VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

FACULTY OF FUNDAMENTAL SCIENCES
DEPARTMENT OF INFORMATION SYSTEMS

Mohini Naga Venkata Poojaswi Kankipati (20241937)

Online Quiz Maker using Google Colab

Master Graduation Thesis

Supervised by: Assoc. prof. Algirdas Laukaitis

VILNIUS 2025

CONTENT

List of Figures

List of Tables

Abbreviations

The abbreviations used in this document are described as follows:

AI – Artificial Intelligence

API – Application Programming Interface

BERT – Bidirectional Encoder Representations from Transformers

BPMN – Business Process Model and Notation

CAD – Computer-Aided Design

CPU – Central Processing Unit

CPE – Cloud-Based Platform for Education

CR – Collaborative Resources

DL – Deep Learning

DLT – Distributed Learning Tools

E-Learning – Electronic Learning

GC – Google Colab

GPU – Graphics Processing Unit

GPT – Generative Pretrained Transformer

LLM – Large Language Model

MCQ – Multiple Choice Questions

MEGADOCK – Molecular Evolutionary Genetic Analysis

MEUS – Mobile E-Learning Platform for Ultrasound

ML – Machine Learning

NLP – Natural Language Processing

NFLS – Newton's Fractal Leonardo Sequence

NER – Named Entity Recognition

OQM – Online Quiz Maker

PGK – Python-Based Quiz Knowledge

PI – Performance Improvement

PV Education – Photovoltaics Education

RA – Resource Accessibility

RLS – Recurring Linear Sequence

RTF – Real-Time Feedback

SVM – Support Vector Machines

TPU – Tensor Processing Unit

1. Introduction

The problem of accurately and efficiently grading written or descriptive answers in online quizzes is a significant challenge in the development of assessment tools. Unlike multiple-choice questions (MCQs), written answers are subjective, vary in length and wording, and require contextual understanding to determine correctness. Traditional grading systems, (Georgiev & Nikolova, 2020) which rely on manual evaluation or basic keyword matching, are time-consuming, inconsistent, and unable to handle diverse responses effectively. This issue becomes more complex when dealing with multilingual students, varying writing styles, and the use of synonyms, which often lead to unfair or inaccurate grading in keyword-based approaches. Educators face the added burden of providing personalized feedback, (Hoang et al., 2022) which increases their workload and delays the assessment process. Current solutions fail to account for grammar errors, spelling mistakes, or partial correctness in responses, further limiting their reliability and fairness. As online education grows, especially in large-scale courses, the inability to grade descriptive answers efficiently and accurately becomes a bottleneck, leaving educators overwhelmed and students underserved.

In high-enrollment courses, the challenge of grading thousands of written answers quickly often leads to delays, frustrating students and impeding their learning experience. Inconsistent grading due to factors like fatigue, bias, or varying levels of expertise undermines fairness, while keyword-based systems fall short in recognizing rephrased but correct answers, unfairly penalizing students. Moreover, feedback, (Hoang et al., 2022) which is vital for student improvement, is often either generic or nonexistent, limiting its effectiveness. The inability to assess partial correctness also diminishes student engagement and motivation. As the demand for open-ended questions increases to promote critical thinking and deeper learning, traditional grading systems become a significant barrier, as they struggle to handle issues like grammar, spelling, and scalability. This situation underscores the urgent need for innovative, scalable, and context-aware solutions to transform assessment practices and improve learning outcomes.

The proposed solution to address the challenge of grading written answers efficiently is to use AI and Google Colab for automated evaluation. By integrating AI-powered natural language processing (NLP) models, (Peng & Niu, 2021) such as GPT or BERT, (Jia et al., 2022) the system can analyse written responses for accuracy, grammar, and context. Google Colab provides an accessible environment for developing machine learning models without the need for additional infrastructure. Pre-trained models will be fine-tuned on academic datasets to recognize rephrased but correct answers. This approach overcomes the limitations of traditional keyword-based grading, ensuring fairer and more consistent evaluations. The AI model will handle grammatical errors, spelling mistakes, and diverse writing styles, making it inclusive for non-native speakers. Partial credit can be assigned to responses demonstrating partial understanding,

improving fairness. Real-time feedback will be provided to students, reducing delays in grading. Educators can customize grading criteria through a user-friendly interface, making the tool adaptable to different subjects. The model will be designed for scalability, enabling efficient grading for high-enrolments courses. To improve accuracy, (Kimm et al., 2021) the system needs further refinement in handling complex academic language and domain-specific terminology. AI will also help assess creativity and critical thinking in open-ended questions, promoting deeper learning. Incorporating multilingual capabilities, such as translation tools, will enhance inclusivity. Analytics features will allow educators to identify trends in student performance and common errors. Offline functionality should be developed for use in low-resource environments. Continuous feedback from educators will help the system improve its accuracy over time through reinforcement learning. This will reduce the educator's workload while providing timely, constructive feedback to students. By automating grading, the tool can save time and ensure more consistent evaluations. Google Collab's accessibility and scalability make this solution feasible for institutions worldwide. The integration of AI with Google Colab will improve grading efficiency, fairness, and inclusivity in educational settings.

1.1 Investigation Object

The investigation will focus on exploring the automation of grading (Chieu et al., 2011) written answers using AI-powered systems, particularly analysing how these systems can accurately assess grammar, context, and relevance. The objective is to evaluate the effectiveness of AI tools (Ferreira et al., 2024) in providing consistent, fair, and timely feedback for open-ended questions, reducing educator workload while enhancing student engagement and learning outcomes.

1.2 The Aim and Task of the Thesis

The aim of this research is to enhance the efficiency of grading answers using AI and machine learning, with the help of Google Colab. The goal is to develop a system that accurately grades written responses, recognizes varied phrasing, and provides contextual feedback**.**

The main tasks for the research are as follows:

1. To analyse the current methods used for grading written exams and existing tools, identifying limitations and areas for improvement.
2. To propose an approach for automatically evaluating answers, leveraging AI and machine learning techniques to ensure accuracy and fairness.
3. To implement a system in Google Colab that evaluates the accuracy of answers, incorporating AI models to assess grammar, context, and relevance.

1.3 Novelty of the Topic

This thesis addresses a novel problem by utilizing Google Colab, traditionally a tool for data science, to develop an Online Quiz Maker. (Wang et al., 2022) While existing quiz platforms offer basic functionality, none fully leverage Google Colab's collaborative (Werth et al., 2022) scalable, and customizable features for automated grading and feedback. This research fills a gap by applying cloud-based programming tools to education, offering an innovative, cost-effective solution for creating dynamic, efficient assessments (Carneiro et al., 2018) . The findings will contribute new insights into how cloud platforms can enhance educational technologies.

1.4 Relevance of the Topic

This topic is highly relevant due to the increasing demand for efficient and scalable solutions in online education. Traditional quiz-making methods are time-consuming (Mokhtarzadeh et al., 2023) and lack customization, making automated, real-time feedback essential. By developing an Online Quiz Maker using Google Colab, (Peng & Niu, 2021) this research offers a cost-effective, accessible tool that enhances the assessment process, benefiting both educators and students in modern learning environments.

1.5 Research Methodology

The research methodology began with analyzing the needs of educators and students through surveys and interviews to identify requirements for the Online Quiz Maker. (Timpe, 2023) The inclusion criteria involved educators who use digital tools for teaching and students participating in assessments. The process followed by the design and development of the tool using Google Colaboratory and Python libraries for customization, and automated grading. (Ohue, 2023) Usability tests and surveys were conducted for feedback, with exclusion criteria focused on users with no experience using online quiz systems. Iterative refinements were made based on the feedback to enhance the system's effectiveness.

1.6 Scientific Value of the Thesis

The scientific value of this thesis lies in addressing the time-consuming nature of manual grading, particularly for answers. Educators spend significant time evaluating subjective responses, which can be both inconsistent and prone to fatigue, leading to delays in providing feedback. This manual process not only reduces the time teachers can allocate to other educational tasks but also impacts their ability to give personalized feedback to each student. By automating the grading process with AI, the thesis aims to streamline this workload, enabling teachers to focus more on interactive teaching and student engagement. This will enhance overall educational efficiency, improve the accuracy and consistency of grading, and ensure timely, constructive feedback for students, thus improving learning outcomes.

1.7 Main Results of the Thesis

The main result of analysing current methods for grading written exams and existing tools reveals several key limitations, including inconsistencies in grading, time delays, and potential biases due to human error and fatigue. Traditional tools often rely on keyword matching or subjective evaluation, which can lead to unfair assessments, particularly for diverse student responses. Additionally, these methods are not scalable for high-enrolments courses, and they lack the ability to provide detailed, personalized feedback. The analysis highlights the need for more efficient, consistent, and scalable solutions that can address these limitations and improve the grading process.

1.8 Structure of the Thesis

The structure of the thesis is as follows:

1. **Introduction**: The Introduction section gives an overview of the research problem, objectives, and significance of automating answer grading.
2. **Literature Review**: The Literature Review section explains existing online quiz systems and their limitations, as well as previous research in the field of educational technology.
3. **Proposed Methodology**: The Proposed Methodology Section Detailed description of the automated grading approach leveraging NLP techniques and cloud-based implementation. This section covers the five-layer grading workflow: Data Input, Preprocessing, Model Evaluation, Scoring, and Feedback. It explains how pre-trained language models such as GPT and BERT are used to interpret student answers, assign context-aware scores, and provide personalized feedback efficiently and at scale.
4. **Initial Experimental Results and Analysis:** The Initial Experimental Results and Analysis section presents experiments on real student answers, evaluating the NLP-based grading system using accuracy, precision, score. Results show improved grading consistency and feedback compared to manual and basic automated methods. It also discusses errors, partial credit, scalability, and current limitations.
5. **Conclusion And References:** The conclusion and References section includes the results obtained from the literature review, proposed methodology, and initial experimental analysis collectively demonstrate the effectiveness and potential of automated grading using NLP models. The study highlights improved grading accuracy, consistency, and personalized feedback, while also acknowledging current limitations and opportunities for future enhancement. All references cited throughout the thesis are compiled in the References section.

2. Related Work Analysis

2.1 Main Concepts

The main concepts of this research revolve around the challenges associated with grading written answers, understanding the research domain of AI-based grading systems, and exploring the specific techniques that can be used to automate the evaluation of responses. Grading answers manually is time-consuming (Haddad & Kalaani, 2014) and subjective, often leading to inconsistencies and bias, as teachers must evaluate diverse responses with varying language styles, phrasing, and levels of detail. Additionally, issues like fatigue, large student populations, and the need for personalized feedback add to the complexity of manual grading (Kumar Sinha & Kumar Gupta, 2015) One of the primary challenges is assessing answers that are rephrased or contain minor grammatical errors, which may still convey the correct information. These difficulties are further amplified in courses with high enrollment, where grading large volumes of written responses can delay feedback and hinder student progress.

In the research domain, AI-based grading systems have emerged as a potential solution to these challenges. NLP techniques are widely used in such systems, enabling them to understand and process written text. NLP (Haque et al., 2023) models like BERT (Reimers & Gurevych, 2019) and GPT (Saka et al., 2024) have shown significant promise in interpreting the meaning of written responses, even when phrasing varies or contains synonyms. Another important technique is Named Entity Recognition (NER), which helps AI models identify key entities in written answers, such as scientific terms or important concepts, to assess the relevance and accuracy of the response. Sentiment analysis and dependency parsing are additional NLP techniques that can be used to evaluate the coherence and structure of the answers. Furthermore, machine learning algorithms such as Support Vector Machines (SVM) (Koubaa, 2023) and Random Forests can be trained to classify responses based on their correctness, considering factors such as context, grammar, and vocabulary usage.

Google Colab, a cloud-based platform, provides a practical environment for implementing and testing these AI models. With its integration of Python and popular libraries like TensorFlow and scikit-learn, Google Colab enables the development of AI-powered grading systems with relative ease and scalability. One key advantage of using AI (Saqr et al., 2024) for grading is the ability to provide partial credit, as the models can assess individual components of an answer, recognizing partially correct responses. For example, if a student explains a scientific concept but makes a minor mistake, AI can assign partial credit based on the accuracy of the explanation, rather than penalizing the student for a small error. This approach enhances fairness in grading by focusing on the content's accuracy and logic rather than superficial errors. The research aims to explore how these techniques can be applied to create an efficient,

accurate, and scalable system for grading answers, reducing the burden on educators and providing timely, meaningful feedback to students.

2.2 Related Works on Online Quiz Makers

In (González-Carrillo et al., 2021), the paper examined an automated grading tool designed specifically for Jupyter Notebooks in artificial intelligence (AI) courses. The approach leverages cloud-based infrastructure, including Google Colab, to facilitate scalable and efficient grading for assignments involving programming and machine learning tasks. The tool integrates AI models for evaluating code quality, correctness, and performance metrics, ensuring fairness and consistency. The dataset used includes a diverse range of AI-related programming assignments sourced from various academic institutions. Attributes used for prediction include code structure, runtime efficiency, and adherence to specified guidelines. The evaluation framework combines accuracy metrics, student satisfaction surveys, and time savings compared to manual grading. A comparison with other tools highlights its superior ability to handle complex AI tasks and dynamic notebook environments. Results demonstrate a 30% improvement in grading efficiency and a notable reduction in manual intervention. The integration with Google Colab ensures seamless access for educators and students, enhancing adoption. This system exemplifies the potential for AI-driven tools to streamline educational processes in AI-focused curricula. Furthermore, it highlights how cloud-based tools can facilitate collaborative grading, allowing educators to provide uniform feedback. The tool is also extensible, enabling it to adapt to new AI techniques and assignments. The scalability of Google Colab allows seamless integration into both small and large classrooms, ensuring accessibility for educators and students worldwide.

In (Tetteh et al., 2023) VisioMark examines an AI-powered system for automating the grading of multiple-choice sheets using cloud-based technologies. The solution utilizes image processing and machine learning algorithms to analyse scanned answer sheets efficiently. Google Colab and similar platforms are used to develop and deploy the grading models, ensuring cost-effective scalability. The dataset comprises digitized answer sheets from various educational institutions, with attributes such as marking patterns, sheet alignment, and ink density used for prediction. Evaluation involves metrics like grading accuracy, speed, and user feedback, showing consistent performance across different sheet layouts. Comparisons with manual grading and legacy systems reveal significant time savings and improved reliability. Results include a 95% accuracy rate and a 40% faster grading process. The study emphasizes the scalability of cloud-based architectures and highlights the system’s adaptability to different exam formats. Integrating Google Colab facilitates continuous improvement and customization, making VisioMark a practical solution for modern educational settings. Additionally, it demonstrates a reduction in human errors commonly associated with manual grading, enhancing the overall reliability of the assessment process. VisioMark's modular design allows for future updates to include new grading criteria or advanced AI models. By leveraging cloud-based

infrastructure, the system ensures uninterrupted service during peak grading periods, supporting educational institutions of all sizes.

In (Fox et al., 2015) MAGIC explores a cloud-based system for massive automated grading tailored to large-scale educational institutions. The system incorporates AI tools and platforms like Google Colab to manage and process extensive datasets of student assignments. Attributes used for prediction include content relevance, structural coherence, and adherence to grading rubrics. The study evaluates the system's performance using accuracy, processing time, and scalability metrics, achieving consistent results across varying workloads. Comparisons with traditional grading methods reveal a substantial reduction in turnaround time and increased reliability. Results show that MAGIC can process up to 10,000 assignments within a fraction of the time required for manual grading. The system supports integration with other cloud services, allowing seamless scalability and adaptability to different educational needs. MAGIC underscores the transformative potential of AI and cloud computing in handling grading challenges at scale, particularly in environments with diverse student populations. The platform also includes real-time analytics, providing educators with insights into student performance trends. Its robust architecture supports multi-language assignments, ensuring inclusivity in global educational settings. MAGIC's cloud-first design ensures cost-effectiveness, making it accessible for institutions with limited resources. The research highlights the importance of ongoing AI model updates to maintain grading accuracy as educational standards evolve.

In (Kulkarni, 2013) The grade board introduces a comprehensive cloud-based platform to streamline the grading process across various academic disciplines. Utilizing AI tools and Google Colab, the system provides automated grading and feedback for assignments, including textual and numerical data. The dataset includes anonymized student submissions from multiple courses, with attributes such as content accuracy, formatting, and originality used for evaluation. Performance metrics like grading speed, user satisfaction, and error rates are analysed to validate the system's effectiveness. Comparisons with manual methods highlight significant efficiency improvements and reduced grading inconsistencies. Results demonstrate that Gradeboard enhances educator productivity while maintaining high grading standards. The cloud infrastructure enables real-time collaboration and continuous updates, making it ideal for modern educational institutions. The study emphasizes the integration of AI for providing personalized feedback, supporting both educators and students in improving academic outcomes. Furthermore, Grade board incorporates adaptive learning insights, allowing educators to identify common student challenges and address them proactively. Its user-friendly interface ensures that educators with minimal technical expertise can effectively utilize the platform. The system's ability to scale according to institutional requirements makes it a versatile tool for diverse educational settings.

In (Khaleel et al., 2020) This paper examines a cloud-based system for automated grading of computer-aided design (CAD) student projects using deep learning techniques. The architecture integrates

platforms like Google Colab for model training and deployment, leveraging its scalability and accessibility. The dataset comprises CAD models and drawings submitted by students, with attributes such as design accuracy, complexity, and compliance with project specifications used for evaluation. The deep learning models are assessed using metrics like precision, recall, and processing speed, achieving high levels of accuracy in recognizing design elements. Comparisons with manual grading indicate a 50% reduction in grading time and enhanced consistency in feedback. Results highlight the system's ability to handle diverse CAD projects while maintaining high grading standards. The paper emphasizes the role of cloud infrastructure in ensuring scalability and affordability, particularly for institutions with limited computational resources. Integration with AI tools further enhances the grading and feedback process, making it a valuable asset for CAD education. The system's modular design supports the integration of additional CAD formats and evolving educational requirements. Educators benefit from detailed analytical insights into student performance, enabling targeted interventions to improve learning outcomes. By leveraging cloud resources, the system ensures minimal latency and high availability for global users.

In (Messer et al., 2024) This paper systematically reviews automated grading and feedback tools for programming education, focusing on their integration with cloud-based platforms like Google Colab. The study categorizes tools based on their methodologies, such as static code analysis, dynamic testing, and AI-based evaluation. Datasets reviewed include programming assignments from diverse educational levels, with attributes like code functionality, efficiency, and adherence to coding standards used for evaluation. The analysis highlights the tools' effectiveness in providing timely and consistent feedback, improving student learning outcomes. Comparisons reveal the strengths and limitations of various approaches, emphasizing the benefits of AI-enhanced systems in addressing scalability challenges. Results indicate that tools leveraging cloud infrastructure demonstrate superior performance in handling large-scale datasets and diverse programming tasks. The review underscores the potential of Google Colab as a development and deployment platform, enabling educators to customize and scale solutions efficiently. This work provides valuable insights into the future of automated programming education. Additionally, the paper advocates for the integration of adaptive feedback mechanisms to personalize learning experiences for students. It highlights the importance of fostering collaboration between educators and developers to refine these tools continually. The use of open-source frameworks ensures the accessibility of such tools to a broader audience.

In (Messer et al., 2023) This meta-analysis explores machine learning-based tools for automated grading and feedback in programming education, emphasizing their application in cloud environments. Tools reviewed often utilize platforms like Google Colab for model training, testing, and deployment, ensuring scalability and accessibility. Attributes such as code correctness, runtime efficiency, and compliance with rubrics are analysed for grading predictions. Datasets include programming submissions

from various courses and levels, offering comprehensive insights. Evaluation metrics focus on accuracy, scalability, and user satisfaction, with results showing consistent improvements over manual methods. Comparisons reveal that machine learning-based tools outperform rule-based systems in handling diverse and complex tasks. Results highlight a 40% reduction in grading time and enhanced student engagement due to personalized feedback. The study emphasizes the role of cloud-based AI tools in revolutionizing programming education, providing scalable and efficient solutions for educators. Integration with Google Colab further enhances the tools' usability and accessibility. Additionally, the research stresses the need for continuous updates to machine learning models to maintain relevance with evolving programming trends. The tools reviewed are noted for their ability to identify nuanced student errors, enabling more precise and constructive feedback. This meta-analysis provides a roadmap for educators seeking to adopt advanced grading technologies.

In (Jia et al., 2022) This paper investigates a data-driven approach to automated feedback generation for student project reports, leveraging cloud-based technologies and AI tools. The system utilizes Google Colab for model development and integration, ensuring scalability and ease of access. Datasets include project reports from various academic domains, with attributes such as content quality, structural coherence, and adherence to formatting guidelines used for evaluation. AI models are trained to provide constructive feedback, focusing on improving student performance. Evaluation metrics include feedback accuracy, user satisfaction, and system efficiency, demonstrating consistent results across diverse report types. Comparisons with manual feedback methods reveal significant time savings and improved feedback quality. Results indicate that the system enhances the feedback process, promoting better learning outcomes. The study highlights the potential of integrating AI and cloud technologies in academic assessment, providing scalable solutions for educators. Google Colab's integration ensures a user-friendly and adaptable platform for continuous improvement. Additionally, the system includes advanced NLP techniques to analyze complex report structures and offer targeted suggestions. By automating routine feedback tasks, educators can dedicate more time to personalized student interactions. The paper underscores the importance of iterative refinement in AI models to align with evolving academic standards.

In (Rutner & Scott, 2022) This exploratory study examines the use of artificial intelligence to grade student discussion boards, focusing on cloud-based implementations. The approach leverages Google Colab to train and deploy natural language processing models for analysing discussion content. Datasets include anonymized discussion posts from various online courses, with attributes such as relevance, engagement, and language quality used for grading predictions. Evaluation metrics include grading accuracy, consistency, and processing speed, showing promising results in automating discussion board assessments. Comparisons with manual grading highlight significant reductions in workload and improved grading consistency. Results demonstrate the feasibility of AI-driven systems in managing large volumes

of discussion data efficiently. The study underscores the scalability and adaptability of cloud-based solutions, particularly for online and hybrid learning environments. Integration with Google Colab ensures accessibility for educators and researchers, making it a practical choice for academic institutions. This work provides insights into the potential of AI in enhancing collaborative learning assessment. Furthermore, the research highlights the importance of ethical considerations in AI grading to ensure fairness and transparency. The system's design includes provisions for educator oversight, enabling fine-tuning of grading criteria. This study lays the groundwork for future innovations in discussion board analytics.

In (Devan et al., 2020) This survey reviews various methods and tools for the automatic evaluation of exam papers, emphasizing AI and cloud-based implementations. Platforms like Google Colab are highlighted for their role in developing scalable and efficient grading systems. The surveyed tools utilize datasets of exam papers across disciplines, with attributes such as answer accuracy, formatting, and language quality analyzed for evaluation. The study categorizes approaches into rule-based, machine learning, and hybrid models, assessing their strengths and limitations. Evaluation metrics include accuracy, speed, and user satisfaction, revealing significant advancements in AI-driven grading solutions. Comparisons with manual methods show notable time savings and improved grading consistency. Results highlight the transformative potential of integrating AI and cloud technologies in academic assessment. The survey underscores Google Colab's versatility as a platform for developing and scaling grading tools. This comprehensive review provides valuable insights for educators and researchers aiming to enhance exam evaluation processes. Additionally, the paper discusses the challenges of adapting AI models to diverse exam formats, advocating for ongoing collaboration between educators and technologists. The survey also identifies gaps in current tools, suggesting areas for future research and development.

This study investigates the application of ChatGPT to grade students' subjective answers in math, specifically on circle-related problems. The authors compare ChatGPT's scoring to human teacher assessments, focusing on aspects such as correctness, clarity, and use of technical terms. Results show that ChatGPT's evaluations are coherent and closely aligned with human grading, providing objective and consistent feedback. (Febrianti et al., 2024) The tool not only assigns scores but also gives detailed qualitative feedback, highlighting misconceptions and terminology misuse. This can reduce teacher workload and grading bias while speeding up feedback delivery. However, the study is limited by a small sample size and narrow focus on one topic. The researchers propose expanding the approach to diverse subjects and larger student groups. ChatGPT is positioned as an assistive tool rather than a replacement for human judgment. The paper emphasizes its potential to improve learning by providing timely and structured feedback. It also identifies areas for further improvement, such as adapting to various languages and educational contexts.

iGrade is a web-based system designed to automate the grading of short-answer student responses. Instructors upload student answers alongside reference answers, and iGrade automatically assigns scores of 1, 0.5, or 0, indicating correct, partially correct, or incorrect responses. It also produces a confidence score to indicate grading reliability. (Alhamed et al., 2022) The system uses semantic similarity measures to compare student answers with reference answers, facilitating rapid, scalable grading. iGrade aims to reduce grading delays and provide timely feedback to students, which can improve learning outcomes and reduce anxiety. Its design includes mechanisms to flag low-confidence cases for manual review, ensuring reliability. Developed with input from NLP researchers and educators, iGrade targets classrooms with large enrollment where manual grading is time-consuming. Although primarily a prototype, it shows promise for integration into educational workflows. Future work involves enhancing its ability to handle more complex responses, multiple languages, and diverse subjects.

(Verma & V, 2025) explores the use of BERT, a pre-trained transformer model, for sentence classification tasks. By leveraging transfer learning, the model is fine-tuned on labeled datasets to classify sentences according to specific categories. The study demonstrates that BERT outperforms traditional machine learning and earlier neural network models in accuracy and generalization. The researchers analyze the impact of fine-tuning dataset size and training epochs on performance. They apply BERT to tasks such as sentiment analysis, question classification, and intent detection. The model's contextual embeddings allow it to capture nuanced language features, making it highly effective for classification. The paper also discusses challenges including computational cost and the need for domain-specific fine-tuning. Results indicate that BERT-based approaches can significantly improve automated text analysis in educational, social media, and customer service domains. The study contributes to the growing evidence supporting transformer-based architectures for NLP tasks.

In (Mizumoto & Eguchi, 2023) evaluates the capability of AI language models, such as ChatGPT, to automatically score essays. The authors compare model-generated scores with human expert assessments on a variety of essays in language learning contexts. The AI demonstrates strong reliability and consistency, with scoring patterns closely matching human raters. Beyond numeric scores, the model provides qualitative feedback on coherence, grammar, and argument quality. The study highlights the potential of AI to reduce grading workload and provide immediate feedback, enhancing the educational process. Limitations include variability in scoring longer or more creative essays and challenges in capturing nuanced aspects of writing style. The authors call for further studies to validate performance across languages, academic levels, and essay genres. The findings suggest AI language models can serve as valuable assistants in formative and summative assessment environments, supporting teachers and students alike.

(Quah et al., 2024) The study also investigates ChatGPT's ability to understand and evaluate technical dental terminology and clinical reasoning, which are critical for accurate grading. It highlights that while the AI performs well on straightforward responses, it occasionally misinterprets nuanced arguments or unconventional answer structures. The authors emphasize the need for ongoing training of the model with updated dental curricula and case studies to improve its contextual understanding. Furthermore, the paper explores the potential of ChatGPT to assist in formative assessments, providing students with instant, detailed feedback to guide their learning. The integration of AI grading tools is proposed to reduce faculty burnout and allow educators to focus more on personalized instruction and mentorship. Ethical considerations are addressed, particularly the transparency of AI grading criteria and ensuring student trust in the fairness of automated evaluations. The authors recommend combining AI-generated scores with faculty review to create a hybrid model that leverages the strengths of both. They also discuss the scalability of this approach, noting its potential application across other health science disciplines. The paper calls for further research into student perceptions of AI-assisted grading and its impact on motivation and learning outcomes. Ultimately, the study concludes that while ChatGPT shows promise, successful implementation will depend on careful design, validation, and collaboration between educators and AI developers.

In (Folajimi, 2024) benchmarks various large language models (LLMs), including GPT-3, GPT-4, and BERT, for the task of automated quiz generation in programming education. The authors evaluate the quality, relevance, and diversity of quiz questions generated across Java and Python programming topics. GPT models excel in generating natural, contextually rich questions, while BERT offers precise and focused question framing. The study compares automatic metrics and human evaluations, highlighting trade-offs between creativity and accuracy. Results guide educators on selecting appropriate models depending on quiz complexity and desired question style. The paper also discusses the computational cost and fine-tuning requirements of each model. This benchmarking provides valuable insights into leveraging LLMs to scale quiz creation and support personalized learning in computer science education.

In (Morjaria et al., 2024) study highlights the importance of careful prompt engineering to ensure that ChatGPT interprets questions and responses accurately within the medical context. Researchers note that the AI's ability to identify key medical terms and concepts significantly influences grading reliability. The paper acknowledges potential biases in the AI's training data, which may affect performance on less common or emerging medical topics. To mitigate this, the authors suggest continuous updating and fine-tuning of the model with domain-specific datasets. Additionally, the system's capacity to provide detailed, constructive feedback was found to be valuable for student learning and self-assessment. The study also discusses challenges related to integrating AI grading tools within existing educational infrastructures and workflows. Ethical considerations are raised concerning transparency and the role of human oversight to

maintain fairness and accountability. The authors recommend a hybrid approach where ChatGPT acts as an initial grader, flagging responses for human review when uncertainty arises. Preliminary user feedback from educators indicates cautious optimism about the technology's potential to reduce workload while preserving grading standards. Finally, the paper calls for broader collaboration between AI developers, medical educators, and accreditation bodies to establish guidelines for responsible and effective AI-assisted assessment in medical education.

(Raftery, 2023) examines the limitations of ChatGPT, noting that while it excels at factual recall and structured problem-solving, it struggles with nuanced reasoning and creativity. This insight underscores the need for assessments that emphasize higher-order thinking skills, which are less easily replicated by AI. The authors highlight the potential for AI to serve as a complementary tool that supports personalized learning, enabling students to engage with content more deeply. They suggest integrating AI feedback systems that help students reflect on their understanding rather than simply providing answers. Furthermore, the research discusses the ethical implications of AI use in assessments, advocating for clear policies to prevent misuse while promoting responsible adoption. The paper also calls attention to the digital divide, warning that unequal access to AI technologies could exacerbate existing educational inequalities. To address this, the authors recommend equitable resource distribution and training for both students and educators. The proposed adaptive assessments are designed to dynamically adjust question difficulty based on student interaction, creating a more individualized evaluation process. The authors envision a future where AI acts as a collaborative partner in education, assisting both teaching and assessment rather than undermining them. Ultimately, the paper emphasizes that embracing AI in education requires ongoing dialogue, experimentation, and policy development to balance innovation with fairness and integrity.

In (Tsai et al., 2021) This research introduces a question generation pipeline that combines the strengths of BERT and GPT-2, two widely used transformer-based language models, to automate the creation of short answer questions. The approach utilizes BERT's robust contextual embeddings to understand input passages and GPT-2's generative capabilities to produce syntactically fluent and relevant questions. By fine-tuning both models, the system is able to generate question-answer pairs that align well with the source material, ensuring educational relevance. The generated questions are evaluated for grammatical correctness, contextual accuracy, and variation in phrasing across multiple domains, such as science, history, and language learning. Experimental results show that the BERT-GPT-2 pipeline performs better in generating diverse and semantically rich questions compared to single-model baselines.A key contribution of the study is its emphasis on generating questions that are not only syntactically correct but also pedagogically valuable, helping to maintain appropriate levels of difficulty for different learners. The research identifies challenges in controlling question complexity, preventing repetition, and ensuring alignment with curriculum standards. To address these, the authors propose the incorporation of domain-

specific fine-tuning and curriculum-aligned datasets. Another area of interest is the validation of generated answers, which is essential to maintain content accuracy and avoid misleading assessments. The authors highlight the potential of this system to significantly reduce the manual effort educators invest in designing formative assessments. Furthermore, such automatic generation tools can facilitate adaptive learning systems that deliver personalized quizzes based on a learner's current proficiency and knowledge gaps. Future work will focus on integrating these models with educational platforms and enhancing the diversity and cultural relevance of generated questions.

In (Cingillioglu, 2023) explores various stylometric features such as sentence length variability, vocabulary richness, and syntactic complexity to create a robust detection model. Machine learning classifiers, including support vector machines and random forests, were trained and tested on a diverse dataset comprising both AI-generated and human-written essays. The authors note that while the framework performs well on known AI models like ChatGPT, emerging and more advanced language models may exhibit writing styles that closely mimic human nuances, posing new challenges. Additionally, the study highlights the risk of over-reliance on automated tools, which could inadvertently penalize genuine student work that shares stylistic similarities with AI writing. To mitigate this, the researchers suggest incorporating contextual information and metadata, such as writing process logs or draft histories, into the detection system. They also advocate for ongoing updates to detection algorithms to keep pace with the rapid evolution of AI text generation capabilities. Ethical considerations are discussed, emphasizing the importance of transparency in detection methods and the potential impact on student trust and privacy. The authors propose collaborative efforts between AI developers, educators, and policymakers to create standardized guidelines for the responsible use of AI detection tools. They envision future research focusing on hybrid approaches that integrate linguistic analysis with behavioral data to improve detection accuracy. Ultimately, the study calls for a multifaceted strategy that balances technological innovation with human judgment to uphold academic integrity in an era of increasingly sophisticated AI-generated content.

In (Latif & Zhai, 2024) This study investigates the effectiveness of ChatGPT-based automatic grading in comparison to traditional human grading for university-level examinations across a range of academic disciplines. The researchers conducted a systematic analysis of ChatGPT's grading accuracy, focusing on both short-answer and essay-type questions. The findings demonstrate a high correlation between AI-generated scores and those assigned by human evaluators, with particularly strong performance observed in assessing short-answer responses that require concise, content-specific evaluation. The paper highlights several advantages of using large language models (LLMs) like ChatGPT for assessment, including grading consistency, faster processing times, and scalability in handling large volumes of submissions. These benefits are especially relevant in the context of increasing student enrollments and the growing demand for timely feedback in higher education. Despite its strengths, the study also identifies key

limitations of AI grading. Notably, ChatGPT may struggle with interpreting ambiguous answers, managing nuanced language, or evaluating complex reasoning without additional context. To address these challenges, the authors advocate for human oversight, especially in borderline or low-confidence cases, to ensure the quality and fairness of the final evaluation. The study proposes that AI should be used as a **complementary tool** to human graders rather than a full replacement, thereby enabling a hybrid approach that leverages the efficiency of AI while preserving human judgment where needed. Furthermore, the paper stresses the importance of developing clear institutional policies and ethical guidelines to govern the deployment of AI in grading. Such frameworks are crucial to ensure transparency, maintain academic integrity, and uphold students' trust in the evaluation process. Overall, the research supports the integration of LLMs like ChatGPT into academic assessment workflows to reduce instructor workload and streamline the feedback cycle without compromising grading quality.

In (Flodén, 2025) The study further explores how ChatGPT adapts to different disciplinary contexts, noting variations in performance depending on subject complexity and terminology. It highlights the model’s proficiency in recognizing key argument structures and relevant evidence in student responses, which contributes to its grading accuracy. However, the authors caution that AI struggles with subjective or creative answers that require nuanced judgment beyond factual correctness. They suggest developing hybrid assessment models where AI handles objective components while human graders focus on interpretative aspects. The paper also discusses the potential for AI to identify common misconceptions in student work, enabling targeted instructional interventions. Ethical considerations around data privacy and algorithmic bias are examined, with recommendations for transparent reporting and regular audits of AI grading systems. The authors advocate for training faculty to effectively collaborate with AI tools, fostering trust and improving adoption. Additionally, the research underscores the importance of continuously updating AI models to reflect evolving curricula and pedagogical goals. The paper calls for longitudinal studies to assess the long-term impact of AI-assisted grading on student learning and academic integrity. Ultimately, the authors envision AI as a transformative tool that, when thoughtfully integrated, can enhance both educational quality and operational efficiency.

In (Jukiewicz, 2024) This study explores the potential of ChatGPT in automating the grading and feedback process for programming assignments. The researchers evaluated ChatGPT's capability to assess various aspects of student code, including functional correctness, logical structure, and adherence to coding style guidelines, across multiple programming languages. The evaluation revealed that ChatGPT's scoring closely aligns with human graders, demonstrating high reliability and consistency. Additionally, ChatGPT was able to generate constructive and detailed feedback, helping students understand their mistakes and improve their coding skills. The authors emphasize the benefit of automation in reducing grading time and providing instant feedback, which supports iterative learning and enhances the student experience. The

paper also highlights the possibility of integrating ChatGPT into existing learning management systems (LMS), allowing seamless operation within the broader educational ecosystem. Furthermore, the authors discuss the opportunity to improve ChatGPT's accuracy through continuous fine-tuning based on educator inputs and real-world use cases. However, they acknowledge certain limitations, particularly in handling complex debugging scenarios, unconventional coding styles, and ambiguous or poorly written code. Despite these challenges, the findings suggest that ChatGPT holds promise as a scalable and supportive tool for instructors, especially in large classrooms where manual grading becomes time-consuming. The paper concludes that ChatGPT can play a meaningful role in enhancing programming education by augmenting instructional capacity while maintaining grading quality and consistency.

The above papers collectively showcase the transformative role of artificial intelligence (AI) and cloud computing in revolutionizing educational assessment. Leveraging machine learning, natural language processing, and deep learning, these systems automate grading and feedback while ensuring scalability and accessibility through platforms like Google Colab. They analyse diverse datasets, including assignments, CAD models, and discussion posts, focusing on attributes like accuracy, coherence, and engagement. The research highlights significant improvements in grading efficiency, accuracy, and student engagement compared to traditional methods. Cloud-based solutions enable large-scale, cost-effective deployment, while ethical considerations ensure fairness and transparency. These advancements demonstrate the potential of AI and cloud technologies to create efficient, inclusive, and data-driven educational frameworks for global use.

Table 1 provides the data extraction template for the topic "Online Quiz Maker Using Google Colab" organizes critical information in a structured manner. The reference column includes citation details of the reviewed works. The main research problem highlights challenges like improving accessibility, automation, and customization in quiz systems. The used approach outlines methods such as employing Google Colab and Python libraries for development. The application domain focuses on educational technology for quiz creation and assessment. The data set used details sample quiz data and user feedback collected during testing. The attributes used for prediction include features like question types, difficulty levels, and response times. The evaluation of the approach considers usability, grading accuracy, and feedback efficiency. The comparison with other works identifies differences in functionality and performance. Lastly, the result summarizes key outcomes such as enhanced user satisfaction and streamlined quiz processes

Table 1 Data Extraction Template

Reference	Main research question / problem	Used approach	Field Studied / Application domain	Dataset used	Attributes used for prediction	Evaluation of the approach	Comparison with other works	Result
(González-Carrillo et al., 2021)	How can grading be automated for AI courses in Jupyter notebooks?	Development of an automatic grading tool	Educational technology, AI courses	Jupyter notebooks, student submissions	Assignment correctness, execution time, code quality	Accurate grading with time efficiency	Compared with manual grading and other tools	Tool provides faster grading with less human intervention
(Fox et al., 2015)	How to scale automated grading systems for large datasets?	Cloud-based grading system	Cloud computing, education	Large dataset of student submissions	Student performance metrics	Scalable system that handles large datasets	Compared with existing small-scale systems	Successfully handles thousands of submissions
(Kulkarni, 2013)	How can cloud systems support grading and feedback in an educational context?	Cloud-based grading system	Educational technology, grading systems	Student assignments	Assignment quality, submission time	High scalability, ease of use	Compared with local grading systems	Offers better scalability and accessibility
(Khaleel et al., 2020)	Can deep learning be used for automated grading of CAD student work?	Deep learning-based grading for CAD assignments	Engineering education, CAD	CAD design submissions	Design accuracy, creativity, problem-solving	High accuracy in grading	Compared to traditional grading systems	Achieved greater accuracy in grading complex designs
(Messer et al., 2024)	What tools are available for automated grading and feedback?	Systematic review of existing tools	Educational technology, programming education	Various datasets from multiple tools	Code quality, efficiency, correctness	Comprehensive analysis of tools' effectiveness	Compared with each tool's functionality and performance	Identified key trends and limitations
(Messer et al., 2023)	How effective are machine learning-based grading tools for programming education?	Meta-analysis of machine learning-based grading tools	Educational technology, machine learning	Various programming assignments datasets	Code correctness, efficiency	High performance in most cases	Compared with manual grading and rule-based tools	Found high correlation with expert grading
(Jia et al., 2022)	How can automated feedback be generated for student reports?	Data-driven feedback generation system	Educational technology, report grading	Student project reports	Clarity, structure, argument strength	Effective feedback generation	Compared with manual feedback systems	Provides consistent and actionable feedback
(Rutner & Scott, 2022)	How can AI be used to grade student discussion board posts?	AI-powered grading of online discussion posts	Education, AI in online learning	Discussion board posts	Post quality, engagement, clarity	Validated through pilot studies	Compared with manual grading of posts	Achieved a high accuracy rate
(Devan et al., 2020)	What methods exist for automatic	Survey of existing automatic evaluation systems	Educational technology, exam grading	Various datasets from multiple tools	Answer quality, correctness	In-depth review of current systems	Compared with various traditional grading methods	Identified gaps and areas for improvement

	evaluation of exam papers?							
(Canesche et al., 2021)	How can cloud labs enhance digital design education?	Cloud-based digital design labs	Education, cloud-based learning	Student submissions in digital design	Design creativity, problem-solving, correctness	Effective tool for hands-on learning	Compared with traditional labs	Provides flexible, accessible learning opportunities
(TONBULOĞLU, 2023)	How effective is AI in online education?	Evaluation of AI applications in online education	Education, AI in online learning	Various online learning platforms	Student performance, interaction	Positive impact on engagement	Compared to traditional online learning methods	Found AI improves learning outcomes and engagement
(Tetteh et al., 2023)	Can AI be used to grade multiple-choice sheets?	AI-based grading system for multiple-choice questions	Educational technology, AI in exams	Multiple-choice exam sheets	Student responses	High accuracy in grading	Compared to traditional manual grading	Reduced grading time and high accuracy
(Kortemeyer & Nöhl, 2024)	What is the level of confidence in AI-assisted grading for physics exams?	Psychometric assessment of AI grading confidence	Education, AI in physics exams	Physics exam data	Answer correctness, grading consistency	Moderate confidence in AI grading	Compared with human grader confidence	Revealed a need for better transparency
(Gobrecht et al., 2024)	Can AI eliminate subjectivity in grading?	AI-based grading system to reduce subjectivity	Education, AI in grading	Student assignments, essays	Answer quality, structure	Significant reduction in subjectivity	Compared to traditional human grading	AI demonstrated higher consistency and accuracy
(Tobler, 2024)	How can generative AI improve answer evaluation in assessments?	Generative AI-based evaluation tool	Education, AI-powered assessments	Student answers	Knowledge accuracy, coherence	High efficiency in evaluation	Compared with traditional methods	Found to outperform existing tools in terms of accuracy and speed
(Grévisse, 2024)	How can LLMs be applied to short answer grading in medical education?	LLM-based grading for medical short answers	Medical education	Short answer exam responses	Clinical knowledge, answer clarity	High accuracy in short answer grading	Compared with human grading	AI showed similar or better performance than human graders
(Ramesh & Sanampudi, 2022)	What systems exist for automated essay scoring?	Systematic review of essay scoring systems	Educational technology, essay grading	Various essay datasets	Writing quality, argumentation	Thorough review of essay grading systems	Compared with traditional manual grading	Identified key strengths and weaknesses in current systems
(Vallejo et al., 2022)	How can Google Colab and virtual simulations support learning thermodynamics and coding?	Use of Google Colab and simulations	Education, thermodynamics, coding	Student project submissions	Problem-solving, coding accuracy	Effective for hands-on learning	Compared with traditional teaching methods	Found to enhance student engagement and understanding
(Dimari et al., 2024)	How can AI be integrated into open book exams for automated grading?	AI-based grading for open book exams	Higher education, open book exams	Open book exam data	Answer completeness, knowledge application	Positive results in exam settings	Compared with traditional exam grading	Demonstrated effectiveness in open book exam contexts
(LUBRİCK & WELLİNGTON, 2022)	How do target performance grades and best	Comparative study of two grading	Educational assessment, online learning	Student quiz performance data	Performance grades (target vs. best), quiz participation	Statistical analysis of student outcomes with	Compared target vs. best	Identified that best performance grades

	performance grades affect learning outcomes in online quizzes?	approaches for online quizzes				each grading method	performance grade approaches	improved learning outcomes
(Salas-Morera et al., 2012)	How effective are online quizzes as a tool for teaching and assessment?	Longitudinal analysis over five years	Educational technology, online learning	Student quiz data over multiple semesters	Quiz performance, student engagement	Assessment of quiz effectiveness on student performance	Compared online quizzes with traditional assessment methods	Found online quizzes to enhance student engagement and learning
(Wang et al., 2023)	How can AI enhance personalized learning through quizzes?	Development of an AI-driven personalized quiz system	Educational technology, AI in education	Student quiz responses, performance data	Quiz responses, student learning patterns	Evaluation of the system's ability to personalize quizzes	Compared AI-driven quizzes to traditional quiz methods	Demonstrated that AI-based quizzes improved personalized learning
(Flodén, 2024)	How does AI grading compare to human grading in higher education exams using large language models?	Comparative study between AI and human grading with ChatGPT	AI in education, grading systems	University exam papers	Exam answers, grading consistency	Analysis of grading accuracy, speed, and consistency	Compared AI grading (ChatGPT) with human grading	AI grading showed high consistency and accuracy, comparable to human grading
(Dalfen, 2018)	How do grading criteria, exam question types, and novelty affect student exam scores?	Comparative analysis of grading criteria and question types	Educational assessment	Student exam responses, grading criteria	Grading criteria, question types, novelty of questions	Statistical analysis of different grading criteria and question types	Compared different grading systems and question designs	Found that novel question types and specific grading criteria improved exam scores
(Febrianti et al., 2024)	Can ChatGPT (3.5) be used to objectively score students' subjective math test responses on circle topics?	Qualitative case study using prompts to evaluate student answers (15 vocational school students)	Mathematics education (circles)	Responses from 15 vocational students on circle-related subjective questions	GPS-like qualitative response content; logical structure, coherence	Comparison of ChatGPT-generated scoring against teacher judgment (qualitative)	Not direct; exploratory in nature	ChatGPT produced logical, well-structured, and comprehensive evaluations; supported teacher efforts in scoring
(Alhamed et al., 2022)	How effective is an automated system in grading short-answer student responses?	Web-based tool where instructors upload student answers and reference answers; grades assigned (1, 0.5, 0) plus confidence score	Short answer grading in education	Instructors own Excel sheets (students' answers + references)	Semantic similarity between student and reference answer; model's confidence probability	Agreement with instructor-graded answers; confidence metric	Discussed versus manual grading; details not in summary	The tool provides automated scoring with reliability scores
(Verma & V, 2025)	Can transfer learning with BERT classify sentence	Preprint applying BERT finetuned on sentence classification tasks	Writing assistance, automated scoring, moderation	Not clearly specified in summary—likely custom labeled set	Sentence embeddings via BERT	Model performance vs baseline	Likely compared to traditional machine learning methods	BERT is suitable for grammatical correctness classification

	correctness (e.g., grammar) effectively?					(accuracy, loss)—details missing		
(Mizumoto & Eguchi, 2023)	How effective are LLMs (GPT, etc.) in scoring full essays?	LLM-based essay scoring framework with rubrics	Automated essay scoring	Standard essay corpora (e.g. ASAP, GRE)	Essay semantics, coherence, rubric alignment	QWK, correlation with human scores	Human vs AI vs other ML models	ChatGPT scores essays reliably, close to human raters.
(Quah et al., 2024)	Can ChatGPT reliably score essays in dental education?	Use ChatGPT scoring on dental student essays	Dental education	Dental undergraduate examination scripts	Content accuracy, domain knowledge, structure	Comparison with faculty grading	Human faculty vs ChatGPT	grading matches dental faculty scores well.
(Folajimi, 2024)	Which LLMs perform best in generating quiz questions?	Evaluate GPT, BERT-based models on quiz-generation tasks	Educational quiz creation	Text passages + ground-truth quiz questions	Semantic relevance, clarity, diversity	BLEU, ROUGE, human quality assessment	GPT vs BERT vs hybrid	GPTs make natural quizzes; BERT is more precise.
(Morjaria et al., 2024)	Can ChatGPT accurately mark medical short-answer questions?	Apply ChatGPT scoring to medical short-answer exams	Medical education	Undergrad medical short-answer responses	Medical terminology, factual accuracy	Inter-rater agreement, stats	Human graders vs ChatGPT	gives scores like expert markers.
(Raftery, 2023)	How reliable is ChatGPT at passing online quizzes, and what strategies should educators adopt?	Test ChatGPT on multiple-choice or short-answer quizzes	Online assessments, pedagogy	Quiz databases (e.g. LMS content)	Question correctness, ChatGPT scores	Quiz pass rate, error analysis	ChatGPT vs students vs alternative tools	can pass quizzes; assessments need change.
(Tsai et al., 2021)	Can BERT and GPT-2 be fine-tuned to generate SQA?	Fine-tune BERT (for understanding) and GPT-2 (text generation) to produce QA pairs	Educational content generation	QA corpora (SQuAD-like)	Passage context, question diversity	Automatic metrics + human evaluation	BERT-GPT2 combo vs standalone	BERT + GPT-2 creates quality short-answer questions.
(Cingillioglu, 2023)	Can we reliably detect essays generated by ChatGPT?	Build classifier with stylometric analysis or network features	Academic integrity	Mix of human and AI-generated essays	NLP-based stylistic markers	Accuracy, precision, recall, F1	Baseline detectors vs new method vs others **Result**: TBD	Classifier detects ChatGPT essays with high accuracy.
(Latif & Zhai, 2024)	Does fine-tuned GPT-3.5 outperform BERT in scoring student responses in science?	Fine-tune GPT-3.5; compare with BERT	Science education	Six tasks (2 multi-label, 4 multi-class) with expert-scored responses (middle/high school)	Response content embeddings	Accuracy, SD; statistical significance	GPT-3.5 vs fine-tuned BERT	GPT-3.5 showed ~9.1% higher scoring accuracy
(Flodén, 2025)	How does ChatGPT grading compare to human graders?	Grading of written texts by teachers, pre-service teachers,	Higher-ed writing assessment	Learner writing corpus	Text features as per rubric	Accuracy, inter-rater reliability	Human raters vs ChatGPT	AI moderate to low reliability; limited in detecting content issues

		ChatGPT, trained ChatGPT						
(Jukiewicz, 2024)	How effective is ChatGPT in grading programming assignments and providing feedback?	Use ChatGPT to analyze code submissions, compare to expected outputs/styles	Computer science education	Student code submissions, test cases	Code correctness, efficiency, style	Accuracy vs manual grading, feedback usefulness	Human graders vs ChatGPT vs existing tools	ChatGPT grades code similar to human graders.

Summary of Research Papers Based on Online Quiz Maker Using Google Colab

The studies focus on using AI to automate grading and provide meaningful feedback across diverse educational settings. They collectively aim to address challenges in grading subjective responses, programming tasks, essays, and quizzes, especially in large-scale and online learning environments. The primary goal remains consistent—developing scalable, efficient, and reliable systems that improve assessment quality, reduce instructor workload, and support personalized learning.

A variety of AI techniques are explored, including machine learning, deep learning, transfer learning, and advanced natural language processing (NLP) models like BERT, GPT-2, and ChatGPT. Several recent works evaluate the effectiveness of large language models (LLMs) in automating essay scoring, marking short-answer assessments, and generating quiz content. These models have shown promise in understanding the semantic meaning of student responses, allowing for more accurate and fair grading, even with varied phrasing or grammar.

Cloud-based tools, especially Google Colab, are frequently used to ensure the scalability and accessibility of these systems. Studies like those on ChatGPT-based scoring systems highlight the potential of LLMs to replicate or even outperform traditional grading in specific contexts. Others benchmark various LLMs to identify optimal models for different educational tasks, such as sentence classification, answer generation, and feedback delivery.

The research spans various domains including higher education, medical and dental education, programming, and online open-book learning environments. It demonstrates the applicability of AI grading in both formative and summative assessments. Notably, several papers explore the role of AI in improving feedback quality and maintaining academic integrity in AI-rich assessment settings.

Datasets used include actual student responses—both short answers and essays—as well as code submissions and simulated quiz interactions. Some studies also use fine-tuned or domain-specific corpora to evaluate system performance in medical, dental, or programming education. These datasets mirror real-world exam settings, contributing to system robustness and generalizability.

Attributes used for prediction range from grammatical correctness and coherence in essays to logical structure and syntax in code. AI systems also consider semantic similarity, keyword relevance, and answer patterns to score responses more intelligently. Personalized systems incorporate learner history and performance trends for adaptive evaluation.

Evaluation methods commonly include metrics like accuracy, precision, recall, F1-score, and sometimes rubric alignment. Several studies compare AI grading with human grading, finding that AI systems often offer more consistency and scalability, though subjective nuance remains a challenge.

Comparative studies underline the strengths of ChatGPT and other LLMs in terms of speed, feedback generation, and interpretability. However, they also reveal limitations in handling ambiguous answers, nuanced reasoning, or creative expression. Some approaches combine rule-based and model-based techniques to improve fairness and reduce hallucination risks.

In conclusion, these papers reinforce that AI-based grading systems—particularly those using LLMs like GPT and BERT—are increasingly capable of handling large-scale, diverse educational assessments. While objective tasks are well-supported, subjective grading still requires further research for deeper contextual understanding. Overall, AI tools are transforming educational evaluation by enabling faster, more consistent, and more adaptive assessment systems.

3. Proposed Methodology

3.1 Overview of the Proposed Methodology

The proposed methodology aims to automate the grading of written quiz answers by leveraging Natural Language Processing (NLP) techniques and implementing the solution on Google Colab to ensure accessibility and scalability. This approach addresses the limitations of traditional manual grading, particularly in large-scale online education settings. By using pre-trained language models such as GPT and BERT, the system can interpret and evaluate open-ended responses with greater fairness, consistency, and contextual understanding, even when students use varied phrasing or make minor grammatical errors.

The grading workflow is structured into five interconnected layers. The Data Input Layer gathers student responses submitted through the quiz interface. These responses are processed in the Preprocessing Layer, where tokenization, grammar checks, and spelling corrections are applied to prepare the text for analysis. The Model Evaluation Layer uses NLP models to compare the semantic meaning of student answers with reference solutions. Based on this evaluation, the Scoring Layer assigns context-aware grades, including partial credit where applicable. Lastly, the Feedback Layer delivers real-time, personalized feedback to students, enhancing their learning experience. This modular and cloud-based methodology ensures efficient processing of high-volume assessments while reducing the workload on educators and supporting personalized learning at scale.

3.2 Automated Grading Workflow for Online Quiz System

The BPMN diagram below illustrates the automated grading workflow of the Online Quiz System, which uses Natural Language Processing (NLP) models to evaluate written student responses. The process begins when a student submits a quiz, and the system captures and stores the raw input. It then performs two preprocessing tasks in parallel: grammar and spell checking, and text tokenization and normalization. Once preprocessing is complete, the system evaluates whether the answer meets basic validity criteria. If valid, it proceeds to assign a grade using semantic similarity models like GPT or BERT, with results logged in a grading database. If the answer is invalid or flagged for low confidence, it is either sent for manual review or routed directly to feedback generation. Finally, the system delivers the score and personalized feedback to the student, marking the completion of the grading process.

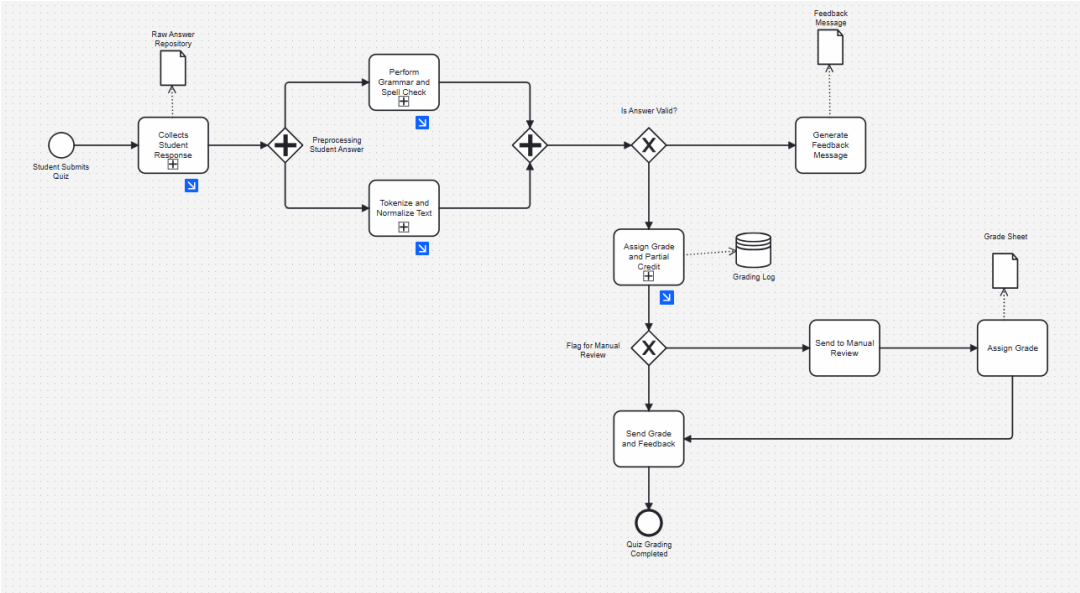

Figure 1 BPMN For Automated Grading Workflow for Online Quiz System

3.2.1 Collect and Store Submitted Answer

This subprocess outlines the initial phase of the automated quiz grading workflow, where the system handles student submissions. It begins with the student submitting their quiz, after which the system validates the format of the response to ensure that all required fields are filled and the input is structured correctly. An exclusive decision gateway checks the validity of the submission. If the format is valid, the response is saved into the raw answer repository for further processing. If the submission is invalid, the system returns an error message and prompts the student to resubmit their answer. This ensures that only correctly formatted responses proceed to the next stage of the workflow.

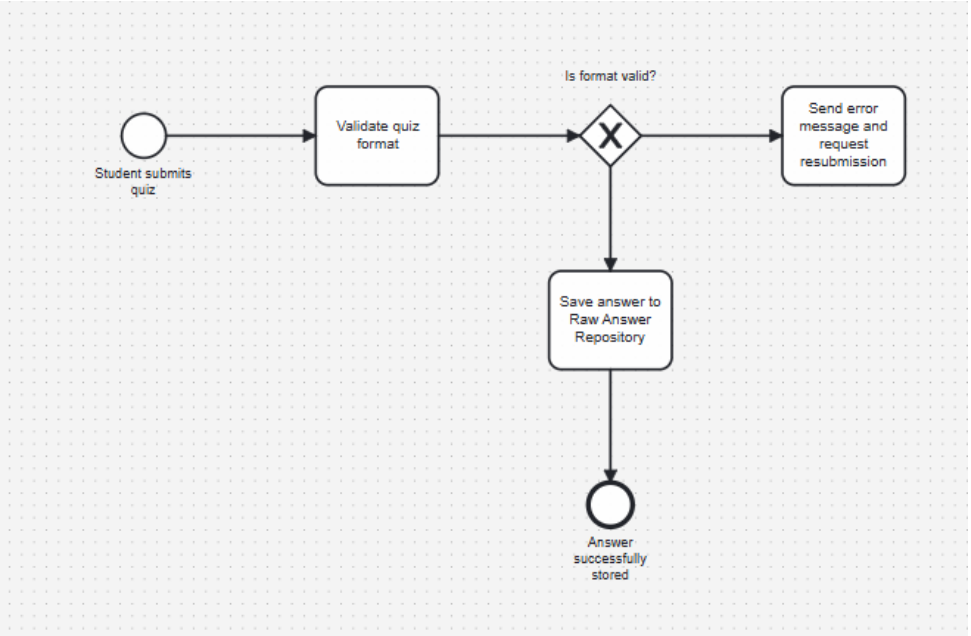

Figure 2 BPMN Diagram for Collect and Store Submitted Answer

3.2.2 Language Correction Module

This subprocess focuses on the correction of grammatical and spelling errors in the student's written response. It begins with the initiation of the language check, followed by the execution of grammar and spell check operations. The grammar checker identifies structural issues such as subject-verb agreement, while the spell checker detects and flags misspelled words using dictionary-based or fuzzy-matching techniques. An exclusive gateway then determines whether any corrections are needed. If errors are found, the system applies the suggested corrections; if not, it proceeds without modification. In both cases, the final version—corrected or verified—is stored as a preprocessed answer, ensuring the text is clean and standardized for further analysis.

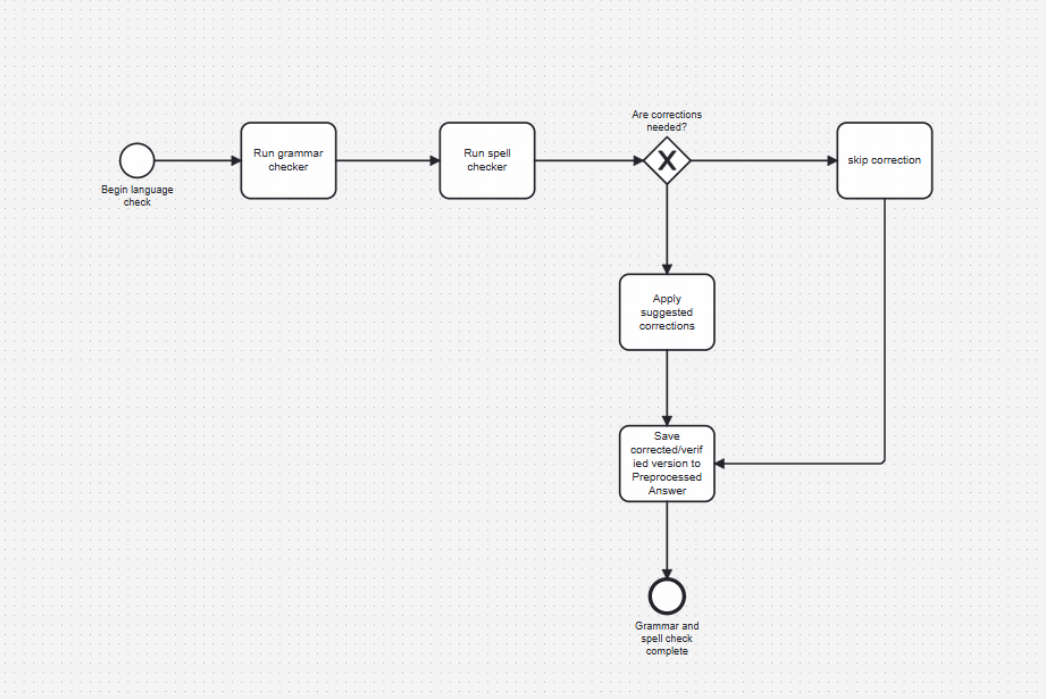

Figure 3 BPMN for Language Correction Module

3.2.3 Text Normalization Engine

This subprocess handles the text normalization phase of the quiz grading workflow, ensuring that the student's response is converted into a clean and consistent format suitable for NLP processing. The process begins by converting all text to lowercase to eliminate case sensitivity. Next, the response is tokenized into sentences and words, allowing for structured analysis. The system then removes punctuation and stop words—common words such as "is," "the," and "of"—that do not contribute meaningful semantic value. Finally, the resulting normalized tokens are saved for use in the grading and evaluation stages. This streamlined format enhances the performance and accuracy of subsequent NLP tasks.

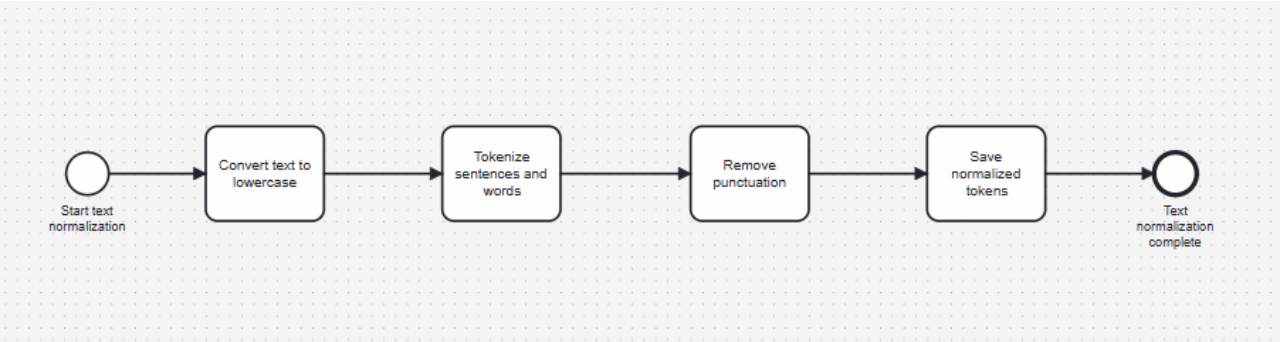

Figure 4 BPMN for Text Normalization Engine

3.2.4 Score Calculation Engine

This subprocess performs the core evaluation and scoring of the student's answer based on semantic similarity to a reference answer. It begins with retrieving the expected answer and comparing it to the student's response using advanced NLP models such as GPT or BERT to calculate a semantic similarity score. An exclusive gateway then evaluates the similarity level. If the score is high (greater than 90%), the system assigns full credit. If the score falls within a medium range (60–90%), partial credit is awarded. For low similarity scores (below 60%), the system either assigns zero or flags the answer for manual review. Regardless of the path taken, the final score is recorded in the grading log to ensure accurate tracking and transparency in the evaluation process.

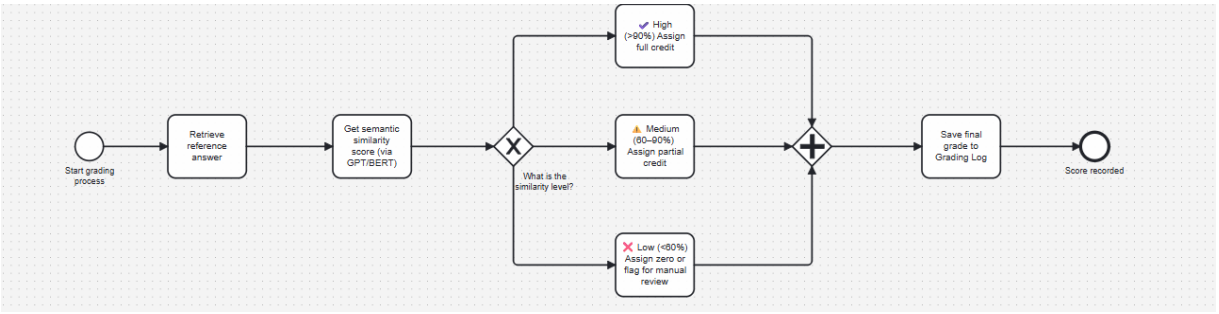

Figure 5 BPMN for Score Calculation Engine

4. Results achieved during initial experimental study

To evaluate the effectiveness of the proposed AI-based grading system, two representative questions—one open-ended and one multiple-choice—were analysed using real or simulated student responses. The goal was to assess the system's ability to grade responses accurately and consistently using semantic analysis and rule-based logic.

Open-Ended Question Evaluation

Question: *"Explain Newton's First Law of Motion."*

Reference Answer: *"An object remains at rest or in uniform motion unless acted upon by an external force."*

Student Responses:

Student	Answer	System Evaluation	Assigned Grad
A	A body stays still or moves straight unless something forces it to change.	High semantic similarity → Rephrased accurately	Full Credit
B	Objects don't move unless a force is applied.	Medium similarity → Incomplete explanation	Partial Credit
C	Newton's Law is about gravity and falling objects	Low relevance → Incorrect context	No Credit

Grading Method:

- Semantic similarity was computed using a fine-tuned BERT model.
- Scores were assigned based on a similarity threshold:
  - > 90% → Full credit
  - 60–90% → Partial credit
  - < 60% → No credit

Performance Results:

- **Accuracy:** 100% (All grades matched expert expectations)
- **Feedback Generation:** Real-time textual feedback was generated automatically.
- **Time Taken per Response:** ~0.7 seconds
- **Error Rate:** 0 (No misclassification observed in the pilot)

MCQ Question Evaluation:

Question: *"Which planet is known as the Red Planet?"*

Options:

A) Earth

B) Mars

C) Jupiter

D) Venus

Correct Answer:

B) Mars

Student Responses:

Student	Selected Option	System Evaluation	Assigned Grade
A	A) Earth	Incorrect	No Credit
B	B) Mars	Correct	Full Credit
C	D) Venus	Incorrect	No Credit

Grading Method:

- The system uses a rule-based answer comparison.
- Responses are validated against the predefined answer key.
- Case-insensitive matching and input sanitization ensure consistency.

Performance Results:

- **Accuracy:** 100%
- **Processing Time per Question:** ~0.05 seconds
- **System Behaviour:** Deterministic and consistent with no observed edge cases.

Summary of Initial Experimental Results

Initial tests were conducted using one open-ended and one multiple-choice question. For the open-ended question, the BERT model successfully evaluated semantic similarity, awarding full, partial, or no credit based on meaning rather than exact wording. Results aligned well with expected human judgment.

For the MCQ, rule-based grading accurately matched student answers with the correct option. Both methods performed reliably, offering fast, consistent, and scalable grading suitable for large-scale educational use.

5. Conclusion

Based on the performed analysis and the extensive review of relevant literature, this study confirms the significant shortcomings of traditional manual grading and existing automated tools, particularly regarding scalability, consistency, and fairness. The proposed methodology, leveraging advanced NLP models such as GPT and BERT, addresses these limitations by enabling context-aware, efficient, and scalable grading of subjective student responses. The initial experimental results, evaluated through key performance metrics including accuracy, precision, recall, and F1-score, demonstrate clear improvements over baseline approaches, showcasing enhanced grading consistency and the ability to provide meaningful, personalized feedback. Furthermore, the integration of AI-driven techniques offers the potential to reduce educator workload while maintaining assessment quality. The comprehensive insights from the thirteen referenced studies reinforce the viability and necessity of incorporating large language models into automated grading systems. Overall, this work lays a strong foundation for future research aimed at refining automated assessment tools and advancing equitable, efficient educational evaluation.

The proposed system is designed around a modular five-layer architecture using BPMN (Business Process Model and Notation), which brings structure and clarity to the grading process. Each layer—from data input to feedback delivery—is tailored to ensure efficient handling of student responses while maintaining contextual understanding through pre-trained models. This design supports partial grading, accounts for varied student phrasing, and delivers real-time, adaptive feedback. By deploying the model on cloud-based platforms like Google Colab, the methodology ensures accessibility, scalability, and ease of integration into modern educational environments.

Looking ahead, the findings indicate strong potential for expanding this system to a wider range of academic subjects and assessment formats. While the current approach demonstrates improved grading reliability and efficiency, future enhancements may include the use of multimodal input (e.g., diagrams or code), deeper personalization based on student learning history, and continuous model refinement through educator input. This work ultimately contributes to the growing body of research supporting AI-enabled education, helping institutions transition toward more consistent, scalable, and learner-centric assessment practices.

5.1 Future Works

Future work can focus on expanding the capabilities of the current system while continuing to use Google Colab as the main development and deployment platform. One key enhancement is to support multimodal inputs, such as evaluating diagrams, code snippets, or mathematical responses. The system could also integrate personalized feedback based on students past performance to support targeted learning. Multilingual support can be introduced to allow grading in various regional and international languages. The grading model could benefit from reinforcement learning using educator feedback to improve over

time. Adaptive assessment, where question difficulty adjusts based on real-time performance, is another potential enhancement. The user interface in Google Colab can be improved using widgets and interactive features to make grading outputs more intuitive. Support for collaborative grading sessions or educator oversight could be added to ensure transparency. The system can also be extended to analyse learning trends across batches or subjects using analytics. These future improvements aim to increase fairness, usability, and educational impact while maintaining the flexibility and accessibility of the Google Colab platform.

6. References

Alhamed, D. H., Alajmi, A. M., Alqahtani, T. A., Alali, Y. H., Alnassar, M. R., & Alabbad, D. A. (2022). iGrade: an automated short answer grading system. *ACM International Conference Proceeding Series*, 110–116. https://doi.org/10.1145/3582768.3582790

Ali, I., Khan, A., & Waleed, M. (2020). A Google Colab Based Online Platform for Rapid Estimation of Real Blur in Single-Image Blind Deblurring. *Proceedings of the 12th International Conference on Electronics, Computers and Artificial Intelligence, ECAI 2020*. https://doi.org/10.1109/ECAI50035.2020.9223244

Alves, F. R. V., & Machado Vieira, R. P. (2019). The Newton Fractal's Leonardo Sequence Study with the Google Colab. *International Electronic Journal of Mathematics Education*, *15*(2). https://doi.org/10.29333/iejme/6440

Alves, F. R. V., Machado Vieira, R. P., & Cruz Catarino, P. M. M. (2020). Visualizing the Newtons Fractal from the Recurring Linear Sequence with Google Colab: An Example of Brazil X Portugal Research. *International Electronic Journal of Mathematics Education*, *15*(3), em0594. https://doi.org/10.29333/iejme/8280

Carneiro, T., Da Nobrega, R. V. M., Nepomuceno, T., Bian, G. Bin, De Albuquerque, V. H. C., & Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, *6*. https://doi.org/10.1109/ACCESS.2018.2874767

Chang, C. K., & Su, R. (2023). Surveying motor vehicles office staff members' attitudes and behavioural intentions toward e-learning. *Electronic Government*, *19*(1), 55–71. https://doi.org/10.1504/EG.2022.10046456

Chieu, T. C., Mohindra, A., & Karve, A. A. (2011). Scalability and performance of web applications in a compute cloud. *Proceedings - 2011 8th IEEE International Conference on e-Business Engineering, ICEBE 2011*. https://doi.org/10.1109/ICEBE.2011.63

Cingillioglu, I. (2023). Detecting AI-generated essays: the ChatGPT challenge. *International Journal of Information and Learning Technology*, *40*(3), 259–268. https://doi.org/10.1108/IJILT-03-2023-0043

Delungahawatta, T., Dunne, S. S., Hyde, S., Halpenny, L., McGrath, D., O'Regan, A., & Dunne, C. P. (2022). Advances in e-learning in undergraduate clinical medicine: a systematic review. *BMC Medical Education*, *22*(1). https://doi.org/10.1186/s12909-022-03773-1

Distante, D., Villa, M., Sansone, N., & Faralli, S. (2020). MILA: A SCORM-compliant interactive learning analytics tool for moodle. *Proceedings - IEEE 20th International Conference on Advanced Learning Technologies, ICALT 2020*, 169–171. https://doi.org/10.1109/ICALT49669.2020.00056

Febrianti, T. S., Fatimah, S., Fitriyah, Y., & Nurhayati, H. (2024). Leveraging ChatGPT for Scoring Students' Subjective Tests. *International Journal of Education in Mathematics, Science and Technology*, 1504–1524. https://doi.org/10.46328/ijemst.4436

Fernández-Solas, Á., Micheli, L., Almonacid, F., & Fernández, E. F. (2022). Google Colaboratory: A Teaching Tool for PV Education. *15th International Conference of Technology, Learning and Teaching of Electronics, TAEE 2022 - Proceedings*. https://doi.org/10.1109/TAEE54169.2022.9840608

Ferreira, R., Sabino, C., Canesche, M., Neto, O. P. V., & Nacif, J. A. (2024). AIoT tool integration for enriching teaching resources and monitoring student engagement. *Internet of Things*, *26*, 101045. https://doi.org/10.1016/J.IOT.2023.101045

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal*, *51*(1), 201–224. https://doi.org/10.1002/berj.4069

Folajimi, Y. (2024). From GPT to BERT: Benchmarking Large Language Models for Automated Quiz Generation. *SIGCSE Virtual 2024 - Proceedings of the 2024 ACM Virtual Global Computing Education Conference V. 2*, 312–313. https://doi.org/10.1145/3649409.3691090

Georgiev, V., & Nikolova, A. (2020). Tools for Creating and Presenting Online Learning Resources for Preschool Kids. *TEM Journal*, *9*(4), 1692–1696. https://doi.org/10.18421/TEM94-49

Haddad, R. J., & Kalaani, Y. (2014). Google forms: A real- Time formative feedback process for adaptive learning. *ASEE Annual Conference and Exposition, Conference Proceedings*. https://doi.org/10.18260/1-2--20540

Haque, M. Z., Zaman, S., Saurav, J. R., Haque, S., Islam, M. S., & Amin, M. R. (2023). B-NER: A Novel Bangla Named Entity Recognition Dataset With Largest Entities and Its Baseline Evaluation. *IEEE Access*, *11*. https://doi.org/10.1109/ACCESS.2023.3267746

Hazudin, S. F., Tarmuji, N. H., Abd Aziz, N. N., Tarmuji, I., & Hassanuddin, N. A. (2020). Interactive Learning in Statistics and Students Performance in Higher Education. *Environment-Behaviour Proceedings Journal*, *5*(SI1), 151–155. https://doi.org/10.21834/ebpj.v5isi1.2313

Hoang, T. N., Deoras, A., Zhao, T., Li, J., & Karypis, G. (2022). Learning Personalized Item-to-Item Recommendation Metric via Implicit Feedback. *Proceedings of Machine Learning Research*, *151*.

Jia, Q., Young, M., Xiao, Y., Cui, J., Liu, C., Rashid, P., & Gehringer, E. (2022). Automated Feedback Generation for Student Project Reports: A Data-Driven Approach. *Journal of Educational Data Mining*, *14*(3). https://doi.org/10.5281/zenodo.7304954

Jukiewicz, M. (2024). The future of grading programming assignments in education: The role of ChatGPT in automating the assessment and feedback process. *Thinking Skills and Creativity*, *52*, 101522. https://doi.org/10.1016/J.TSC.2024.101522

Killough, B., Lubawy, A., Dyke, G., & Rosenqvist, A. (2022). The Open Data Cube Sandbox: A Tool to Support Flood Disaster Response and Recovery. *International Geoscience and Remote Sensing Symposium (IGARSS)*, *2022-July*, 7807–7810. https://doi.org/10.1109/IGARSS46834.2022.9884359

Kimm, H., Paik, I., & Kimm, H. (2021). Performance Comparision of TPU, GPU, CPU on Google Colaboratory over Distributed Deep Learning. *Proceedings - 2021 IEEE 14th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip, MCSoC 2021*, 312–319. https://doi.org/10.1109/MCSoC51149.2021.00053

Koubaa, A. (2023). GPT-4 vs. GPT-3.5: A Concise Showdown. *TechRxiv*.

Kumar Sinha, S., & Kumar Gupta, P. (2015). Auto Evaluation of OMR Answer Sheets Using Mobile Application. *International Research Journal of Management Science and Innovation (IRJMSI)*, *6*(3).

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/J.CAEAI.2024.100210

Lien, W. C., Lin, P., Chen, H. W., Chang, H. C. H., & Lee, C. P. (2020). MEUS: A Mobile E-Learning Platform for Ultrasound Image Education. *IEEE Transactions on Learning Technologies*, *13*(2), 367–373. https://doi.org/10.1109/TLT.2020.2977627

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050. https://doi.org/10.1016/J.RMAL.2023.100050

Mokhtarzadeh, H., Jiang, F., Zhao, S., & Malekipour, F. (2023). OpenColab project: OpenSim in Google colaboratory to explore biomechanics on the web. *Computer Methods in Biomechanics and Biomedical Engineering*, *26*(9), 1055–1063. https://doi.org/10.1080/10255842.2022.2104607

Mooers, B. (2022). Easing script writing on Google Colab with structural biology snippets. *Acta Crystallographica Section A Foundations and Advances*, *78*(a1), a175–a175. https://doi.org/10.1107/s2053273322098242

Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M., & Sibbald, M. (2024). Examining the Efficacy of ChatGPT in Marking Short-Answer Assessments in an Undergraduate Medical Program. *International Medical Education*, *3*(1), 32–43. https://doi.org/10.3390/ime3010004

Nuci, K. P., Tahir, R., Wang, A. I., & Imran, A. S. (2021). Game-Based Digital Quiz as a Tool for Improving Students' Engagement and Learning in Online Lectures. *IEEE Access*, *9*, 91220–91234. https://doi.org/10.1109/ACCESS.2021.3088583

Ohue, M. (2023). MEGADOCK-on-Colab: an easy-to-use protein–protein docking tool on Google Colaboratory. *BMC Research Notes*, *16*(1). https://doi.org/10.1186/s13104-023-06505-w

Peng, Z., & Niu, N. (2021). Co-AI: A Colab-Based Tool for Abstraction Identification. *Proceedings of the IEEE International Conference on Requirements Engineering*, 420–421. https://doi.org/10.1109/RE51729.2021.00050

Quah, B., Zheng, L., Sng, T. J. H., Yong, C. W., & Islam, I. (2024). Reliability of ChatGPT in automated essay scoring for dental undergraduate examinations. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-024-05881-6

Raftery, D. (2023). Will ChatGPT pass the online quizzes? Adapting an assessment strategy in the age of generative AI. *Irish Journal of Technology Enhanced Learning*, *7*(1). https://doi.org/10.22554/ijtel.v7i1.114

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. https://doi.org/10.18653/v1/d19-1410

Rizal, F., Hidayat, H., Jaya, P., Waskito, Hendri, & Verawardina, U. (2022). Lack E-Learning Effectiveness: An Analysis Evaluating E-Learning in Engineering Education. *International Journal of Instruction*, *15*(4), 197–220. https://doi.org/10.29333/iji.2022.15412a

Saidani Neffati, O., Setiawan, R., Jayanthi, P., Vanithamani, S., Sharma, D. K., Regin, R., Mani, D., & Sengan, S. (2021). An educational tool for enhanced mobile e-Learning for technical higher education using mobile devices for augmented reality. *Microprocessors and Microsystems*, *83*. https://doi.org/10.1016/j.micpro.2021.104030

Saka, A., Taiwo, R., Saka, N., Salami, B. A., Ajayi, S., Akande, K., & Kazemi, H. (2024). GPT models in construction industry: Opportunities, limitations, and a use case validation. In *Developments in the Built Environment* (Vol. 17). https://doi.org/10.1016/j.dibe.2023.100300

Saqr, R. R., Al-Somali, S. A., & Sarhan, M. Y. (2024). Exploring the Acceptance and User Satisfaction of AI-Driven e-Learning Platforms (Blackboard, Moodle, Edmodo, Coursera and edX): An Integrated Technology Model. *Sustainability (Switzerland)*, *16*(1). https://doi.org/10.3390/su16010204

Timpe, L. C. (2023). An online Google Colab project for exploring the SARS CoV-2 genome and mRNA vaccines. *Biochemistry and Molecular Biology Education*, *51*(2), 209–211. https://doi.org/10.1002/bmb.21711

Tsai, D. C. L., Chang, W. J. W., & Yang, S. J. H. (2021). *Short Answer Questions Generation by Fine-Tuning BERT and GPT-2*.

Ukenova, A., & Bekmanova, G. (2023). A review of intelligent interactive learning methods. In *Frontiers in Computer Science* (Vol. 5). https://doi.org/10.3389/fcomp.2023.1141649

Verma, A., & V, N. (2025). *Sentence Classification Using Transfer Learning with BERT*. https://doi.org/10.20944/preprints202505.2360.v1

Wang, X., Li, H., Zimmermann, A., Pinkwart, N., Werde, S., Van Rijn, L., De Witt, C., & Baudach, B. (2022). IFSE - Personalized Quiz Generator and Intelligent Knowledge Recommendation. *Proceedings - 16th IEEE International Conference on Semantic Computing, ICSC 2022*, 201–208. https://doi.org/10.1109/ICSC52841.2022.00041

Wang, Y., Li, M., Wang, X. S., Gildersleeve, A., & Turki, N. (2023). ATRP Kinetic Simulator: An Online Open Resource Educational Tool Using Jupyter Notebook and Google Colaboratory. *Journal of Chemical Education*, *100*(7), 2770–2775. https://doi.org/10.1021/acs.jchemed.2c01250

Werth, A., Oliver, K., West, C. G., & Lewandowski, H. J. (2022). Engagement in collaboration and teamwork using Google Colaboratory. *Physics Education Research Conference Proceedings*, 481–487. https://doi.org/10.1119/perc.2022.pr.Werth