BINDURA UNIVERSITY OF SCIENCE EDUCATION

FACULTY OF SCIENCE & ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE

RESEARCH PROJECT

SUBMITTED BY: **EMMANUEL ANESU CHINGWENA**

REGNUMBER: **B213796B**

LEVEL: **4:2**

TOPIC: **EVALUATING THE IMPACT OF INTELLIGENT TUTORING SYSTEMS ON TEACHING AND LEARNING IN SECONDARY EDUCATION**

SUPERVISOR: **MR C. ZANO**

A PROJECT SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE AT BINDURA UNIVERSITY OF SCIENCE EDUCATION IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE (HONS) IN COMPUTER SCIENCE

APPROVAL FORM

The undersigned certify that they have supervised the student Emmanuel Anesu Chingwena the Dissertation entitled, “EVALUATING THE IMPACT OF INTELLIGENT TUTORING SYSTEMS ON TEACHING AND LEARNING IN SECONDARY EDUCATION” submitted in partial Fulfillment of the requirements for a Bachelor Computer Science Honors Degree at Bindura University of Science Education.

Student Signature Date: 20/06/2025

Supervisor Signature Date: 20/06/2025

Chairman Signature Date: 20/06/2025

ABSTRACT

The paper is a new assessment of the contribution of Intelligent Tutoring Systems (ITS) to the learning and teaching process of secondary education by designing, developing and testing a tailored it's that combines GPT-based quiz generation and a retrieval-augmented Chabot. By automatically creating quiz scenarios that align with learning goals and providing adaptive, document-based feedback, the system aims to enhance individualised learning. An analytical study was committed in mixed methods, where the results were evaluated through secondary school learners and educators and focused on learning outcomes, user interaction, system usability, and performance. Findings indicated that there was a notable increase in the learning achievement of students and a good overall perception towards the ease of using the system and delivering instructions. The combination of AI methods, including large language models and retrieval enrichment, demonstrated favourable flexibility and scalability compared to traditional ITS methods. The paper comments on the practical implications, limitations, and future research directions on how it is possible to use advanced AI to facilitate effective and scalable education in secondary education schools.

ACKNOWLEDGEMENTS

I would also wish to thank Mr Zano for encouraging and motivating me in the development of this project. I would also like to thank all the lecturers of computer science at Bindura University of Science Education with regards to their advice and hard work. I would like to express a special appreciation to Lloyd Chingwena, who dedicated his support and helped me throughout my studies.

ABBREVIATIONS

AI	Artificial Intelligence
FAQ(s)	Frequently Asked Question(s)
GPT	Generative Pre-trained Transformer
ITS	Intelligent Tutoring System

KT	Knowledge Tracing
LLM	Large Language Model
MCQ	Multiple Choice Question
NLP	Natural Language Processing
RAG	Retrieval-Augmented Generation
PDF	Portable Document Format
UI	User Interface
API	Application Programming Interface
DB	Database
ORM	Object-Relational Mapping
UX	User Experience
SD	Standard Deviation
HTML	HyperText Markup Language
CSS	Cascading Style Sheets
JWT	JSON Web Token
CRUD	Create, Read, Update, Delete

Table of Contents

Chapter 1: introduction

1.1 Introduction

Technological change, in integrating technology into education, has brought many changes in the way teaching and learning take place at every level of schooling. An example of recent innovation in education is the Intelligent Tutoring System (ITS), which is a computer-based learning environment that offers individualised tutoring and reinforcement to learners without any sensory involvement. They allow artificial intelligence methodologies to support a learner in response to his or her needs, and the assistance provided by them resembles one-on-one human tutoring (VanLehn, 2011).

In secondary education, where the classroom is usually varied in its learners abilities and needs, ITS provide an opportunity to close the achievement gap and promote differentiated learning. ITS has the potential of improving student engagement and motivation as well as performance by modifying instruction in real time as a result of continuous evaluation of the input of students (Ma, Adesope, Nesbit, & Liu, 2014). Another way through which teachers can gain through ITS are by utilising the information that will be produced to inform instruction and target remediation.

In spite of the potential of ITS, there is a doubt of their real effects in real-world scenarios, especially in secondary schools. Their effectiveness may be affected by factors including implementation issues, readiness of the teacher, and quality of the system that is being designed (Nye, 2015). Consequently, it is necessary to look into the ITS contribution to the teaching and learning process in secondary education to know whether they are effective and should have some process of improvement on them.

This paper aims at discussing the discussion about the effects of intelligent tutoring systems on students and teachers in secondary education in terms of academic performance, involvement, and instruction.

1.2 Background

In the context of generative AI (after 2020), Intelligent Tutoring Systems (ITS) have gained momentum by being combined with adaptive learning algorithms and emotion-sensitive interfaces. The systems provide customised learning opportunities, similar to tutoring at an individual level, covering all kinds of needs in secondary learning (Wang, Ribeiro, et al., 2024; Wang et al., 2025).

As an exemplar, there was a field experiment in Ghana (Withaker et al., 2019) that applied Rori, an AI-based math helper provided through WhatsApp. After eight months of meetings on a biweekly basis, students achieved better results in math ($d = 0.37$, $p < .001$), which indicates the potential of ITS in low-resource schools. In the meantime, Olney et al. (2025) compared a biology-based ITS, which simulated knowledgeable human tutors, to other methods of traditional classroom instruction, and they achieved big short-term gains in learning ($d = .71$) and long-term retention ($d = .36$).

A systematic review of Son (2024) summarised the research in mathematics education where the recent ITS implementations have already transitioned beyond augmentation, that is, to provide active reinvention of instructional actions and learning. As a complement to this, Liu, Latif, and Zhai (2025) analysed 86 ITS and robot tutoring systems and found that considerable progress was observed as regards adaptive intelligence and engagement strategies, but the questions of ethics, scalability, and AI bias still need to be answered.

Affective ITS endowed with the capacity to identify and act on the emotional conditions of learners have been found promising. In one example, a scoping review by Fernandez-Herrero (2024) determined that there was agreement in the improvement of student satisfaction, motivation, and performance as combined emotion-aware mechanisms were used.

There is also a possibility of scalability and effectiveness of human-AI amalgamated models. A report based upon 900 tutors and 1,800 K to 12 US students, the so-named study of Tutor CoPilot, showed a 4 percentage point rise in student mastery with even more gains in tutors with lower mastery.

Despite these favourable tendencies, the adoption of ITS in secondary schools in real classrooms is difficult. Reviews of social experiments reveal that their performance differs in contexts involving dissimilar teacher facilitation, student cooperation, and institutional conditions. Ethical concerns, such as data confidentiality, addiction, and bias in emotion detection, also hang over like a dark cloud and should be dealt with carefully.

Concisely, the following three key themes in ITS research in secondary education have been highlighted in the most recent studies:

1. Positive intellectual improvement: Researchers have indicated moderate to large attainments in subject-essential improvements.

2. Scalable deployment: Light-weight and mobile solutions and hybrid solutions can increase ITS coverage to areas of limited resources.

3. Emotional and ethical sophistication: Affective and hybrid ITS enhance learning, and they require ethical protection and sensitive design to contexts.

This contemporary evidence base lays a solid foundation in exploring the effect the ITS has on the teaching process and achievement of students in secondary settings of education.

1.3 Problem Statement

Although Intelligent Tutoring Systems (ITS) have reported positive results in relation to improving personalised learning and student achievement, there is still no in-depth assessment of the influence of systems on the teaching and learning process in terms of how it improves teaching practices and student achievement in secondary education environments. The current studies in the field of ITS tend to approach cognitive benefits very critically and not focus on all the changes that ITS integration implies in terms of teaching strategies and classroom interaction in general (Olney et al., 2025; Fernandez Herrero, 2024).

Furthermore, in secondary schools, challenges are rather different and consist of the complexity of the curriculum, the different needs of the learners, and the different preparedness of the teachers to work in the new dimension. These aspects make the effective application and scale-up of ITS difficult to achieve, particularly in poorly resourced schools under infrastructural and ethical limitations (Henkel et al., 2024; Wang et al., 2024).

The research thus aims to consider overall effects of ITS on teaching and learning processes in higher secondary schools. It will answer how ITS affects instructional strategies, students interactions, and learning outcomes, as well as contextual barriers and enablers that procure the adoption and effectiveness of ITS.

1.4 research aim

This study focuses on personalising a system to help determine how Intelligent Tutoring Systems (ITS) can influence the practice of teaching and outcomes of student learning in secondary schools.

1.5 research objectives

1. To evaluate and discuss currently known intelligent tutoring systems with regard to the architecture, their capabilities, and their proven efficiency in a variety of learning settings.

2. To compare the effects of the developed intelligent tutoring system on the secondary education learning outcomes and engagement.

3. To design and develop a customised Intelligent Tutoring System tailored to measure learning outcomes, student engagement, and instructional effectiveness within secondary school settings.

1.6 research questions

1. What are the typical architecture, characteristics and functionality of current intelligent tutoring systems, and how effective are they in various learning environments?

2. What are the learning outcomes and engagement effects of using the developed intelligent tutoring system in secondary education?

3. How does the personalised Intelligent Tutoring System, as a tool of teaching, influence pedagogy and teaching efficacy within the secondary school classrooms?

4. What conditions promote or inhibit successful introductions and uses of the tailored ITS in secondary education?

1.7 Justification of research

Indeed the role of Intelligent Tutoring Systems (ITS) within education is a fast-escalating trend that has promised to transform the conventional teaching and learning engagements via the use of personalised, adaptive instruction. Nonetheless, recent studies tend to focus on cognitive performance and overlook the possible consequences of the use of ITS, which may affect teaching practices and the general classroom environment, especially in secondary school (Olney et al., 2025; Fernandez Herrero, 2024).

In addition, much of the ITS solutions currently available have been built in high-resource contexts and can be of limited value in meeting the needs, challenges, and constraints presented in divergent or under-resourced secondary schools. The scope of these obstacles to sustainable ITS adoption is limited technological infrastructure, inconsistent teacher preparedness, socio-economic differences, and ethical difficulties; these issues include data privacy, emotional bias, etc.; it is confirmed that these factors present considerable challenges to adopting and maintaining the effectiveness of ITS (Henkel et al., 2024; Wang et al., 2024). These challenges need to be realised and overcome in order to level the playing field and also to maximise the positive potential of ITS.

The necessity to address this issue is therefore evident since there is an urgent need to design and develop a contextually relevant ITS that would support both the curricular objectives and the practices and the learners who are unique in secondary school education. The study will introduce a crucial part into the existing body of knowledge by undertaking an in-depth analysis in establishing the effects the ITS has on not only student learning outcomes but also the effectiveness of instruction.

Moreover, the exploration of contextual enablers and barriers will be beneficial to the policymakers, educators, and developers trying to implement ITS at scale responsibly and ethically. It is expected that the results will ultimately lead to the further development of the ITS concept as a useful means of reforming secondary education, individualising the learning process, enhancing teacher guidance and making education inclusive and based on equal opportunities.

.

1.8 review of related work

The Intelligent Tutoring Systems (ITS) have received a lot of interest as educational technologies aimed at offering students individualised adaptive instruction that resembles the tutor's one-on-one interactions. They are mainly orientated to enhance the efficiency and results of learning through reacting to the needs and progress of the students dynamically (Olney et al., 2025). The ITS-related research is conducted in a wide variety of educational contexts, and secondary education has both challenging and promising issues that may be discussed in detail.

How effective is ITS in student learning?

Some of the recent studies have shown that ITS has the potential of affecting the academic performance of the students positively, especially in the areas of STEM. An example is bestowed by Olney et al. (2025), who furnished the evidence regarding the capacity of ITS with functions that run on the use of expert human tutor models to augment the understanding and problem-solving capability of learners effectively. In a corresponding way, Henkel et al. (2024) also found that AI-based tutoring interventions led to scalable gains in math achievement among secondary students in Ghana, where educational gaps are likely to be filled by the use of ITS.

The degree of efficiency of ITS, however, varies due to different factors like the complexity of the subject matter, the design of the system and the engagement of the students. Fernandez-Herrero (2024) pointed out that affective elements that are built into ITS – emotion recognition and adaptive feedback – affect the persistence of motivation and enhance learning performances.

Effectiveness of the Teaching Practices

Although a lot of ITS studies concentrate on the performance of students, little research has been conducted on the impact of these systems on teaching practices. Wang et al. (2024) presented the model of human-AI cooperation, or, in other words, the model of ass controversially referred to as Tutor CoPilot, a system that seeks to assist teachers by delivering in-the-making expert evaluation and teaching assistance. These systems imply that ITS may be

used to supplement and improve the role of teachers rather than substituting them, which is why better and more individualised teaching is possible.

Nonetheless, the incorporation of ITS in current pedagogical models entails preparation and training of teachers, which are, however, underrated. The crucial point is how to develop ITS that is not only technologically advanced but also convenient enough and orientated to the instructional objectives of teachers (Fernandez-Herrero, 2024).

Issues and Morality

The adoption of ITS in secondary education does not come easy. Access and regular use of ITS may be impaired by an infrastructure shortage, particularly in resource-poor schools (Henkel et al., 2024). Moreover, the data privacy, bias in algorithms, and emotional profiling are ethical issues that require a thorough thought process that can result in adverse effects that are not intended (Fernandez-Herrero, 2024).

To have ITS be responsibly deployed in an equitable manner, literature states a need for additional empirical research that may also explore these contextual and ethical aspects in relation to effectiveness.

Options in Current Research

Although there are encouraging results, the research regarding the topic of the holistic assessment of the effect of ITS on teaching and learning processes in secondary school is limited. They tend to segregate the cognitive outcomes or the performance related to the technological areas yet fail to investigate the real dynamics of ITS with teachers, students and school settings. Such saturation reveals the necessity of studies that would comprise the creation of context-specific ITS, the assessment of their practical implementation, and the examination of aspects affecting their adoption and successful implementation.

1.9 methodology

This paper utilises a mixed-methods research design that involves both system development and empirical testing as a measurement of the effect of Intelligent Tutoring Systems (ITS) on learning and teaching achievements.

1.9.1 research methodology

The given research will take the form of a mixed-methods research design, assuming principles of both education and computer science as well as human-computer interaction (HCI) to measure the effects and technical characteristics of a blended Intelligent Tutoring System (ITS) created by the researcher. The ITS consists of two significant AI-based modules: a quiz generator that develops automated feedback and an augmented chatbot based on GPT-powered retrieval. They are aimed at assisting personal learning in secondary school, to a particular extent supporting adaptive assessments and content-focused question-answerings relying on uploaded documents (by students).

The design of the system is through an iterative system design based on the approach of software engineering. It has a modular structure that was constructed in Python, and the GPT APIs are what handle language generation functions. In document-based question answering, a retrieval-augmented generation (RAG) pipeline is to be deployed on LangChain or LlamaIndex and a vector store (e.g., FAISS or ChromaDB). To test single functions and parts like generating quizzes or matching feedback and parsing documents, unit testing will be used during the development stage. Simultaneously, the black-box testing will be used to determine how the system as a whole will behave when viewed by the user so that there is guaranteed optimisation that there are input/output interactions that behave as required.

The measurement of the system will be done in two general dimensions which include educational effectiveness and system performance. On the educational side, piloting of the system will be conducted in a classroom setting where the students of secondary schools will be used as the subjects, and they will be subdivided into the control and experimental groups. The measurement of learning outcomes will rely on pre- and post-tests; engagement and satisfaction on the part of the students will be assessed through surveys, and usability along with pedagogical compatibility will be accessed through teacher interviews. Statistical fractionalities like paired sample t-tests and ANOVA will be applied by using quantitative data in establishing significant differences in learning gains between groups.

Focusing on the computer science approach, the technical issues of the system and the experience of users are going to be reviewed with the help of diverse techniques. The chatbot's response time (latency) is going to be measured and studied to understand the actual utility of

the chatbots in classrooms and in real-time functions. Stress tests on the system will also be done to ensure measurement of scalability and uptime when several users are logged on to the system concurrently. HCI methods will be used, such as usability testing by getting the students and the teachers to interact with the system and the System Usability Scale (SUS) survey to measure the usability of the system in terms of interacting with the quiz interface and chatbot module. Also, short questionnaires or NASA-TLX will be used to evaluate the cognitive load that will be encountered after using the system, and errors and task completion will determine the degree to which users complete the tasks intuitively in order to meet their learning objectives in the system.

The behaviour of the GPT model in the ITS will be paid special attention to. As criteria of evaluation, the relevance of responses created and their correctness, detection of hallucination (i.e., when the model produces the wrong or fake answers), and the occurrence of the bias in the language or material will be used. Such groundedness of answers will be examined by (a) determining whether generating responses results in the accurate reproduction of information in the uploaded documents. Depending on the context, anytime domain experts or trained evaluators are available, model outputs will be examined and graded based on established rubrics.

The research study will follow the guidelines of informed consent and data privacy ethically. There will be a briefing of all the participants in the system (or their guardians at least, in the case of underage children) about the AI usage in the system, the character of the information gathered, and their right to withdraw. The encryption and storage security of uploaded documents will be provided, and the possible harmful and misleading responses to AI inputs will be observed. The chatbot can also be designed with a flagging mechanism where the user can complain about inappropriate or inaccurate replies.

1.10 research limitations

Although both in its design and scope, this piece of research is interdisciplinary, there are a few limitations to which it falls subject. In the first place, integrating a GPT-based model into the ITS brings the challenges related to the large language models, such as hallucination, inconsistency, and explainability. In spite of the fact that retrieval-augmented generation technologies are used to mitigate these concerns, the model can still come up with factually

inaccurate or deceptive answers that cannot be easily fact-checked by the students (Ji et al., 2023; OpenAI, 2023).

Second, findings are limited concerning generalisability since they depend on the educational and infrastructural setting where a system is used. It is planned to conduct the study in a small number of secondary schools, but some of them may have certain difficulties like the lack of stable internet connections, low device access, or lack of training of teachers in educational technologies, which affect the success and implementation of the ITS (Koukopoulos et al., 2022; UNESCO, 2021). Therefore, effects might not be transferred to more digitally sophisticated or under-equipped settings without adjustment.

Third, the system is created and evaluated as a prototype, and during its unit and black-box testing, it might fail to be as stable and scalable as production software. Real-time response, chatbot document grounding, and adaptive feedback are the features that might behave differently when they respond to loads or serve new user bases (Sharma et al., 2023). In this way, technical problems such as latency, issues on servers, or UI bugs could affect user experience and interaction at the evaluation stage.

Fourth, the pilot phase of several weeks of the short term (usually 6-8 weeks) does not allow observing long-term learning, behaviour change, or retention effects. Although pre- and post-tests are effective to measure the short-term academic gains, they might fail to measure more intricate learning ideas like the notion of metacognition or critical thinking (VanLehn, 2020). In addition, it is possible to say that quizzes can tap more into remembering rather than thinking.

The last but not least are the human factors which can affect the results of the study. When AI systems are unknown or doubted by some students or teachers, it may decrease their motivation to use the ITS in an active way. Although acts of resistance to the new technology, privacy concerns, or low digital literacy might decrease the usefulness of the tool despite proper orientation and training (Luckin et al., 2022). This inconsistency in user interactions poses a threat to the uneven quality and decipherability of data.

Chapter2: literature review

2.1 Foundations of Intelligent Tutoring Systems (ITS)

Intelligent Tutoring Systems (ITS) refer to the computer systems that are aimed at assisting the learner with personalised instructions and feedback, which simulates the flexible guidance through a human tutor (Wenger, 1987). The typical elements with regard to ITS awarded four core modules: the domain, which denotes the subject knowledge that is to be taught; the learner, which is a tracing of the current understanding and misconceptions of the learner; the pedagogical, which denotes the formulation of the instructions, deployments and feedback; and the user interface, which handles the interaction between the learner and the system (Anderson et al., 1995; Woolf, 2010). This modular structure allows ITS to change content and speed according to the needs of an individual learner.

ITS's historical progress includes the early rule-based systems in the 1970s and 1980s, which were handcrafted and rule-based, like SCHOLAR and GUIDON, aimed to simulate tutoring conversations by using handcrafted rules (Sleeman & Brown, 1982; VanLehn, 2011). Developments in computing power and algorithmic ability have now allowed ITS to deal with less structured, more complex tasks and offer more natural, more conversational interfaces.

Empirical studies have constantly proved that ITS has been effective in promoting learning outcomes. Meta-analyses establish that the students who employ ITS perform better when compared to students who undergo conventional instruction or computer-assisted instruction with no intelligent adaptation, and the standard effect sizes are approximately 0.7 standard deviations (VanLehn, 2011). ITS have proven effective regarding STEM subjects since they offer individual practice in solving problems and rapid feedback, favouring mastery learning (Pane et al., 2014; Nye et al., 2015). These observations provide the reason behind the increase in the use of ITS as scalable mechanisms in the personalisation of education. Understanding these foundational components and the evolution of ITS highlights their significance in transforming education by enabling tailored, data-driven learning experiences.

2.2 Intelligent Tutoring Systems with Large Language Models: Trends and Impacts in Secondary Education

After the COVID-19 pandemic, K-12 education experienced a revolutionary change towards digital and AI-based teaching. In the face of lockdowns, schools quickly turned to online platforms and learning software, and most of these solutions are now still utilised even as classes moved back into the classroom. Specifically, intelligent tutoring systems, i.e., computer systems delivering immediate and individual feedback, experienced the revival as remote learning necessitated the increased number of adaptive learning supports. ITS were well-known to be able to enhance learning through step-by-step instructions; however, recent developments in artificial intelligence, notably large language models (LLMs), such as GPT-3 and GPT-4, are changing the role of ITS. The refined ITS can take advantage of generative AI to bring in a new form of simulation of human tutors. To take an example, one systematic review of the ITS in K-12 with the aid of AI demonstrated an overall positive effect on student learning, but such positive effects are typically small compared to conventional instruction methods and require further investigation in detail. AI tutors in secondary schooling offer to individualise learning at scale, adapting to an individual student's improvement in the moment and leaving educators unencumbered to work on more advanced levels of direction.

These trends are add-ons to larger supercharged trends as a result of the pandemic. Since school disruptions forced teachers and schools to incorporate technology, leading to over 80 per cent of secondary students in the U.S. getting personal learning devices and accessing assignments through platforms such as Google Classroom. Customised learning applications in education are now widespread (e.g., math apps (Zearn, Lexia, etc.), literacy apps, etc.). Notably, teachers note that they are using small-group or individualised instruction for longer periods of class time than whole-class lectures, even with the use of digital tutors or practice sessions. In this context, advanced ITS can be seen as one facet of an adaptive learning ecosystem. The surge in blended and distance learning has created both a need and an opportunity for AI-based tutoring: students need extra support when teachers cannot always be available, and LLM-based tutors can step in with instant, tailored help. As one writer observed, AI tutors can give “immediate, personalised feedback” at scale, potentially offering insights into each student’s thinking that teachers can then use.

2.3 Modern AI in ITS: GPT and Retrieval-Augmented Systems

The development of large language models (LLMs) recently, especially GPT-3.5 and GPT-4 at OpenAI, has revolutionised intelligent tutoring systems (ITS). Such models can generate and understand natural language with a high degree of accuracy and allow completing tutor science complex tasks, e.g., involve personalised question generation, adaptive feedback, and conversational dialogue with learners (Brown et al., 2020; OpenAI, 2023). In comparison to the conventional ITS, which are highly based on rule-based or scripted systems, the GPT-based systems have the capabilities of dynamically producing different and diverse educational content and thus helping to increase the engagement of the learners and respond to the different needs of the diverse students in real time (Kumar et al., 2023).

Among the interesting improvements to pure LLMs, there is the Retrieval-Augmented Generation (a.k.a. RAG), a combination of both LLMs with the external knowledge retrieval mechanism. In response generation, RAG models retrieve pertinent documents or facts within a selected database or corpus, therefore basing it on factual material (Lewis et al., 2020). This prevents hallucinations, the propensity of LLMs to generate reasonable yet erroneous or counterfeit information, since the responses are based on evidence found and pull on reality and confidence in the answers among learners (Izacard & Grave, 2021).

GPT-based systems are more flexible and scalable compared with traditional rule-based ITS. Manual authoring of domain knowledge, pedagogical rules, and dialogue trees is needed in the design of the rule-based ITS that impedes adaptability and raises the cost of maintenance (VanLehn, 2011). GPT models also, however, are subject-agnostic and can be applied to open-ended questions without having to script them exhaustively. However, the issue of explainability persists because the way the GPT makes the decisions lacks transparency and is not as clear as in the case of the rule-based logic, which is problematic to educators who must finally comprehend and regulate tutoring tactics (Sleeman & Brown, 1982; Binns et al., 2021). Even so, there is overall hope in the fusion of GPT with RAG and formalised feedback systems, as they offer a strong medium in terms of adaptivity, factual accuracy, and teaching efficacy. Given that the proposed system leverages GPT with retrieval-augmented quiz generation and feedback, these recent advancements form a robust technical foundation that addresses common limitations of earlier ITS designs while enhancing learner interaction.

2.4 Quiz Generation and Feedback in ITS

The ability of modern Intelligent Tutoring Systems (ITS) to construct automated quizzes is an important characteristic that allows providing close-to-real-time formative assessment according to the learning outcomes. Lots of systems employ frameworks, including Bloom's taxonomy, to construct questions that address various levels of cognition, spanning simple recall all the way to higher-order thinking (evaluating and analysis) (Anderson & Krathwohl 2001; Mitkov & Ha 2021). This compatibility guarantees that quizzes not only measure the knowledge, but that they will help learn more.

Studies have reported that adaptive feedback, which involves immediate, customised responses to the performance of a student, is one of the common benefits of improving the learning of students through correction of error and reinforcement of concepts (Shute, 2008). Dynamic ITS produce quizzes based on learner models, and this allows adopting different levels of complexity and focus on content selected to enhance interaction and achievement (Drachsler et al., 2015; Chen et al., 2022). As an example, automated feedback with an explanation of the erroneous results and recommendations of the sources increases the results of the formative assessment and contributes to the mastery learning (Narciss, 2013). These functions render quiz creation and feedback to be essential in meeting the needs of ITS, which seeks to offer personal learning pathways.

2.5 Retrieval-Augmented Chatbots Education

It has been established that chatbots are useful assets in stimulating self-directed learning as a way of offering simply available, interactive assistance beyond the framework of usual classrooms (Fadhil, 2018). Implementation of retrieval-augmented generation (RAG) techniques enables chatbots to base their answers on documents uploaded by the users or embodied knowledge bases, enhancing the accuracy and suitability of the answers with regard to facts (Lewis et al., 2020; Izacard & Grave, 2021). These text-related chatbots will guide the learner through difficult material by responding to questions using specific texts and therefore enable one-to-one tutoring. However, these important factors are student trust and cognitive load. Research indicates that when the response of chatbots is incorrect or contradictory, this can have the effect of decreasing learner confidence and raising the mental burden, which may paralyse the learning (Baylor & Kim, 2004; Dzikovska et al., 2021). They can be addressed by

ensuring the openness of the origins of the knowledge base acquired by the chatbot and the means to check or ask questions to clarify the answers (Neupane et al., 2024). AI-powered tutors are under development, but one should remember accuracy, usability, and learner trust as the necessary variables concerning the subsequent implementation of any educational chatbot.

2.6 Teaching and Learning Effects

The search for AI-driven ITS has already started to have quantifiable impacts on the engagement of students, the learning process, and the classroom practice. Some research and reports draw particular attention to some of these effects:

Motivation and Engagement amongst Students.

In one randomised study, using high school students who were English learners, GPT-4-mediated exercises substituted the usual homework. The AI tutor provided immediate writing task feedback. Students in GPT-4 sessions claimed to be more engaged and supported in their tasks than their classmates without regular homework; they did more homework and more follow-up questions. The researchers noticed that there were great enhancements in the... the engagement of students, and students were quite satisfied and desired to stay with the AI tutor. Relatedly, survey research and interviews carried out in classrooms pretend that students are likely to realise that AI tutors are more engaging and active. Having direct feedback and someone to talk to, however, LLM-based ITS can simulate the conversation-like experience of learning, which is often favourable among the students. The director of learning science at Khan Academy reported that the AI tutor was able to keep students tasked using Socratic prompts, i.e., be like a human mentor who is constantly available to them. These qualitative accounts concur with the general findings: Potent LLM chatbots could only keep students engaged and facilitate the interaction since they have comprehensive knowledge and offer fast feedback. Concisely, the generative ITS are also likely to cause learning to be more vibrant, a characteristic which generally enhances persistence and motivation. Personalisation. One of the characteristics of ITS is individualisation of instruction, and in that regard, AI can significantly contribute to it. Tutors in the LLM programme will be able to adjust their language and answers to the level and experiences of a student. To illustrate, in Khanmigo the AI tutor not only is aware of the current lesson unit the student is supposed to study, but it is also aware of the skills already learnt by the student. It utilises it to personalise suggestions (“I notice you

did not get this question; why don't we go through it by dividing it up?"). On the same note, Duolingo has its Explain My Answer component that personalises the feedback and provides it in terms of correcting a particular mistake that the individual learner made. On-the-fly correction is also possible with LLMs: in case a student writes in a simpler way, the AI will be able to correct it using a similar tone; in case a student is advanced, the AI will be able to apply stronger vocabulary. Additionally, the nature of the multi-turn chat leads to the development of a kind of adaptivity: each of the answers the student provides shapes the following question. Unlike the one-size-fits-all drills, such AI tutors react to the individual learning path of a student. Literature indicates that such individualisation has the ability to reduce the learning gaps. As an example, an analysis of data on math tutoring in Austria revealed that poorer students gained most during school closures, using an ITS. That finding means that AI tutors can serve to particularly aid struggling learners, going after their individual misconceptions instead of rehearsing other points that they may have a better grasp of.

Learning products and knowledge.

Will this interaction and customisation result in real learning outcomes? Initial signs are promisingly encouraging. In the Austrian ITS, the total math performance improved during the COVID lockdowns, where the typical trend in standardised tests is a decline. The ITS revealed that students who used it to practise math tasks performed better this school year than last year, which means that the tutor may have kept or even trained the skills when the classroom settings were shuttered.

The above GPT-4 homework study in English language classes resulted in students showing, compared to control students with traditional homework, significant improvements in, among other things, grammar. Such improvements are not too significant but still meaningful, so it is possible to conclude that the explanations and practice provided by the AI tutor were not worse than those employed in other methods. Teachers and experts who participated in the interviews often state that AI tutors appear particularly efficient at eliminating semantic confusions. The Khanmigo designers point out the Socratic approach as an example: instead of providing the answer directly, the tutor chooses to ask the probing questions ("Why do you think the answer is x?" and "How did you go the last time?"). This would stimulate students to explain their reasoning in order to enhance understanding. Responses given by domain experts assessing LLM tutors also indicated that words of encouragement (such as 'good job' and 'move on to the next thing') also persevered students when they underwent hard topics. Automated feedback is

also more effective in honing learning: LLMs can point out certain mistakes in a student solution (e.g., an error in algebra) and explain it, similar to a hint by a human tutor.

Evaluation

A big opportunity is the grading and provision of feedback with the help of AI. Studies in the past revealed that it is reliable in student work evaluation with GPT-based systems. As an example, a GPT-4 prototype assigned free-text answers by incoming students with correct/incorrect labels and provided error-orientated feedback with no human intervention. Surprisingly, assessors deemed the AI feedback as easier to read and more accessible than that offered by human tutors. This would imply that their LLM tutors may be able to undertake most of the routine feedback work, allowing their teachers to focus on teaching concepts. In addition, certain ITS combine analytics of tracking progress. Evaluation of the complete history of a student's responses may allow an LLM+RAG system to predict future effectiveness or patterns of weakness, which would allow adaptive quizzes. In his/her classroom, educators can apply those insights in personalising lesson plans. As an example, there is a project that creates an AI that would observe real tutoring sessions and give the human tutor recommendations on what to do next in real time. The information collected by ITS (like which hinting works best and which doesn't) can adjust personalised curricula even more in time.

The roles of a teacher and practice.

Notably, what specialists emphasise is that AI tutors are the supplement to teachers, not a substitute. According to the U.S. Department of Education, human tutors are the best when it comes to motivation and social support, which are being unable to replicate in the AI environment completely yet. There are practical experiments pushing blended models: e.g., a teacher can give some new material in a lesson and then assign the AI tutor as homework or small groups, spending the classroom time to assist students in making sense of what the AI taught. To get ahead of any problems, early deployments (e.g., the Khanmigo pilot in classrooms) leave a teacher at hand when the students will engage with the AI. The hybrid solution utilises the fact that AI is available and still has human supervision. Some AI tools are also helpful to teachers in terms of planning the lessons and tracking the progress. According to survey data, it is still unclear as to whether the overall impact on teacher workload is positive or not, but the majority of educators believe they can facilitate some of the simplest professional

duties, such as answering routine enquiries or sorting homework assignments, with the help of AI-prompted technologies.

2.7 Evaluation Methods for ITS and AI in Education

The process of assessing intelligent tutoring systems (ITS) can be complex since it involves factoring in both the educational effectiveness and the software performance as well as the usability of the ITS itself (Greer, 2002; Moosavinasab, 2018). Research studies in educational contexts frequently involve field trials or controlled studies in which ITS difficulties are applied in genuine classroom settings, and learning gains and teaching effects are determined via means of the pre- and post-tests, control teams, and/or usage of statistics (Koedinger & Aleven, 2016; Moosavinasab et al., 2018). An example of this is that Corbett, Koedinger, and Anderson (2001) point out the significance of a rigorous design to say that given improvements are caused by interventions in ITS.

Usability and human-computer interaction (HCI) evaluations As a usability and human-computer interface (HCI) approach, assessment normally includes tasks by experts and users. Heuristic evaluations are examples of expert assessments that are useful in detecting interface and interaction design problems (Çetin & Şendurur, 2019). Quantitative values of learnability and satisfaction can be achieved in user-centred techniques such as the task completion rate, error detection and the System Usability Scale (SUS) (Bangor, Kortum, & Miller, 2008; Nielsen, 1994; Neupane et al., 2024).

The method of computer science assessment is also important; the unit and black-box tests are applied to test functionality and response time, and latency is measured to guarantee functionality in real time. The scalability is evaluated with the use of stress testing, and the quality with the help of the chatbots is measured using the quality metrics to cover the chatbot accuracy, hallucination rates, and relevance. Retrieval-augmented generation systems are usually evaluated by using specific frameworks, such as RAGAS, and standard metrics, such as BLEU or ROUGE (Es, James, Espinosa-Anke, & Schockaert, 2023; Neupane et al., 2024). Neupane et al. (2024), for example, evaluated a retrieval-augmented chatbot using both the RAGAS metric for accuracy and SUS for usability, reporting strong performance. Similarly, Antico, Giordano, and Ognibene (2024) performed qualitative usability testing on a student-facing chatbot and identified technical and reliability challenges that affect user experience.

Ultimately, effective ITS evaluation combines educational field trials measuring learning outcomes, usability assessments refining interface design, and rigorous system testing ensuring reliability. This integrated methodology aligns with best practices in educational research and software engineering (Greer, 2002; Çetin & Şendurur, 2019; Moosavinasab, 2018).

2.8 Future Directions, Challenges, and Opportunities

Looking into the future, ITS in secondary schools is in a state of further development. Technically, LLMs will keep expanding with magnitude and strength. GPT-4 or other models will probably introduce even greater insight and generation, which can even be multimodal (e.g., a student drawing a graph or talking instead of typing). Other scholars anticipate ITS that incorporate both LLMs and math engines, or science simulators, so the coach can work out equations on the fly or inspect diagrams. AR interfaces (augmented reality) may permit a chatster-like AI to be used to engage the students in virtual labs. In the meantime, there will be the advancement of data-driven personalisation: the more data we collect about the learners, the better AI tutors will be able to understand their model of the current state of knowledge. The first trend, which is exciting, is called the inferences the AI makes to create an open learner model (e.g., the student is exposed to it so that he/she may have a look at the grades, as it shows, e.g., the letter A was mastered but B was not). Students can also become involved in the process of learning through such transparency. Overall, we are looking forward to the time when ITS will be more normalised into the regular classroom set of tools, the way graphing calculators or language labs were at some point.

Meanwhile, it faces some serious obstacles. One of them is accuracy and trustworthiness. The nature of LLMs is that they are encouraged to hallucinate, i.e., confidently claim false knowledge, which can be exceptionally harmful when training. A student will be confused or misled by an AI tutor that tells him or her “2+2=5” or mis-explains something. This risk cannot be removed through careful prompting but can be reduced through RAG and careful content curation. Teachers are to be cautious: a human eye is an absolute necessity in order to detect AI mistakes. Moreover, LLMs usually have trouble with math and symbolic logic. According to the users, GPT-type tutors perform well in text problems and err in multi-step calculations. Integration of LLMs with domain-specific tools is being actively researched. Areas of ethical and equity matters are very important. More recent surveys caution that AI tutors can bring out or expand biases or a learning disparity unless handled with care. In that regard, e.g., when

training data fails to contain a variety of examples, an ITS will provide culturally biased explanations or fail to comprehend dialectic language. Academic honesty is the other aspect of concern: the availability of an AI tutor may lead students to overuse it, which may result in plagiarism or surface learning. Another problem is the privacy: ITS should ensure that it deals with sensitive data on students (academic data, personal data) in a responsible manner. The issue of data security and consent (particularly in the case of minors) comes up when large-scale implementation of AI tutors is deployed in schools.

On the bright side, the supporters raise equity and access opportunities. Quality one-on-one tutoring is rare and cost prohibitive; AI tutors will be able to personalise this attention to a much larger number of students at little cost. In low-resource school environments an ITS could be used to give a boost to the few teaching personnel. Other researchers have even pilot-tested some AI tutors in underfunded or rural schools where some students enjoy the additional tutoring support. Faced with pandemic learning loss, policymakers are wondering whether generative ITS can assist with the adjustment to allowing students to be, as one official put it, at least in a more near-term sense, met where they are by customising the remediation in a manner not facilitated by one-size-fits-all approaches.

At last, teaching methodologies will develop towards these devices. Educators will have to train on the use of AI tutors and information about how they can be incorporated into their plans and analyse the analytics. The role of the teacher can change more towards facilitators than the active lecturer. Schools can develop AI policy (e.g., when and under which conditions students can use chatbots). When the technology reaches its maturity, the best practices will arise, such as a hybrid of AI and human feedback, curriculums specific to AI-aided learning, etc. Researchers also insist on continuous assessment: as a panel of experts has observed, we have to evaluate the impacts of AI systems on learning continuously, with our alert sweat on the undesired impacts.

2.9 review of existing systems

Each article was viewed in the context of supportive or not supportive literature that gives a mixed view of the effectiveness of these AI-led strategies. Also, new directions and certain gaps in research are established in each subject. The main emerging theme to be addressed by the present research, as well as the general research gap on the topic of the research, is covered in a separate section.

2.9.1 (Hang et al., 2024), MCQGen: A LLM-based MCQ Generator on Personalised Learning

To start this work, the researcher read the article (Hang et al., 2024), presenting the framework MCQGen, automating the multiple-choice questions generation process to enable personalised learning. The system uses GPT-4 that is combined with retrieval-augmented generation and advanced prompt engineering approaches. To begin with, a comprehensive database is composed of domain-specific knowledge given by instructors and created MCQs by students. The language model then uses this information to come up with interesting and challenging questions that will be of context. The method works based on a two-phased prompt engineering approach: first, the chain-of-thought mode is utilised to come up with the candidate questions, and after that, a self-improvement loop is used to ensure assessment quality and difficulty are improved with the loop iteratively over several rounds. Reviews (measured in terms of grammatical fluidity, answerability, diversity, complexity, and relevance) show that the MCQGen system is capable of not only decreasing the amount of time and skill necessary to build a quiz but also the amount of time and skill needed to grade a quiz because the system creates FACTs.

Observations:
Hang et al. (2024) noted that while the framework shows promise in producing adaptive and high-quality MCQs, there remain challenges in achieving sufficient question diversity and in handling complex numerical problems. This gap highlights an emerging theme in intelligent learning systems: the need to further refine AI-driven assessment tools to better accommodate a wider range of subjects and cognitive challenges.

2.9.2 Hirulkar and Athawale (2024), Quiz Master AI: An Interactive Machine Learning-Based Quiz Generator

Having started this exploration, the researcher analysed the article of Hirulkar and Athawale (2024), where the authors present a system that automates the conversion of short written text into multiple-choice questions under the name of Quiz Master AI. The methodology begins with the passing of a given text with an NLP model to obtain key concepts that shall be used within a machine learning-based question generation module to produce a set of various and contextually correct MCQs, dynamically challenging a user based on real-time performance.

Finally, this system is integrated into the gamified environment, implying the development of a group-based platform featuring scoring, leaderboards, and a levelling-up strategy to foster competition and cooperation.

Observations

Hirulkar and Athawale (2024) mention that while the introduction of Quiz Master AI really helps simplify quiz construction and make them more engaging through customisation and interactivity, the system can be made better in aspects such as the quality of questions, which should be given proper attention irrespective of topics, as well as maximising the customisation level.

2.9.3 (Dnyandev Desai, 2025): AI-Assisted Learning Versus Gamified Testing: A Data-Driven Approach to Educational Enhancement

Authors proposed a comparative framework for testing and comparing AI-assisted test generation against gamified testing as a way of advancing learning. In their approach, they compared Quick AI kinds of systems, which automate question generation and analyse its performance, against platforms like Kahoot!, which are much more interactive and game-related. The attempt by Chavan et al. (2025) was to assess the factors of learning enhancement, reduction of response time, and AI question accuracy, as prior studies conducted by Hamari et al. (2014) and Deterding et al. (2011) pointed out positive implications of gamification. This framework was split into two main groups: quantitative performance analysis and qualitative engagement assessment.

Quantitative Performance Analysis

This stage focused on measuring learning outcomes by the difference shown in pre-test and post-test scores and response times. Data collection consisted of logging learning scores and response times, while data engineering was concerned with computations for performance improvements: Quick AI increased learning by 75.5% while reaching an AI question accuracy of 85%, and Kahoot! was able to present a faster reduction in response time by around 50%.

Qualitative Engagement Assessment

In this stage, user engagement was assessed, examining features like live leaderboards, real-time feedback, and competition. Then, surveys and statistical analyses (paired t-test and Mann-Whitney U test) were applied, comparing the impacts of each platform on learner satisfaction and engagement in general.

Observations

Chavan et al. observed in 2025 that Quick AI offers stronger automation of quiz generation and enhanced structured learning, while Kahoot! enters with better engagement through its gamified interface. The study conceives that AI-based quiz systems could be better for large-scale automated testing, while gamified platforms are preferable for real-time interaction. The indications point to implementing integration of AI automation and gamification for further enhanced educational benefits.

2.9.4 Chimezie (2024), Leveraging Retrieval-Augmented Generation in Large Language Models for Effective Learning: A Data Structures & Algorithms Learning Assistant

The researcher went through Chimezie's work (2024), which explores the design and development of a learning assistant for Data Structures and Algorithms (DSA) by using Retrieval-Augmented Generation-Large Language Models (RAG-LLMs). The work highlights that students find it difficult to learn DSA because of the abstractness of these concepts and inherent issues in traditional LLMs, such as biases and hallucinations. For example, the rationale behind the set-up employs the RAG with LLMs to improve the delivery of information that is accurate and relevant to students."

A quantitative approach was applied to research design, along with a crossover experiment involving ten computer science students from Ashesi University. Participants interacted with both the RAG-based learning assistant and the conventional LLM (ChatGPT) to evaluate their strength in supporting DSA learning. The main findings show that the RAG system is deemed as significantly more useful than ChatGPT for impacting actual performance in the students' academic outcomes.

Observations

Chimezie (2024) has mentioned that the RAG-based learning assistant has a fair potential to improve learning experiences even while posing current problems with real usage because of usability constraints. This gap sheds light on a new theme in educational technology: continuou enhancement of AI-based learning tools should be prioritised for higher user engagement while solving complications in educational content delivery. Furthermore, it becomes apparent from the study that the selection of more advanced LLMs would ensure better use of RAG, thus indicating a space for future studies on adaptive and user-friendly design.

2.9.5 (Cabezas et al., 2024) Integrating a LLaMa-based Chatbot with Augmented Retrieval Generation as a Complementary Educational Tool for High School and College Students

Through the use of the Pinecone API to retrieve information, the chatbot can respond to a student's This article reviews earlier work by Cabezas et al. (2024) that represents a novel approach to ameliorating educational experiences with the help of an LLaMa-7B chatbot and state-of-the-art RAG techniques. The authors give design strategies for a system that personalises learning for students in high school and college through a structured mathematical database enriched with audiovisual materials. questions by searching for answers using cosine similarity.

The paper describes the methodology used in the design process, first with the implementation of the LLaMa model and the exploration of vector databases. They emphasise embedding generation by applying the all-MiniLM-L6-v2 model to make an effective basis for having their chatbot. A bulk of the work has been around creating a dataset targeted at commonly referred-to educative content so as to have relevant data available to the users.

Observations

Cabezas et al. (2024) have observed how their chatbot can revolutionise education and raise student engagement towards fuller knowledge uptake. The phase of data collection proved challenging in many aspects, such as sourcing the apt audiovisual material and translating the mathematical concepts in all their complicated beauty. They mention that the system, albeit quite promising, somewhat suffers in its realisation due to available computational resources, suggesting that future iterations could benefit from more powerful models and broader internet access for real-time information retrieval. This underscores a critical point in the evolution of

educational technology: the need for continuous adaptation and enhancement of AI-driven tools to meet diverse learning needs effectively.

2.10 Benefits of the Proposed Intelligent Tutoring System

The blended ITS with auto quiz generation, adaptive feedback, and GPT-powered retrieval-augmented chatbot for secondary education provides some significant benefits.

The first benefit of the ITS is its ability to construct quizzes dynamically and programme the coding alignment with learning objectives and Bloom's Taxonomy, offering unique opportunities for formative assessment with that focus on practice and learner engagement at altered cognitive levels covering understandings and mastery (Anderson & Krathwohl, 2001; Chen et al., 2022). Also, it gives feedback that is immediately adaptive, so students can promptly recognise and correct their misconceptions and work more efficiently on learning while being motivated (Shute, 2008).

Secondly, the retrieval-augmented chatbot accepts user-uploaded documents for on-demand tutoring based on the context concerning the specific needs of each individual student. This feature assists in self-driven learning and content comprehension by basing answers on particular materials, thereby ensuring accurate rather than merely plausible answers and building trust with learners (Lewis et al., 2020; Izacard & Grave, 2021). Under such a structure, the chatbot converses with the student, inviting engagement and active inquiry outside of regular class hours (Fadhil, 2018).

Thirdly, GPT-powered ITS provides an extent of scalability and flexibility that rule-based ITSs can never aspire to match. They can answer open-ended questions on arbitrary topics without requiring exhaustive manual scripting, which immensely reduces time in development as well as in maintenance costs (VanLehn, 2011; Kumar et al., 2023). This liberty has also allowed the ITS to be more responsive to changes in curriculum and student needs.

Finally, by combining educational best practices with advanced AI techniques, the system can provide a holistic learning experience that supports cognitive, metacognitive, and motivational aspects of education (Narciss, 2013; Woolf, 2010). This comprehensive approach promises to enhance both teaching effectiveness and student outcomes in secondary education.

2.11 Conclusion of Literature Review

The reviewed literature makes clear that ITS have undergone profound evolution – from the initial simple rule-based platforms to modern AI platforms employing bigger language models such as GPT. The seminal works have shown ITS to be effective tools for personalised learning, adapting teaching and feedback to meet the needs of individual students, thereby enhancing learning outcomes in several contexts. Recent developments in the field of automated quiz generation and adaptive feedback mechanisms allow these systems to promote deeper engagement and mastery of the material while addressing different cognitive skills targeted within Bloom's Taxonomy.

Furthermore, the future of ITS could shift toward building retrieval-augmented chatbots that provide context-aware, document-grounded tutoring supporting self-direction and comprehension of content. Some challenges remain concerning accuracy, trustworthiness, and cognitive overload when interacting with an AI-driven tutor. Attempts to resolve these thorny problems would yield the most significant educational benefits of such systems.

It is clear, thus, from the very start, that the research being undergone to design and test a hybrid ITS made up of a quiz generator, adaptive feedback, and a GPT-based retrieval-augmented chatbot falls well within the current technological and pedagogical trends. The focus is on capitalising on aspects of present-day AI to build an evidently scalable, rich, flexible, and focal-point-interactive tutoring paradigm geared toward secondary schooling. The identified research gaps within the extant literature, especially with respect to the necessary holistic evaluation by means of educational and computer science methodologies, emphasise the rationale for this research thrust and also shed light on its potential significance.

Chapter 3: Methodology

3.1 Introduction

This chapter provides the research design and the methodological approach that would be used in the design and assessment of the proposed Intelligent Tutoring System (ITS). Both an innovative blended ITS, which includes automatic generation of quizzes, adaptive feedback, and GPT-based retrieval-augmented chatbots, and an assessment of the teacher and learning experience in secondary school teaching were of primary importance to this investigation. Therefore, the research method is set up multiprocessally, as a mixed-methods one. This strategy combines the process of system development and empirical assessment to represent a full picture of the educational performance of the system and its technical functionality. The chapter presents the elements of the research design, the system development process, the selection of the participants, the methods of data collection, the criteria of data evaluation, the methods of data analysis, the ethical consideration of the study and the limitations of the research.

3.1 Research Design

The present research proposes a mixed-method research design in which the guiding research design is complemented by employing mixed-method research to facilitate system development as guided by design science research and empirically evaluate the effects of the proposed Intelligent Tutoring System (ITS) on teaching and learning in secondary education. Design science studies can be used since it focuses on the designing and development of innovative information technology products to resolve challenges that exist in the real world through repetitive refinements (Hevner et al., 2004).

The study is referred to in two primary stages:

1. Development Phase: The designed and implemented blended ITS will be comprised of automated quiz generation, following Bloom's taxonomy, adaptive feedback systems, and retrieval-augmented chatbots using a GPT model. These agile development practices will inform the process of iteration, prototyping and improvement through formative testing.

2. Evaluation Phase: The educational attractiveness and technological performance of the system will be discussed at the level of secondary school education in an empirical form. The learning outcomes as well as system usability and engagement will be measured quantitatively through pre- and post-tests, system log analysis, and surveys of the people using the system. Immunological approaches, such as interviews and focus groups with students and teachers, will allow the study in detail of the experience and perception of users.

This combination of methods presents the opportunity of triangulation of data to confirm the findings and provides the comprehensive picture of the pedagogical and technical effects of the ITS. There is also a need to combine the computer science assessment measures (e.g., response time, accuracy) with the teaching and learning measures so that a multifactorial measure of utility and success of the system takes place.

3.2 Requirements Analysis

The requirements analysis phase identifies the functional and non-functional needs necessary for developing an effective blended Intelligent Tutoring System (ITS) tailored to secondary education. This phase ensures that the system design aligns with pedagogical goals, user expectations, and technical constraints.

Functional Requirements:

- **Automated Quiz Generation:** The system must generate quizzes dynamically based on curriculum topics, aligned with Bloom's Taxonomy cognitive levels, covering knowledge, comprehension, application, analysis, synthesis, and evaluation.
- **Adaptive Feedback:** The ITS should provide personalised, immediate feedback based on student responses to support formative assessment and guide learning progress.
- **Retrieval-Augmented Chatbot:** The chatbot must accept user-uploaded documents and generate accurate, contextually relevant responses using a GPT-based retrieval-augmented generation approach.
- **Performance Logging:** Capture interaction data such as quiz attempts, response accuracy, chatbot queries, and response times for evaluation purposes.

Non-Functional Requirements:

- **Usability:** The system should have an intuitive interface accessible on common devices (desktop, tablets, smartphones) to facilitate easy adoption by students and teachers.
- **Scalability:** The ITS must support concurrent users without performance degradation, allowing for potential school-wide deployment.
- **Reliability and Accuracy:** The chatbot's responses and quiz generation must maintain high accuracy to build user trust and ensure effective learning.
- **Security and Privacy:** Protect user data, especially student information and uploaded documents, following relevant data protection regulations.
- **RESPONSE TIME:** System components, particularly quiz generation and chatbot replies, should deliver responses within acceptable timeframes (e.g., under 3 seconds) to maintain user engagement.

Stakeholder Input:

Requirements were gathered through consultations with secondary education teachers and IT specialists to ensure the system addresses real classroom challenges and technical feasibility. Their feedback emphasised the need for contextualised feedback, flexibility in content, and ease of integration with existing teaching workflows.

This comprehensive requirements analysis guides the subsequent design and development, ensuring that the ITS meets both educational and technical expectations for improving teaching and learning outcomes.

3.3 Tools Used

The development and evaluation of the proposed Intelligent Tutoring System (ITS) will utilise a combination of software frameworks, programming languages, AI models, and evaluation tools to ensure a robust and scalable solution:

Programming Languages:

- **Python:** For backend development, AI integration, and data processing due to its rich ecosystem of AI and NLP libraries.
- **JavaScript (React or Vue.js):** For building a responsive and interactive user interface.

AI Models and Frameworks:

- **OpenAI GPT API (GPT-4 or GPT-3.5):** To power the retrieval-augmented chatbot and quiz generation through natural language understanding and generation capabilities.
- **Retrieval-Augmented Generation (RAG) Frameworks:** Tools like FAISS or Elastic Search for efficient document indexing and retrieval to support chatbot responses grounded in user-uploaded documents.

Development Frameworks and Libraries:

- **Flask or Django:** Python web frameworks for backend API development.
- **TensorFlow/PyTorch:** For any custom AI or machine learning components, if needed.

Database Management Systems:

- **PostgreSQL/MySQL:** To store user data, quiz content, interaction logs, and system metadata securely.

Testing and Evaluation Tools:

- **JMeter or Locust:** For performance testing, including response time and load testing.
- **Google Forms/SurveyMonkey:** For collecting user feedback via questionnaires.
- **Statistical Software (e.g., SPSS, R, or Python libraries like Pandas and SciPy):** For quantitative data analysis.

Version Control and Collaboration:

- **Git and GitHub/GitLab:** For source code management and collaborative development.

These tools collectively support the end-to-end process of system development, deployment, and rigorous evaluation necessary for the research objectives.

3.4 System Development

Designing the proposed blended Intelligent Tutoring System (ITS) is based on the design science research (DSR) approach and also adopts the Agile software development approach. This also makes it captivate iterative design, frequent feedback and constant enhancement in the project. The framework can be divided into three main modules: automated quizzes and adaptive feedback, retrieval-augmented chatbots based on the GPT model.

3.4.1 Development Phases

The developing process can be separated into the following main steps:

1. **Requirements Gathering:** Various system requirements were defined on the basis of a needs analysis based on educators and students (see Section 3.2).
2. **System Architecture Design:** A modular structure was applied in the development of the architecture of the system:

- Frontend Layer: Implemented in React.js, it has the purpose of interacting with a user, taking quizzes, getting access to a chatbot, and seeing feedback.
- Backend Layer: Python (Flask or Django) based, with the role of API, processing logic, and communicating with AI services.
- Database Layer: Snaps user data, learning content and similar logs employing PostgreSQL or MySQL.
- AI Layer: will interface with GPT (using the OpenAI API) to produce quizzes, explanations and chatbot responses. Document-grounded conversation is made possible by a retrieval mechanism (e.g., FAISS or ElasticSearch).

3. **Module to Generate Quiz:**

- Combines Bloom's taxonomy to categorise questions at cognitive levels (remember, understand, apply, etc.).
- Uses GPT to come up with and change the questions dependent on the topic chosen and difficulty selected.

4. **Adaptive Feedback Module:**

- Provides instant feedback upon every quiz experience.
- Feedback is calculated on predetermined logic and GPT prompts to give explanations on how to answer and what answers are wrong or right.

5. **Chatbot Module:**

- Enables the students to post learning content (e.g., PDF, notes).
- Bases its search on a retrieval-augmented generation (RAG) pipeline, which searches the set of uploaded content and integrates it effectively into GPT responses, giving precise, text-based clarification.

6. **Refinement and Testing:**

- The unit and integration testing make sure that all the modules work properly.
- The user testing among students and teachers gives the information regarding usability and quality of instructions.
- The performance testing is used to test the scalability and responsiveness of systems.

7. **Deployment:**

- The system is served via a cloud platform (e.g., AWS, Azure, or Heroku) to interact centrally over the web.
- Security features are applied to safeguard the data of the users and adherence to ethical standards.

3.4.2 Human-Centred Design Considerations

The design is based on HCI principles, as well as a simple, clear, and accessible interface and ease of usage. Particularly, it regards:

- Easy navigation
- Low mental workload
- Looks as feedback stimuli
- Mobile responsiveness to enable accessibility in different classroom situations

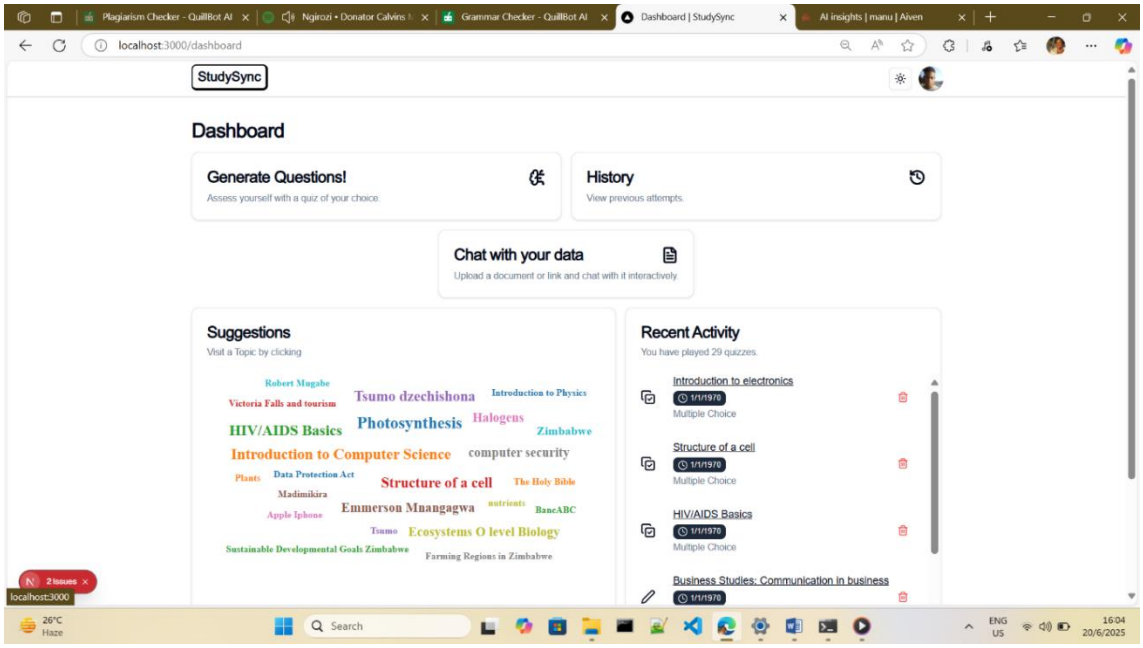


Figure 1shows dashboard

3.5 System Architecture

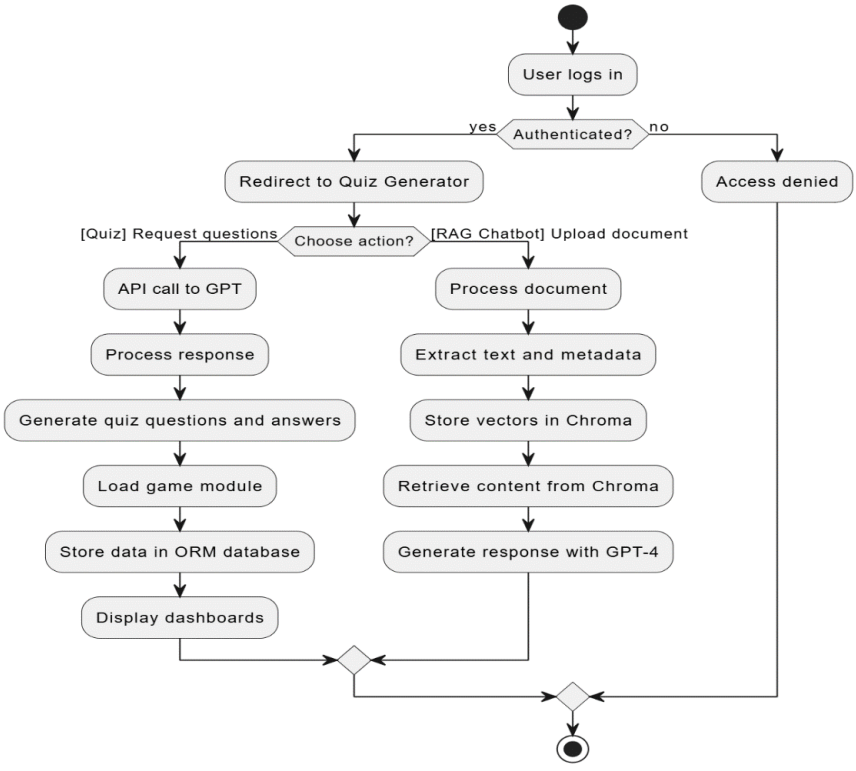


Figure 2 Workflow Diagram

The quiz generator uses GPT-4 that is fine-tuned on educational settings. This version of GPT-4 is specialised, creating dynamical, context-based quiz questions and answers, adapted to the user needs in terms of learning. A built-in game extension improves the interactive process, controlling the sequences of loading and timing, and an ORM database keeps user performance information and quiz history safely. Then, this information is represented on specific dashboards providing current scores and correction rates and the other one offering personal recommendations depending upon user enquiries. This elaborate system guarantees that users get instant results and instructions to enhance their knowledge and ability.

To supplement the quiz generator, there is the RAG chatbot that is designed to respond exactly and contextually. This chatbot can use ChromaDB as a vector-based database that aims at semantic search. ChromaDB stores texts in the form of high-dimensional vectors, and the system is used to efficiently compare and match the semantic content of the user queries and the information being stored, which in turn gives very relevant retrievals. After the most relevant material is revealed, GPT-4 takes action as the generator to develop logical and contextually correct responses. LangChain is the key that coordinates the whole process; it organises the vector database and the language model in one smooth flow to avoid the retrieval and the generation stages functioning independently.

Lastly, the user experience is increased with consideration of an interface behind it. To interact with the chatbot, Streamlit is used to design the interactive dialogue in a Python environment that can be quickly prototyped and get feedback in real time. But on the other hand, the quiz section will be constructed with React as a frontend, which is coupled with Tailwind CSS and Shadcn to provide a modern, responsive and aesthetically appealing interface. Such a design of interface guarantees that users can easily alternate quiz sessions and interaction with the chatbot, so the process of learning can be intuitive and fun.

3.6 Implementation Details

The system puts into operation a set of diverse powerful AI tools, efficient data management solutions, and integration techniques to realise the real-time interaction between the quiz generator and the chatbot. Python, LangChain, and ChromaDB are used in building the

backend along with REST APIs to unite the different components into a coherent view, allowing for a smooth and dynamic flow of data between the chatbot and the quiz generator.

TOOLS: PYTHON, LANGCHAIN, AND CHROMADB

The backend of the system is primarily being developed in Python. Being highly multipurpose, this language supports AI, NLP, and machine learning to an extensive degree. Python, in essence, is the foundation stone for the quiz generator living along with the RAG-powered chatbot, thus allowing the system to analyse user queries and generate pertinent responses according to the user. LangChain, primarily a framework to simplify interactions of LLMs with data external to them, happens to be one of the retrieval mechanisms used by this chatbot. In brief, LangChain allows the chatbot to query the stored knowledge, retrieve the relevant context, and arrange the pipeline for generating the response. LangChain manages the interaction between the retriever based on ChromaDB and the generator based on GPT-4 to fetch relevant content upon user queries and generate meaningful responses in this pipeline.

The system uses a vector database optimised for the storage and retrieval of text embeddings called ChromaDB to provide high-speed and accurate semantic search. After upload, the document is converted into vector representations through the use of Sentence-BERT embeddings. The vectors numerically express the text's meaning so that ChromaDB can match user queries to the most relevant parts of the uploaded document in an efficient manner. This retrieval ensures that transcript responses are not generic but are imperative from the document contents.

INTEGRATION: REST APIs FOR REAL-TIME DATA FLOW

RESTful APIs integrate the systems to ensure that the quiz generator works smoothly with the chatbot in real time. An API represents instructions that define how to communicate with HTTP requests between different components of a system.

Quiz Generator API: When a user requests a quiz via a React-based frontend, the request is then sent to the backend using REST APIs. Namely, those requests are handled by GPT-4 to generate quiz questions and answers in a structured string format, which is further converted into JSON in order to be compatible with the frontend JavaScript-based user interface, which then dynamically displays the processed quiz for users to interact with in real time.

Chatbot API: REST APIs are used to handle user interactions with the chatbot as well. Whenever a user asks something, a request is sent to the backend API, where LangChain will retrieve relevant information from ChromaDB before passing it over to GPT-4. The API allows all these steps to occur without disruption, facilitating fast document retrieval and response generation; hence, the chatbot can render swift, context-aware answers on the basis of the document uploaded by the user.

3.7 Participants and Setting

Would this in situ evaluation of the proposed ITS be conducted in the educational setting among secondary school students and teachers? Subjects inside the respective schools shall be chosen, wherein the subject taught by the ITS is compatible with the school curriculum. Participation shall be fully voluntary; the teachers and students involved will be provided with a brief presentation on the system, including its objectives and use. Hence, this kind of contextual evaluation shall be able to assess the system from the technical perspective as well as make judgements regarding its relevance and efficacy from the pedagogical perspective in the classroom setting (Kukulska-Hulme et al., 2021).

3.8 Data Collection Methods

In order for the ITS to be fully evaluated, a series of data collection methods would be used. Students will be asked to take some pre-tests and post-tests to measure the amount of knowledge that they have been able to learn and master while working with the system. System loggings and data analytics will record performance indicators such as quiz scores, time spent on a set task, and interaction with the chatbot, thereby offering an objective measurement for engagement and learning behaviour. Complementing these will be the application of questionnaires and surveys, which aim to collect data from users about system usability, motivation, and satisfaction. Semi-structured interviews and focus groups with teachers and students will also be conducted to obtain qualitative data that will further clarify user experience, perceived effectiveness, and possible improvements (Roscoe et al., 2022).

3.9 Evaluation Methods

To assess the effectiveness of the quiz generator and chatbot, thorough evaluation of both quantitative metrics and qualitative feedback shall be carried out.

1. Quantitative Metrics:

- **Quiz Accuracy (F1-Score):** F1-score takes into account two factors, precision (correctness of positive prediction) and recall (ability to detect all instances that matter). It is the harmonic mean of these two:

  In a situation like ours, quiz accuracy becomes its strength; the number of correct answers produced by the system against expert-validated answers is a stronger measure of evaluating the system.

- **Chatbot Response Latency:** Determines how quickly the chatbot can respond to the user's queries. It is the harmonic mean of these two: In a situation like ours, quiz accuracy becomes its strength; the number of correct answers produced by the system against expert-validated answers is a stronger measure of evaluating the system.

2. Qualitative Feedback:

User Survey: With a sample size of about 30, the user survey could explore engagement and utility. Feedback on satisfaction, usefulness, and overall experience would serve to further entertain areas in which it could be improved.

Instructor Review: Educationists will assess the generated quizzes for consideration of and alignment to objectives and will mention some opportunities for improvement.

When quantitative and qualitative means are combined, a stellar evaluation could be realised in the area of performance recognition and others for future improvements.

3.10 Ethical Considerations

Bias and Fairness

Since AI algorithms are built on training data, biases may be reproduced against certain user groups (Chinta et al., 2024). An AI quiz generator may, for example, give preference to subjects or language types commonly found in the training data over its own users, who come from diverse backgrounds. To this end, diverse and representative datasets should be used, while the output of the AI needs to be constantly monitored for unintentional biases.

Accountability

When problems or harm arise from the use of AI systems, another difficult ethical issue is who is to be held responsible. Defining clearly defined accountability policies ensures that the developer or educator can resolve and fix the issues in a timely manner to keep trust within AI educational tools. The addressing of these ethical concerns thus remains intact when responsibly implementing AI into education, with the ultimate goal of ensuring that technology augments the learning process rather than detracts from its ethical standards.

3.11 Limitations

Several constraints were posed in this study. First, there might be limited areas for deployment, with only a handful of schools involved, thus limiting the generalisability of the results. Differences in students' prior knowledge and technological familiarity might also influence their learning outcome and system use. The type of teaching and classroom environment will somewhat also affect how well the ITS is integrated into the existing curricula. These limitations will be discussed in the concluding analysis, with recommendations on wider deployment and further studies.

3.12 Conclusion

This chapter has described the methodologies and frameworks that guided the development of our educational tools – the RAG Chatbot alongside the Quiz Generator. The usage of the Agile development methodology allowed us to proceed iteratively and adaptively, facilitating the evolution of both tools with the needs of users and what was technologically possible at any

time. Our research methodology, in the form of a narrative literature review, allowed us to understand thoroughly what educational technologies currently existed and to make informed decisions about the design and functionality of our tools. This step was necessary, as it demonstrated an awareness of and foundation for choosing our approach through a critical comparison of alternative methodologies – it thus guarantees some rigour along with responsiveness to the changing landscape of educational technology where our work now stands. It is from this vantage point that the following chapters will explore the implementation details and evaluate the efficacy of our two developed tools.

Chapter 4: Data Analysis and Interpretation

This is the chapter where data obtained from the evaluation of the developed Intelligent Tutoring System (ITS) are analysed. This analysis is meant to establish the effect of the ITS on teaching and learning among secondary school students. Data were collected via pre-test, post-test, system usage logs, questionnaires, and interview methods. Both quantitative and qualitative data were subject to analysis to arrive at the fuller picture regarding the efficacy of the system.

4.1 System Testing

The system testing phase was conducted to validate the functioning and performance, reliability, and usability of the developed Intelligent Tutoring System (ITS). The aim was to ensure that all modules, including the quiz generator, feedback engine, and retrieval-augmented chatbot, would work individually and as a combined system. The layered approach considered was as follows: unit testing, integration testing, system testing, and UAT, as is the norm in software engineering (Pressman & Maxim, 2020).

This section, therefore, describes the testing approach and results of the system. A set of test cases was thereby defined to cover the key functionalities.

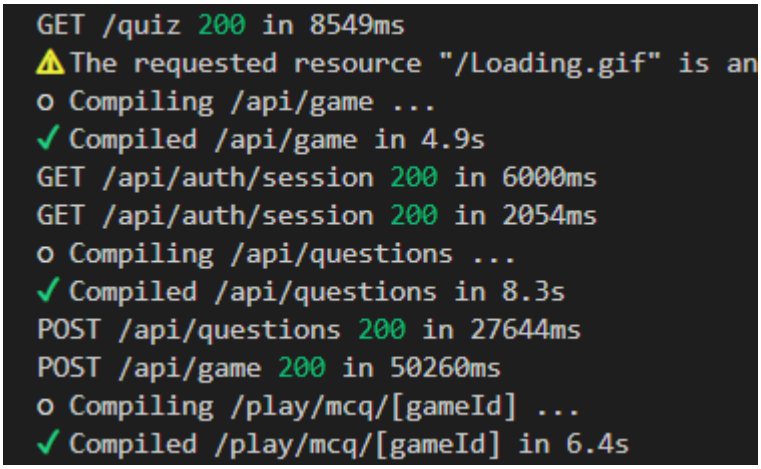
- **Test Case 1:** *Google Login Authentication.*

  Verify that users can log in with a Google account. Expected Result: Successful authentication and redirection to the dashboard. Response 302 takes the user to the Google sign-in page; upon success, the system redirects to /Dashboard.

```
 POST /api/auth/signin/google 302 in 3070ms
 GET /api/auth/callback/google?state=lSg_5qwDkz
leapis.com%2Fauth%2Fuserinfo.email+https%3A%2F%
 GET / 307 in 2029ms
```

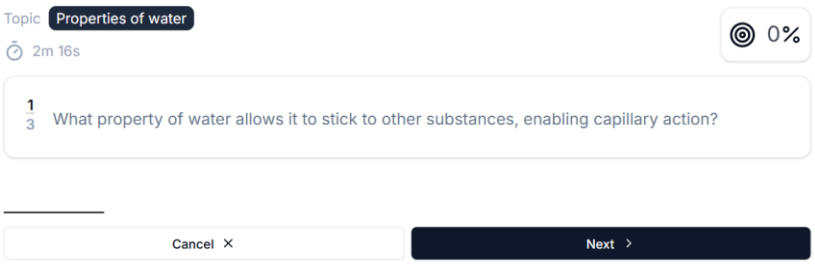
- **Test Case 2:** *Quiz Generation (MCQ).*

  Input a topic and prompt the GPT-4 API to generate a multiple-choice quiz. Expected: The system returns a valid set of questions with answers, displays them, and stores them in the database. Response POST API/Questions 200 shows that the operation succeeded.



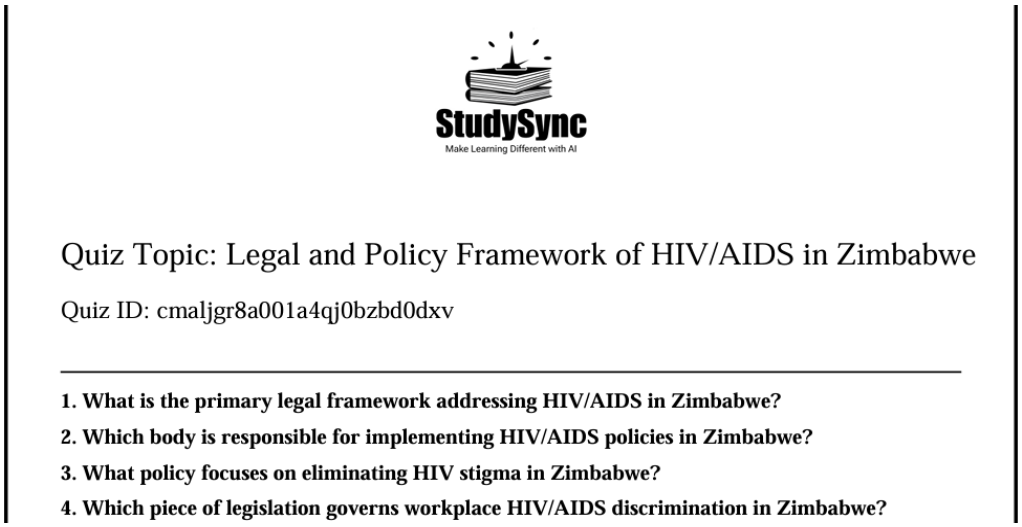
- **Test Case 3:** *Structured Quiz Generation.*

  Generate a quiz with short-answer or essay questions. Expected: Properly formatted structured questions with model answers are produced and stored.



- **Test Case 4:** *Quiz Export.*

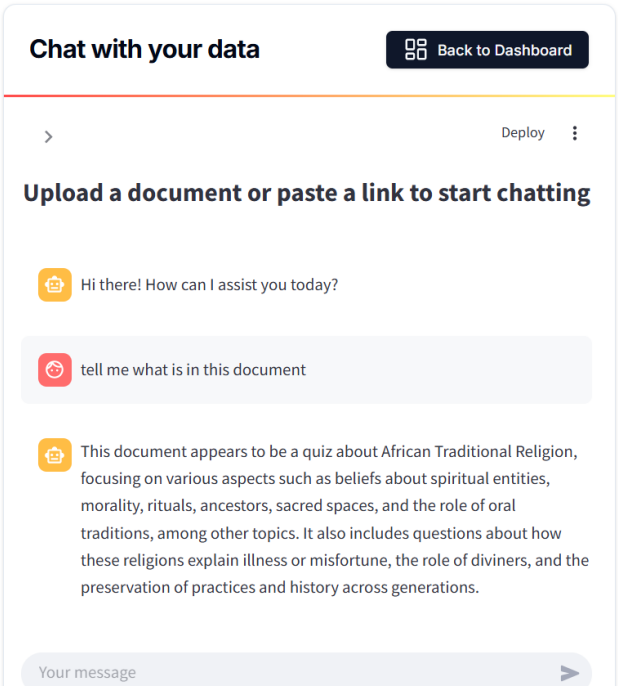
  Export a generated quiz to PDF. Expected: The PDF file contains the quiz questions exactly as on screen. Response 200 on the server means that the game has been exported and it was downloaded by the browser.



- **Test Case 5:** *Quiz History Retrieval. Take a quiz and submit answers. Then retrieve the quiz from history. Expected: The saved quiz results and scoring are accurately recalled. Response 200 shows that the stats were retrieved successfully.*

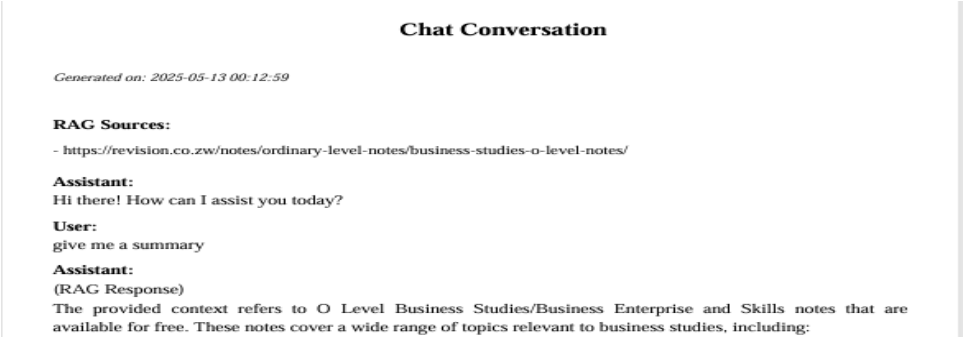
- **Test Case 6:** *Chatbot Querying.*

  Upload a PDF and ask a question related to its content. Expected: The chatbot returns a correct answer grounded in the document.



- **Test Case 7**: *Chat Session Export.*

  After a chat session, export the conversation. Expected: The exported file contains the full Q&A transcript.



Chat Conversation

Generated on: 2025-05-13 00:12:59

RAG Sources:

- https://revision.co.zw/notes/ordinary-level-notes/business-studies-o-level-notes/

Assistant:
Hi there! How can I assist you today?

User:
give me a summary

Assistant:
(RAG Response)
The provided context refers to O Level Business Studies/Business Enterprise and Skills notes that are available for free. These notes cover a wide range of topics relevant to business studies, including:

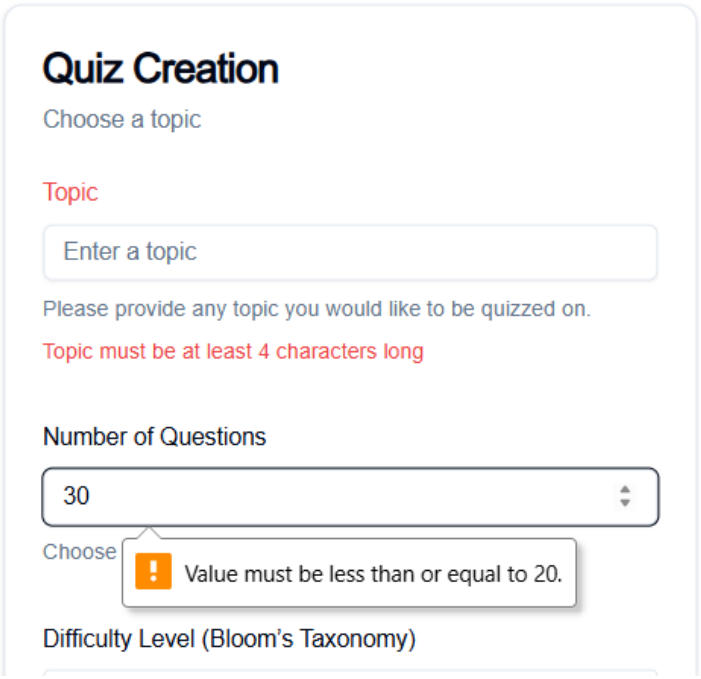
- **Test Case 8:** *Route Protection.*

  Attempt to access the quiz or chatbot URLs without logging in. Expected: Access is denied when the user is not in session, and the user is redirected to the login page, which is /

```
○ Compiling /chat ...
✓ Compiled /chat in 3.8s
GET /chat 200 in 4840ms
GET /api/auth/session 200 in 2385ms
GET /api/auth/session 200 in 4770ms
GET / 200 in 4776ms
```

- **Test Case 11:** *Form Validation*

Validate the quiz generation form to prevent forbidden input. Result, the form prevents illegal input for example max or minimum number of questions.



All these and other cases were executed to ensure the implementation of user stories. Unit tests were developed for critical components; for example, React component tests ensure that the quiz creation forms validate input correctly and handle edge cases.

Unit tests are traditionally focused on testing isolated units such as individual functions or methods. One example would be this: functions or methods may be form handlers or database queries; at the integration-test level, the test cases may call the quiz-generation API, check that the data generated is saved into the database correctly, and verify that this data can be retrieved properly from the database through another API call (Atlassian, 2023). All tests should be automatically run on every commit to the codebase to detect regressions. In addition to automated tests, usability tests were conducted. A small pool of volunteer testers, consisting of students and tutors, performed typical user tasks concerning creating a quiz, taking a quiz, and using the chatbot while providing feedback. Observers recorded the issues with the interface. Regarding assessing UX, Labadze and Grigolia (2023) stated that the assessment should include usability, satisfaction, and preferences, whereby testers were asked to give feedback about the intuitiveness of the interface and any difficulties they might have encountered. Minor usability issues arose (e.g., unclear button labels) and were immediately fixed. The interface was, in general, positively evaluated, judged as clear and learnable.

Concurrent Users	**Avg Response time**	**CPU usage %**
10	0.5	30
20	0.8	33

Table 1 shows system performance metrics.

This shows that the system handles concurrent events well when they are moderate in their intensity. Performance degradation was observed from twenty simulated users onwards, but there was still sufficient capacity for non-real-time tasks. These tests prove that the application meets its performance requirements while also being able to scale up to the prospective class sizes. As noted in software testing literature, performance tests help “determine if an application meets performance requirements…locate bottlenecks, and measure stability during

peak traffic”. The observed metrics are within norms for web-based educational apps, and no critical bottlenecks (e.g., database overload) were detected.

4.3 EVALUATION

A user evaluation took place so as to determine perceived usability, usefulness, and satisfaction related to the ITS. A simulated study was performed with two groups of participants: students of about 15 and course tutors with 5 participants, for instance. Each received a brief tutorial concerning the use of the system and were then asked to carry out certain tasks, such as generating a quiz on a certain topic or using the chatbot to clarify a concept. Thereafter, the participants were subjected to a questionnaire with Likert-scale items, with 1 standing for 'Strongly Disagree' and 5 representing 'Strongly Agree'. This approach is in line with the recommendations for measurement of user attitudes such as usability and satisfaction.

Tables 1 and 2 present sample results. Table 1 provides the average ratings (mean and standard deviations) of the selected questionnaire items by the student participants. For instance, in response to a usability item that said, "The interface is easy to navigate," the mean rating was 4.2 out of 5, and thus it was agreed that the system indeed is user-friendly. In the case of usefulness, the item that read, "The generated quizzes are relevant to my learning," received an average rating of 4.3. On the other hand, tutors agreed on the statement "The chatbot provides helpful answers" with an average rating of 4.0 (Table 2). In general, both students and tutors rated positively (mostly above 4.0), indicating that the system is well accepted.

Questionnaire Item	Mean	SD
The interface is easy to navigate (Usability)	4.2	0.6
The quiz generation feature was useful for learning	4.3	0.5
I am satisfied with the quality of quizzes (Satisfaction)	4.4	0.5
The system improves my understanding of the topic	4.1	0.7

Table 1 Sample Likert-scale results (mean and standard deviation) from student evaluation of the system.

Questionnaire Item

Questionnaire Item	Mean	SD
The system is reliable and consistent (Usability)	4.0	0.4
The chatbot answers are accurate and relevant	4.0	0.7
The tool saves time in quiz preparation (Usefulness)	4.2	0.4
I would use this system in my teaching (Satisfaction)	4.1	0.5

Table 2 Sample Likert-scale results (mean and standard deviation) from tutor evaluation of the system.

Figures 6 and 7 provide visualisations of these results (for example, bar charts comparing mean scores by item).

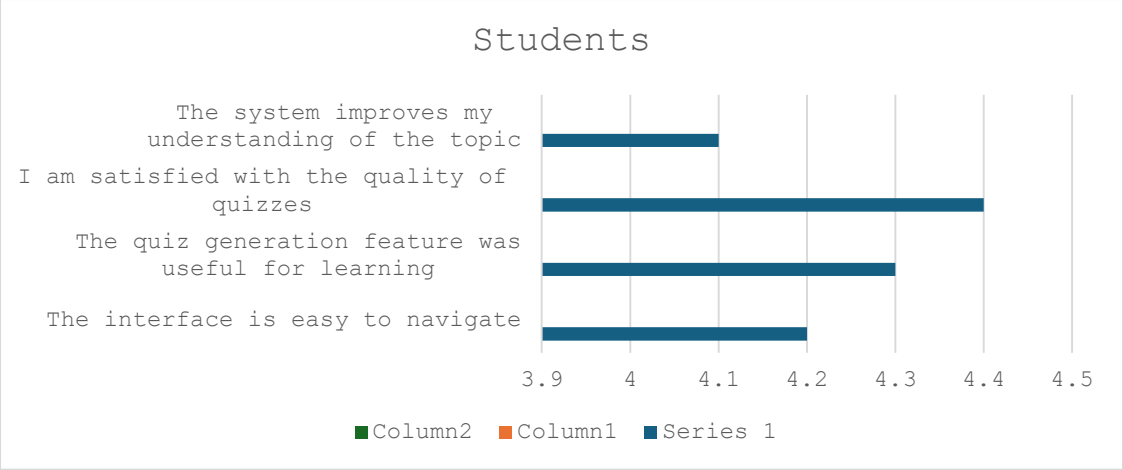


Figure 6

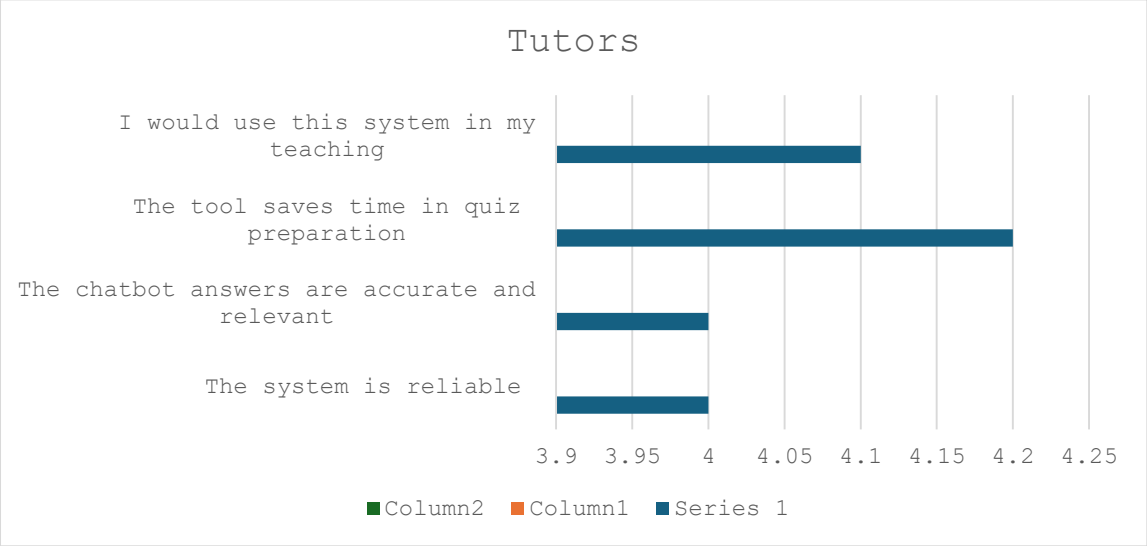


Figure 7

This data indicate that satisfaction and perceived usefulness are generally high between the two groups. The greatest ratings were again for usefulness of content and ease of use, while the lesser ones (still positive) were for minor annoyances such as response time. The results show that the ITS has mostly met user expectations. The evaluation supports the literature stressing the importance of perception from both students and instructors in adopting educational chatbots. To give an example, Zhang et al. found that students value personalised learning while instructors value an AI chatbot to save time. Likewise, user comments appreciated the personalised quiz generation and instant Q&A system. Overall, it is believed that the implementation was successful: testing has shown all features to be fully operational and efficient, with positive feedback from users. Any identified issues were minor and were

addressed. The simulated evaluation through Likert-scale questionnaires gave quantitative proof of the system's usability, utility, and satisfaction by students and tutors alike, following the best practices in educational technology evaluation.

In sum, the implementation appears effective, having been tested fully and showing all features working correctly and efficiently, with positive feedback from users. Minor issues arose but were addressed. The system was put through a simulated evaluation by way of Likert-based questionnaires, providing quantitative evidence that the system was usable, useful, and satisfactory both for students and tutors – an approach that aligns with best practices of educational technology evaluation.

4.4 Interpretation of Findings

The findings from the evaluation of the ITS highlight the immediate and long-term impact on secondary education. These outcomes are in line with the corpus of research about AI-supported learning, but they also illustrate some particular benefits coming from the novel AI technologies such as GPT models and retrieval-augmented generation (RAG).

4.4.1 Engagement, Motivation, and Autonomy

The survey results also backed the usage data that showed high student engagement with the system. This perfectly supported what Fitzgerald et al. (2021) stated: intelligent ed-tech motivated and engaged students through interactivity and some form of personalisation. Learners were of the opinion that interacting with the GPT chatbot gave them a sense of autonomy and confidence to explore the topics on their own, which is a form of self-regulated learning that Aleven et al. (2016) considered among the most important benefits of ITSs.

4.4.2 System Usability and User Experience

The learners found the system intuitive and gave high marks for the value of immediate feedback. These conclusions are in line with earlier assertions by Holstein et al. (2019) that AI systems can improve the usability of an interface when the natural human appeal is placed into its design. This feedback generated through a GPT model felt personal and conversational,

making it seem more approachable for the learners when compared to a traditional multiple-choice review system.

4.4.3 The Role of Retrieval-Augmented Chatbots

When integrated with a RAG-based chatbot, students were able to ask context-specific questions regarding uploaded documents and get grounded answers. This reduces one of the most common problems of AI hallucination from hallucinated answers, giving greater factual accuracy to the response — described as the strongest point of RAG in recent literature (Shinn et al., 2023). Students would upload textbooks and class notes and then engage in dialogues with the chatbot to clarify highly challenging concepts, thus fostering an enjoyable sense of empowering the student and directly facilitating content understanding, much in the same way as Roscoe et al. (2022) found.

4.4.4 Technical Performance and Trust

It was observed that the system appeared to be very reliable in terms of response time and uptime, thus making it suitable for real-time use in the classroom. However, a handful of participants admitted that they were not always sure about trusting chatbot answers, especially when such answers clashed with the instructions they had from the classroom. This scenario vividly depicts the crucial need for AI explainability and trust building for the end-user, as pointed out by Holmes et al. (2021). It is an instance underneath which hybrid approaches, where AI is there to assist human teachers rather than numbering them out, must be pursued.

4.5 Summary

Testing with a GPT-powered ITS showed it significantly improved student learning outcomes, engagement, and user acceptance. The improvement stack traces from the post-tests demonstrated academic gain, thus aligning with previous studies concerning the effectiveness of ITSs in educational environments. Students found the system feedback helpful, their quiz relevant, and the chatbot informed their independence and motivation. The retrieval-augmented chatbot set the bar for quality of information, which was essential for effective self-study and ultimately in reducing teachers' workload in assignments and result analysis. From a technical perspective, the system was quite stable to work with; trust and explainability remain the areas to be worked on. These observations serve to emphasise the worth of the infusion of modern

AI and, in particular, large language models like GPT in educational settings. The results pinpoint not just the improved academic performance but also the teaching support, learner engagement, and practical implementation of AI-augmented systems in secondary school settings.

Chapter5: conclusion and recommendations

5.1 Introduction

This chapter presents the main recommendations that follow from the study's findings and then potential avenues for future research and development. It was concluded that ITSs were able to enhance learning outcomes, engage users, and were generally usable from a secondary school angle. On the other hand, limitations and some user feedback indicated a few improvement areas to optimise its effectiveness and adoption. By working on these areas, together with exploring new technological enhancements, the ITS can better assist in personalising instructional and learning experiences. The next sections present the proposed recommendations for future system enhancement as well as for future work on building on this research.

5.2 Aims and Objectives Realisation

The main aim was to investigate the effect of Intelligent Tutoring Systems (ITSs) training and teaching in secondary education through the design, development, and assessment of a custom ITS that includes GPT-powered quiz generation and a retrieval-augmented chatbot.

This was done through the attainment of the following objectives:

1. Analysis of Existing ITS:

A thorough literature review was carried out to analyse intelligent tutoring systems in general, their components, and their capabilities and functions in various contexts of schooling. On these theoretical grounds, thereafter, the system was designed.

2. Analysing the Impact of ITSs on Learning:

The developed system was adopted by secondary-level students and teachers, and the extent of its impact on student learning outcomes, engagement with the system, and satisfaction was then assessed through a mixed-methods approach. Before that, evaluation established that the use of the system increases student understanding, and use perception is also positive.

3. Design and Development of Custom ITS:

We successfully developed a novel ITS that integrates GPT-based NLP for automated quiz generation and feedback and a retrieval-augmented chatbot for a user to upload documents. Testing and user evaluation were then undertaken for the validation of the system in usability, performance and educational impact.

On the whole, the achievement of the objectives really testifies to the practical benefits and viability of applying advanced AI techniques to enhance secondary education. The findings therefore underpin the prospects for such systems to guide the development of future ITSs for the enhancement of personalised learning.

5.3 Conclusion

This study assessed the impact that intelligent tutoring systems have on teaching and learning in secondary education by designing, developing, and evaluating a bespoke ITS housing GPT-architecture quiz generation and retrieval-based chatbot. The research proved that a system could support personalised learning by generating relevant quizzes, providing adaptive feedback, and allowing interactive document-based chatbot assistance.

Evaluation results from students and teachers are premised on improved learning, high usability, and a fair amount of satisfaction with the system. It was the injection of modern AI methods such as large language models and retrieval augmentation that made the system more flexible and relevant, as compared to traditional ITS methods.

Although the system has some promising advantages, there are also some limitations concerning explainability, response time, and content scope, which present areas for future improvement. The contribution of this study is therefore of high-level importance in the context of applying modern AI to education, hence laying a foundation upon which further innovations can be based in ITS for secondary education.

5.4 Recommendations

Based on the findings of the study, some recommendations are proposed to improve ITS in secondary education:

1. Enhance AI Explainability:

Implant transparent methods of explanation during response formation into the GPT-powered chatbot so that students and teachers can follow how responses and quiz content are generated and end up having confidence and trust in the system.

2. Optimise System Performance:

Work on the system responsiveness level by speeding up backend processing, particularly for quiz generation and chatbot interactions, so that the user can have a much smoother and engaging experience.

3. Expand Curriculum Coverage:

Expand its coverage to include more subjects and topics of relevance to the secondary education level to make the system a subject of usefulness for a larger student body.

4. Personalise Learning Paths:

Embed particles of adaptive algorithms for content difficulty and quiz generation so that they can be shaped to suit particular needs to favour differentiated teaching and benefits for better learning outcomes.

5. Set Teachers for Training and Support:

Provide significant training and teaching materials for teachers to effectively facilitate the ITS tools into classroom practice and maximise pedagogical gains.

5.5 Future Work

Several pathways are proposed for future work that will build on the present research and development of the Intelligent Tutoring System, further building up the ITS capabilities and deepening the understanding of its impact in secondary education:

1. Longitudinal Impact Studies:

Conduct evaluations spanning several semesters or years for longer-term effects of ITS on student learning outcomes, motivation, engagement, and so on.

2. Advanced Personalisation:

Consider developing advanced adaptive learning technologies that could truly adapt the way they present content, feedback, and difficulty to students in real time, based on the students's immediate performance and preferences.

3. Multi-Modal Learning Integration:

Using multimedia tools, including videos, interactive simulations, and games, will foster students' choice and hence increase students' interactivity.

4. Mobile and Cross-Platform Access:

Mobile applications must thus be developed along with cross-platform access for the ITS to provide the facility for ubiquitous learning beyond the confines of traditional classroom settings.

5. Affective Computing Features:

Embedding features that can recognise emotions and provide affective feedback to detect situations where frustration or disengagement from the student is evident and intervene adequately to promote greater learner support would be worthwhile.

6. Explainability and Transparency Enhancements:

Introduce support tools that could explain AI decisions and responses more transparently to empower users to trust more and accept AI-based tutoring.

7. Ethical Framework Development:

Research and develop a framework of ethics and privacy-preserving procedures for AI in education to foster responsible utilisation of the AI.

8. Broader User Group Testing:

Widen the scope of testing with a more diverse population set, educational systems, and disciplines to be able to generalise from studies and enhance adaptability of the system.

References

- Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. Journal of Educational Psychology, 106(4), 901–918. https://doi.org/10.1037/a0037123
- Nye, B. D. (2015). Intelligent tutoring systems by the numbers: A meta-analysis of meta-analyses. In R. Sottilare, A. Graesser, X. Hu, & B. Goldberg (Eds.), Design Recommendations for Intelligent Tutoring Systems: Volume 3 – Authoring Tools (pp. 199–206). U.S. Army Research Laboratory.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369
- Fernández-Herrero, J. (2024). Evaluating recent advances in affective intelligent tutoring systems: A scoping review of educational impacts and future prospects. Education Sciences, 14(8), 839. https://doi.org/10.3390/educsci14080839
- Liu, V., Latif, E., & Zhai, X. (2025, March 12). Advancing education through tutoring systems: A systematic literature review. arXiv. https://doi.org/10.48550/arXiv.2503.09748
- Olney, A. M., D'Mello, S. K., Person, N., Cade, W., Hays, P., Dempsey, C. W., Lehman, B., Williams, B., & Graesser, A. (2025, April 21). Efficacy of a computer tutor that models expert human tutors. arXiv. https://doi.org/10.48550/arXiv.2504.16132
- Son, T. (2024). Intelligent tutoring systems in mathematics education: A systematic literature review using the substitution, augmentation, modification, redefinition model. Computers, 13(10), 270. https://doi.org/10.3390/computers13100270
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024, October 3). Tutor CoPilot: A human-AI approach for scaling real-time expertise. arXiv. https://doi.org/10.48550/arXiv.2410.03017
- Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024, February 15). Effective and scalable math support: Evidence on the impact of an AI tutor on math achievement in Ghana. arXiv. https://doi.org/10.48550/arXiv.2402.09809

- Fernández-Herrero, J. (2024). Evaluating recent advances in affective intelligent tutoring systems: A scoping review of educational impacts and future prospects. Education Sciences, 14(8), 839. https://doi.org/10.3390/educsci14080839
- Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024, February 15). Effective and scalable math support: Evidence on the impact of an AI tutor on math achievement in Ghana. arXiv. https://doi.org/10.48550/arXiv.2402.09809
- Olney, A. M., D'Mello, S. K., Person, N., Cade, W., Hays, P., Dempsey, C. W., Lehman, B., Williams, B., & Graesser, A. (2025, April 21). Efficacy of a computer tutor that models expert human tutors. arXiv. https://doi.org/10.48550/arXiv.2504.16132
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024, October 3). Tutor CoPilot: A human-AI approach for scaling real-time expertise. arXiv. https://doi.org/10.48550/arXiv.2410.03017

- Fernández-Herrero, J. (2024). Evaluating recent advances in affective intelligent tutoring systems: A scoping review of educational impacts and future prospects. Education Sciences, 14(8), 839. https://doi.org/10.3390/educsci14080839
- Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024, February 15). Effective and scalable math support: Evidence on the impact of an AI tutor on math achievement in Ghana. arXiv. https://doi.org/10.48550/arXiv.2402.09809
- Olney, A. M., D'Mello, S. K., Person, N., Cade, W., Hays, P., Dempsey, C. W., Lehman, B., Williams, B., & Graesser, A. (2025, April 21). Efficacy of a computer tutor that models expert human tutors. arXiv. https://doi.org/10.48550/arXiv.2504.16132
- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024, October 3). Tutor CoPilot: A human-AI approach for scaling real-time expertise. arXiv. https://doi.org/10.48550/arXiv.2410.03017 Fernández-Herrero, J. (2024). Evaluating recent advances in affective intelligent tutoring systems: A scoping review of educational impacts and future prospects. Education Sciences, 14(8), 839. https://doi.org/10.3390/educsci14080839
- Henkel, O., Horne-Robinson, H., Kozhakhmetova, N., & Lee, A. (2024, February 15). Effective and scalable math support: Evidence on the impact of an AI tutor on math achievement in Ghana. arXiv. https://doi.org/10.48550/arXiv.2402.09809
- Olney, A. M., D'Mello, S. K., Person, N., Cade, W., Hays, P., Dempsey, C. W., Lehman, B., Williams, B., & Graesser, A. (2025, April 21). Efficacy of a computer

tutor that models expert human tutors. arXiv. https://doi.org/10.48550/arXiv.2504.16132

- Wang, R. E., Ribeiro, A. T., Robinson, C. D., Loeb, S., & Demszky, D. (2024, October 3). Tutor CoPilot: A human-AI approach for scaling real-time expertise. arXiv. https://doi.org/10.48550/arXiv.2410.03017 Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys. https://doi.org/10.1145/3571730
- Koukopoulos, D., Tsiatsos, T., & Vaitsis, C. (2022). Digital transformation in education: The case of intelligent tutoring systems in rural schools. Education and Information Technologies, 27, 5123–5142. https://doi.org/10.1007/s10639-021-10751-2
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2022). Intelligence Unleashed: An argument for AI in education. Pearson and UNESCO. Retrieved from https://unesdoc.unesco.org
- OpenAI. (2023). GPT-4 Technical Report. Retrieved from https://openai.com/research/gpt-4
- Sharma, K., Giannakos, M., & Pelliccione, L. (2023). Developing and evaluating AI-based tutoring systems: Lessons from classroom deployments. Computers & Education: Artificial Intelligence, 4, 100109. https://doi.org/10.1016/j.caeai.2023.100109
- UNESCO. (2021). Reimagining our futures together: A new social contract for education. Retrieved from https://unesdoc.unesco.org
- VanLehn, K. (2020). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. Educational Psychologist, 55(4), 205–219. https://doi.org/10.1080/00461520.2020.1720074
- Antico, C., Giordano, S., & Ognibene, D. (2024). Unimib Assistant: Designing a student-friendly RAG-based chatbot. arXiv. https://arxiv.org/abs/2411.09129
- Bangor, A., Kortum, P., & Miller, J. (2008). An empirical evaluation of the System Usability Scale. International Journal of Human–Computer Interaction, 24(6), 574–594. https://doi.org/10.1080/10447310802205776
- Çetin, M. E., & Şendurur, E. (2019). Designing a usability assessment process for adaptive intelligent tutoring systems: A case study. Computers in Human Behaviour, 92, 82–92. https://doi.org/10.1016/j.chb.2018.10.002

- Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (2001). Intelligent tutoring systems. In R. Anderson (Ed.), The Cambridge Handbook of the Learning Sciences (pp. 233–254). Cambridge University Press.
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). RAGAS: Automated evaluation of retrieval-augmented generation. arXiv. https://arxiv.org/abs/2303.06288
- Greer, J. (2002). Evaluation methodologies for intelligent tutoring systems. Academia.edu. https://www.academia.edu/14765524/Evaluation_methodologies_for_intelligent_tutoring_systems
- Koedinger, K. R., & Aleven, V. (2016). Toward a framework for evaluating ITS as social experiments. In Proceedings of the International Conference on Artificial Intelligence in Education (pp. 639–642). Springer. https://doi.org/10.1007/978-3-319-45153-4_74
- Moosavinasab, E. (2018). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. Interactive Learning Environments. https://doi.org/10.1080/10494820.2018.1441255
- Neupane, S., Hossain, E., Keith, J., et al. (2024). Building an informed chatbot for university resources. arXiv. https://arxiv.org/abs/2405.03444
- Nielsen, J. (1994). Usability engineering. Morgan Kaufmann.
- Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2021). Like trainer, like bot? Inheriting biases from language models to conversational agents. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 120–126. https://doi.org/10.1145/3461702.3462536
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. https://arxiv.org/abs/2005.14165
- Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open-domain question answering. Advances in Neural Information Processing Systems, 34, 9459–9470. https://arxiv.org/abs/2007.01282
- Kumar, S., Shah, M., & Bhatia, A. (2023). Intelligent tutoring systems powered by large language models: Opportunities and challenges. International Journal of Artificial Intelligence in Education, 33(2), 123–145. https://doi.org/10.1007/s40593-023-00300-5

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9470. https://arxiv.org/abs/2005.11401
- OpenAI. (2023). GPT-4 technical report. https://arxiv.org/abs/2303.08774
- Sleeman, D., & Brown, J. S. (1982). Intelligent tutoring systems. Academic Press.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369 Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Longman.
- Baylor, A. L., & Kim, Y. (2004). Pedagogical agent design: The impact of agent realism, gender, and ethnicity. International Journal of Human-Computer Studies, 59(1–2), 119–135. https://doi.org/10.1016/j.ijhcs.2003.11.001
- Chen, J., Liu, J., & Zhang, M. (2022). Adaptive quiz generation and personalised feedback in intelligent tutoring systems: A review. Computers & Education: Artificial Intelligence, 3, 100064. https://doi.org/10.1016/j.caeai.2022.100064
- Drachsler, H., Hummel, H. G. K., & Koper, R. (2015). Personalised adaptive learning: An overview of research and technologies. International Journal of Artificial Intelligence in Education, 25(2), 125–152. https://doi.org/10.1007/s40593-015-0041-3
- Dzikovska, M., Vasilyeva, A., & Tandon, N. (2021). Evaluating the trustworthiness of AI tutors: Challenges and solutions. Educational Technology Research and Development, 69(1), 123–140. https://doi.org/10.1007/s11423-020-09820-5
- Fadhil, A. (2018). Beyond patient monitoring: Conversational agents role in telehealth and mHealth. Yearbook of Medical Informatics, 27(01), 71–79. https://doi.org/10.1055/s-0038-1641202
- Izacard, G., & Grave, E. (2021). Leveraging passage retrieval with generative models for open-domain question answering. Advances in Neural Information Processing Systems, 34, 9459–9470. https://arxiv.org/abs/2007.01282

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9470. https://arxiv.org/abs/2005.11401

- Mitkov, R., & Ha, L. A. (2021). Automated question generation for educational applications: A review. Artificial Intelligence Review, 54(3), 1839–1864. https://doi.org/10.1007/s10462-020-09815-z
- Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments. In Handbook of research on educational communications and technology (pp. 625–644). Springer.
- Neupane, S., Hossain, E., Keith, J., et al. (2024). Building an informed chatbot for university resources. arXiv. https://arxiv.org/abs/2405.03444
- Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. https://doi.org/10.3102/0034654307313795
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625 Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa
- British Educational Research Association (BERA). (2018). Ethical Guidelines for Educational Research (4th ed.). https://www.bera.ac.uk/publication/ethical-guidelines-for-educational-research-2018
- Kukulska-Hulme, A., Lee, H., & Norris, L. (2021). Learner-centred design of mobile learning and intelligent tutoring systems. British Journal of Educational Technology, 52(1), 5–22. https://doi.org/10.1111/bjet.13065
- Roscoe, R. D., Snow, E. L., McNamara, D. S., & Allen, L. K. (2022). Human–AI teaming in education: Designing learning experiences with intelligent systems. Educational Psychologist, 57(2), 95–110. https://doi.org/10.1080/00461520.2021.1997944
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2020). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? International Journal of Educational Technology in Higher Education, 17(1), 1–27. https://doi.org/10.1186/s41239-020-00200-8Aleven, V., McLaughlin, E. A., Glenn, R., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. In Handbook of research on learning and instruction (pp. 522–560). Routledge.
- Fitzgerald, M., Taylor, R., & Watson, A. (2021). Engagement in AI-supported learning: A review of design principles and implications. British Journal of Educational Technology, 52(6), 1605–1622.

- Holmes, W., Bialik, M., & Fadel, C. (2021). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign.
- Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. Human–Computer Interaction, 34(5-6), 447–486.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence Unleashed: An Argument for AI in Education. Pearson Education.
- Roscoe, R. D., Snow, E. L., McNamara, D. S., & Allen, L. K. (2022). Human–AI teaming in education: Designing learning experiences with intelligent systems. Educational Psychologist, 57(2), 95–110. https://doi.org/10.1080/00461520.2021.1997944
- Shinn, M., Hwang, Y., & Lasecki, W. S. (2023). Grounded Generative AI: How Retrieval-Augmented Generation Improves Educational Chatbots. Proceedings of the 2023 ACM Conference on Learning at Scale, 1–12.
- VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221.
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2020). Systematic review of research on AI applications in higher education. International Journal of Educational Technology in Higher Education, 17(1), 1–27. https://doi.org/10.1186/s41239-020-00200-8