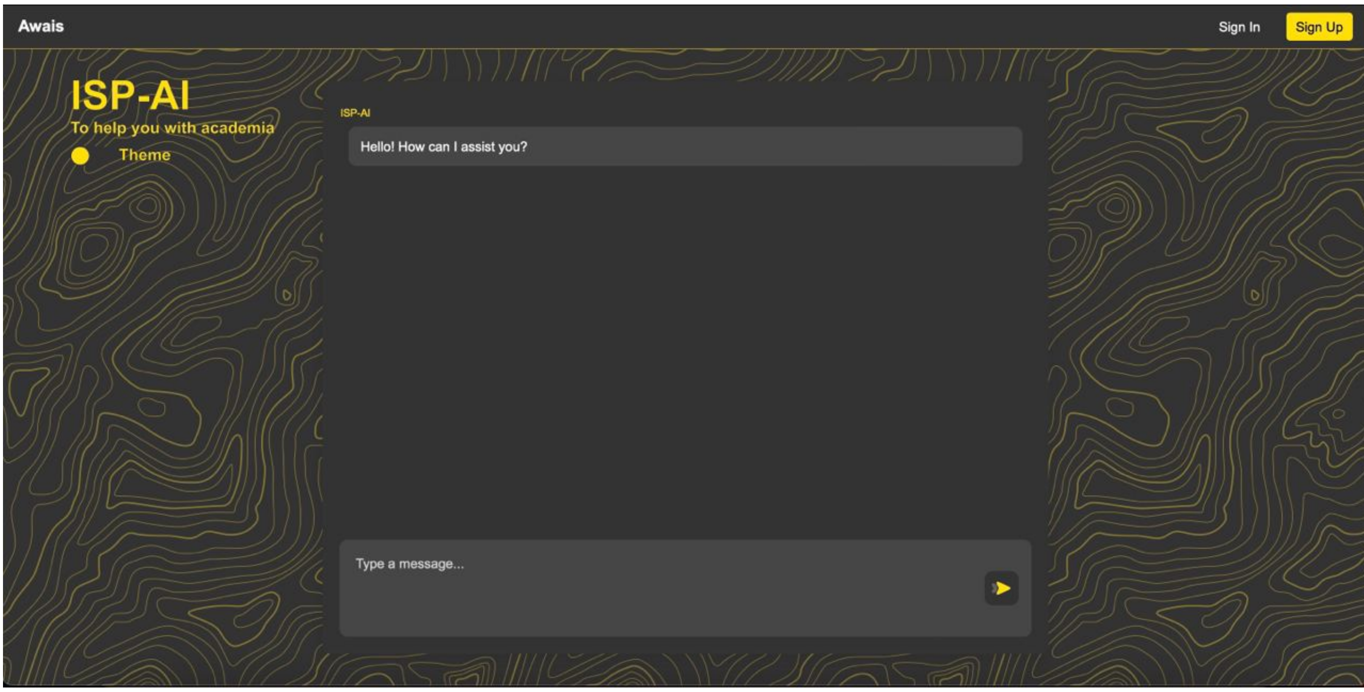
Consistency	
Question	**Response**
Analyze the impact of climate change on biodiversity.	The impact of climate change on biodiversity is significant and multifaceted, including increased risk of plant and animal species extinction, shifts in ecosystem composition, and alteration of water availability [0:2 Introduction to Modern Climate Change]; warmer temperatures will drive some species towards extinction due to changes in foraging times and food availability, ecosystems will evolve with new species relationships and ecological surprises, and there will likely be losses of ecosystem types like alpine systems, coastal mangrove systems, and coral reefs [0:2 Introduction to Modern Climate Change]; increased sea levels and ocean acidification from higher carbon dioxide levels will further disrupt both terrestrial and marine life and associated human activities [12:12 Introduction to Modern Climate Change].
Analyze the impact of climate change on biodiversity.	The impact of climate change on biodiversity is significant, causing increased extinction risks, altered species interactions and ecosystems, loss of particular ecosystem types such as alpine and coastal mangroves, and disruptions to both land and marine life due to factors like sea level rise and ocean acidification.
Analyze the impact of climate change on biodiversity.	The impact of climate change on biodiversity is extensive, with a variety of consequences such as increased extinction risks for many species due to temperature changes impacting their survival, changes in ecosystems leading to new species relationships and potential loss of certain ecosystem types like alpine and coastal mangroves, disruptions in marine and terrestrial life due to sea level rise and ocean acidification, and human societies feeling the effects through the loss of services provided by these ecosystems [0:2 Introduction to Modern Climate Change].

Continued on next page

Consistency (continued)	
Question	**Response**
Analyze the impact of climate change on biodiversity.	The impact of climate change on biodiversity arises from multiple facets such as increased extinction risk for numerous species as temperatures rise, altered species interactions, the evolution of ecosystems with potential loss of specific types like alpine and coastal mangroves, sea level rise, and ocean acidification affecting marine and terrestrial life, with human activities being impacted by these ecological changes [0:2 Introduction to Modern Climate Change-Cambridge University Press (2011)]. Additionally, changes in precipitation patterns and water availability, along with human health impacts due to increased air pollution and disease risks as temperatures rise, present further challenges [12:12 Introduction to Modern Climate Change-Cambridge University Press (2011)].
Analyze the impact of climate change on biodiversity.	The impact of climate change on biodiversity includes increased extinction risks for plant and animal species, altered species interactions and ecosystems, potential loss of specific ecosystems such as alpine and coastal mangroves, and disruptions to both land and marine life due to sea level rise and ocean acidification, which can also affect human societies dependent on these biological resources [0:2 Introduction to Modern Climate Change].

Continued on next page

F Appendix F



Hello! How can I assist you?

Figure 6.11: *CHAT4-ISP UI*

G Appendix G

Mohammed Taiye, Christopher High, Johanna Velander, Khaled Matar, Rihards Okmanis, and Marcelo Milrad. 2024. Generative AI-Enhanced Academic Writing: A Stakeholder-Centric Approach for the Design and Development of CHAT4ISP-AI. In Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24). Association for Computing Machinery, New York, NY, USA, 74–80. https://doi.org/10.1145/3605098.3636055