# DEVELOPING AN AI-ASSISTED GRADING SYSTEM USING LARGE LANGUAGE MODELS (GPT-4)

by

Andrei Modiga

A PROJECT PROPOSAL

Presented to the Faculty of

The School of Computing at the Southern Adventist University

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Anderson

Collegedale, Tennessee

August, 2024

# DEVELOPING AN AI-ASSISTED GRADING SYSTEM USING LARGE LANGUAGE MODELS (GPT-4)

Andrei Modiga, M.S.

Southern Adventist University, 2024

Adviser: Scot Anderson, Ph.D.

Grading written student work is a critical component of education, directly influencing the learning experience and providing essential feedback to both students and educators. This project focuses on building an AI-assisted grading system that leverages Large Language Models (LLMs), specifically GPT-4, to automate and enhance the grading process. By classifying student responses based on semantic similarity, the system streamlines grading and allows educators to provide more consistent and efficient feedback. The project aims to reduce the manual workload on teachers, improve grading consistency, and provide timely feedback to students.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

In the educational landscape, grading written student work is a task of high importance; it directly influences the learning experience and provides crucial feedback to both students and educators. Traditional grading methods, while effective, can be time-consuming and labor-intensive, especially for teachers managing large volumes of student submissions. The demand for efficient and scalable grading solutions has become increasingly evident as educators seek to streamline their workloads without compromising assessment quality.

Recent advancements in artificial intelligence, particularly in the development of Large Language Models (LLMs), offer a promising approach to this challenge. LLMs like OpenAI's GPT-4 have demonstrated remarkable proficiency in understanding and generating human-like text by analyzing vast datasets and recognizing patterns in language. These models have the potential to transform the grading process by automating the evaluation of student responses, thereby reducing the workload on teachers.

This project explores the development of an AI-assisted grading system that incorporates GPT-4 to classify student responses by grouping those that are semantically similar. By identifying clusters of answers that convey the same meaning, regardless of phrasing or structure, the system can assist educators in grading more efficiently.

The primary goal of this project is to build and implement this AI-assisted grading system, including the ability to recognize handwritten text, and integrate it into an educational platform like OICLearning.com. By focusing on building a practical solution, the project aims to enhance the grading process, making it quicker and more consistent, thereby allowing teachers to focus on more critical aspects of instruction while ensuring that students receive fair and accurate feedback.

However, integrating LLMs into the grading process presents challenges that must be carefully addressed. Key considerations include the model's ability to accurately differentiate nuanced meanings and appropriately group responses that, while semantically similar, may vary in correctness. Additionally, the effectiveness of these models in handling a diverse range of student expressions and potential errors needs thorough evaluation.

# Chapter 2

# Background

The integration of artificial intelligence (AI) into education has revolutionized teaching and learning processes. A significant area of impact is the automation of grading, where AI technologies, particularly Large Language Models (LLMs) like GPT-4, have shown great promise. This chapter delves into the advancements in AI-assisted grading, focusing on GPT-4's role, and discusses existing research, challenges, and the potential benefits of such systems.

## 2.1 Artificial Intelligence in Education

Artificial intelligence has become a significant part of modern education, offering tools and solutions that enhance learning experiences and streamline administrative tasks. AI-driven educational platforms can personalize learning paths, adapt content to individual student needs, and provide real-time feedback [1]. This personalization is crucial in accommodating diverse learning styles and improving student engagement.

Moreover, AI systems can analyze vast amounts of data to identify patterns and trends in student performance. Teachers can use these insights to tailor instruction, address knowledge gaps, and improve overall educational outcomes. The automation of administrative tasks, such as scheduling and attendance tracking, allows teachers to focus more on instruction and less on paperwork [2].

### 2.1.1 Personalized Learning and Feedback

One of the most significant contributions of AI in education is the ability to offer personalized learning experiences. Intelligent tutoring systems powered by AI can adjust the difficulty level of exercises based on a student's performance, ensuring optimal learning progression [1]. These systems can also identify areas where a student is struggling and provide targeted interventions.

In the context of grading, AI can provide detailed feedback on assignments, highlighting specific areas of strength and weakness. GPT-4, for instance, can generate personalized comments that help students understand their mistakes and learn from them. This immediate and individualized feedback is more effective than generic responses and can significantly enhance the learning process [7].

### 2.1.2 Administrative Efficiency

AI's ability to automate routine administrative tasks is transforming the educational landscape. Tasks such as grading, scheduling, and record-keeping can be handled efficiently by AI systems, reducing the workload on educators [1]. This automation not only saves time but also minimizes human errors that can occur in manual processes.

Furthermore, AI systems can monitor student engagement and participation, alerting educators to potential issues such as decreased performance or absenteeism. By identifying these problems early, interventions can be implemented promptly, supporting student retention and success [13].

## 2.2 Advancements in Large Language Models

### 2.2.1 Introduction to GPT-4

GPT-4, developed by OpenAI, is one of the most advanced LLMs available today. Building upon the success of its predecessors, GPT-4 has been trained on an extensive dataset comprising diverse textual content, enabling it to understand and generate human-like text with remarkable fluency and coherence [1].

The architecture of GPT-4 allows it to capture complex language patterns, understand context, and infer meaning beyond surface-level interpretations. This deep understanding is particularly beneficial in educational applications where more than simple comprehension of student responses is required.

### 2.2.2 GPT-4 in Educational Applications

GPT-4's capabilities extend to various educational applications. It can assist in content creation, such as generating lesson plans, educational materials, and practice questions tailored to specific learning objectives [1]. In language learning, GPT-4 can provide conversational practice, corrections, and explanations to learners.

In grading, GPT-4's ability to understand and evaluate open-ended responses makes it a valuable tool. It can analyze essays, short answers, and problem-solving explanations, providing not only grades but also constructive feedback. This level of assessment requires a deep understanding of the subject matter and the ability to interpret diverse student expressions [8].

## 2.3 AI-Assisted Grading Using GPT-4

### 2.3.1 Automating Short Answer Grading

Short answer questions are challenging to grade automatically due to the variability in student responses. Traditional grading systems rely on keyword matching or predefined answer patterns, which can miss correct answers phrased differently or accept incorrect answers that contain the right keywords.

GPT-4 overcomes these limitations by understanding the semantic meaning behind the text. It can recognize correct answers even when they are expressed in unconventional ways and identify incorrect answers that superficially appear correct [6]. This semantic understanding allows for more accurate grading of open-ended questions.

Liu et al. (2023) [6] conducted a study demonstrating GPT-4's effectiveness in grading university-level mathematics exams. The AI model was able to evaluate complex mathematical reasoning and provide grades that closely aligned with human graders. This study highlights GPT-4's potential in handling subjects that require critical thinking and problem-solving skills.

### 2.3.2 Grading Handwritten Responses

Grading handwritten assignments poses additional challenges due to the need for accurate handwriting recognition. Optical Character Recognition (OCR) technologies have improved, but they still struggle with illegible handwriting or complex symbols commonly used in subjects like mathematics and physics.

Kortemeyer (2023) [3] explored the feasibility of using AI to grade handwritten physics solutions. By integrating OCR technologies with GPT-4, the study achieved a significant correlation between AI-generated grades and human grading. However, the study also noted limitations in accurately interpreting diagrams and notations, indicating areas where further improvement is needed.

### 2.3.3 Semantic Understanding and Contextual Evaluation

GPT-4's advanced natural language processing capabilities enable it to understand context and semantics deeply. It can evaluate not just the correctness of an answer but also the reasoning process behind it. This is important in subjects where the method is as important as the final answer, such as mathematics and science.

For example, GPT-4 can assess a student's problem-solving approach, identify logical errors, and provide specific feedback on where the reasoning went astray [1]. This level of detailed evaluation helps students understand their mistakes and learn more effectively.

## 2.4 Challenges in AI-Assisted Grading with GPT-4

### 2.4.1 Accuracy and Reliability

While GPT-4 demonstrates high proficiency in understanding and evaluating text, it is not immune to errors. Misinterpretations can occur, especially with ambiguous or poorly constructed responses. Liu et al. (2023) [6] emphasized the importance of human oversight to ensure the reliability of AI-generated grades.

Moreover, AI models may sometimes be overconfident in their assessments, potentially leading to incorrect grading. Implementing mechanisms for uncertainty estimation and flagging ambiguous cases for human review can mitigate this issue.

### 2.4.2 Bias and Fairness

AI models can inadvertently exhibit biases present in their training data. This raises concerns about fairness, as certain groups of students might be disadvantaged by biased grading. Tossell et al. (2023) [11] highlighted the importance of addressing these biases to ensure equitable treatment of all students.

Strategies to mitigate bias include using diverse and representative training data, implementing fairness-aware algorithms, and regularly auditing AI systems for discriminatory patterns [9].

### 2.4.3 Handwriting Recognition Limitations

Despite advancements in OCR technologies, accurately recognizing handwritten text remains challenging. Variations in handwriting styles, the use of non-standard symbols, and poor scan quality can hinder recognition accuracy.

Kortemeyer et al. (2023) [4] found that AI struggled with interpreting hand-drawn diagrams and complex notations, which are common in STEM subjects. Improving OCR algorithms and combining them with context-aware models like GPT-4 can enhance performance, but human intervention may still be necessary in some cases.

### 2.4.4 Ethical and Privacy Concerns

The use of AI in grading involves processing sensitive student data, raising ethical and privacy considerations. Compliance with regulations such as the Family Educational Rights and Privacy Act (FERPA) is essential.

Alto (2023) [1] emphasizes the need for robust data protection measures, transparency in AI operations, and obtaining informed consent from users. Ensuring that AI systems are secure and that data is handled responsibly is crucial for maintaining trust.

## 2.5 Existing Research on AI Grading Systems

### 2.5.1 Machine Learning Approaches

Before the advent of LLMs like GPT-4, machine learning approaches to grading primarily involved training models on labeled datasets to recognize correct answers. Weegar and Idestam-Almquist (2022) [12] explored such methods for grading short answers in computer science exams.

Their research demonstrated that machine learning could significantly reduce grading workload by clustering similar answers and automating scoring. However, these systems often required extensive preprocessing and were limited in handling the full variability of human language.

### 2.5.2 Feasibility Studies with GPT-4

Studies utilizing GPT-4 have shown promising results in various subjects. Kortemeyer (2023) [3] conducted a feasibility study on using GPT-4 for grading physics problems. The AI's grades correlated highly with human graders, indicating its potential effectiveness.

However, the study also noted that GPT-4 sometimes missed nuanced aspects of the solutions, particularly in assessing the quality of reasoning and the appropriateness of assumptions. This suggests that while GPT-4 can assist in grading, it should complement rather than replace human judgment.

### 2.5.3 Student Perceptions of AI Grading

The acceptance of AI-assisted grading by students is crucial for its successful implementation. Tossell et al. (2023) [11] examined student perceptions of using AI tools like ChatGPT in academic settings.

While students appreciated the promptness and consistency of AI feedback, they expressed concerns about the transparency of the grading process and the AI's ability to understand their unique perspectives. Building trust requires clear communication about how the AI operates and opportunities for students to contest or discuss their grades [13].

## 2.6 Potential Benefits of GPT-4 in Grading

### 2.6.1 Efficiency and Scalability

Implementing GPT-4 in grading systems can significantly reduce the time educators spend on evaluating assignments. This is especially beneficial in large classes where manual grading is impractical. AI systems can handle vast amounts of data quickly, providing timely feedback to students [1].

### 2.6.2 Consistency and Fairness

AI grading systems apply the same criteria uniformly, reducing inconsistencies that can arise from human graders' subjective judgments or fatigue. This consistency contributes to fairness, ensuring that all students are evaluated on the same basis [5].

### 2.6.3 Timely Feedback

Prompt feedback is essential for effective learning. AI systems can provide immediate evaluations, allowing students to understand their performance and address any issues while the material is still fresh in their minds [10].

### 2.6.4 Enhanced Learning Outcomes

By analyzing patterns in student responses, AI systems can identify common misconceptions and areas where many students struggle. Educators can use this information to adjust their teaching strategies, focus on problematic topics, and improve overall learning outcomes [13].

## 2.7 Summary

The integration of GPT-4 into grading systems offers significant advantages, including increased efficiency, consistency, and the ability to provide personalized feedback. While challenges such as accuracy, bias, and ethical considerations exist, ongoing research and development are addressing these issues.

By combining GPT-4's capabilities with human oversight and ethical practices, AI-assisted grading can enhance educational experiences for both students and educators. The potential for improved learning outcomes and reduced workloads makes this an important area of development.

## 2.8 Conclusion

The existing body of research highlights the transformative potential of AI-assisted grading using GPT-4. By understanding the capabilities and limitations of these systems, educators can implement AI tools that complement their teaching practices.

The proposed AI-assisted grading system aims to build upon this foundation, addressing current challenges and harnessing GPT-4's full potential. By doing so, it seeks to improve the grading process, enhance learning experiences, and contribute to the advancement of AI in education.

# Chapter 3

# Proposal

This chapter outlines the key requirements for developing an AI-assisted grading system using Large Language Models (LLMs) like GPT-4 and GPT-4 Vision, implemented in Python. The goal is to automate the grading process for various types of assignments on OICLearning.com by utilizing GPT-4 Vision to read images directly, performing both Optical Character Recognition (OCR) and Natural Language Processing (NLP) in a unified model. Additionally, the system will incorporate grading algorithms to assign grades based on expected answers for each question. This integration simplifies the workflow and enhances the system's ability to accurately interpret and grade student submissions.

# Appendix A

# Requirements Specification

# Bibliography

[1] Valentina Alto. *Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4*. Packt Publishing Ltd, 2023. 2.1, 2.1.1, 2.1.2, 2.2.1, 2.2.2, 2.3.3, 2.4.4, 2.6.1

[2] Li Chen, Gang Chen, and Xiaodong Lin. Artificial intelligence in education: A review. *IEEE Access*, 8:75264–75278, 2020. 2.1

[3] Gerd Kortemeyer. Toward ai grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research*, 19(2):020163, 2023. 2.3.2, 2.5.2

[4] Gerd Kortemeyer, Julian Noh, and Daria Onishchuk. Grading assistance for a handwritten thermodynamics exam using artificial intelligence: An exploratory study. *arXiv preprint arXiv:2306.17859*, 2023. 2.4.3

[5] Robert Könnecke and Torsten Zesch. Automated scoring of content and style in short essays. *Frontiers in Education*, 5:90, 2020. 2.6.2

[6] Tianyi Liu, Julia Chatain, Laura Kobel-Keller, Gerd Kortemeyer, Thomas Willwacher, and Mrinmaya Sachan. Ai-assisted automated short answer grading of handwritten university level mathematics exams. *arXiv preprint arXiv:2308.11728*, 2023. 2.3.1, 2.4.1

[7] Rosemary Luckin, Wayne Holmes, Mark Griffiths, and Laurie Brooks Forcier. Intelligence unleashed: An argument for ai in education. *Pearson Education*, 2016. 2.1.1

[8] Benjamin D. Lund, Daniel Wang, Kyle Chao, and Matthew S. Gerber. Chatgpt's ability to provide timely, novel, and impactful ideas for research. *arXiv preprint arXiv:2305.06566*, 2023. 2.2.2

[9] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 2.4.2

[10] David J. Nicol and Debra Macfarlane-Dick. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218, 2006. 2.6.3

[11] Chad C. Tossell, Nathan L. Tenhundfeld, Ali Momen, Katrina Cooley, and Ewart J. de Visser. Student perceptions of chatgpt use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Education*, 2023. 2.4.2, 2.5.3

[12] Rebecka Weegar and Peter Idestam-Almquist. Reducing workload in short answer grading using machine learning. *International Journal of Artificial Intelligence in Education*, 32:611–643, 2022. 2.5.1

[13] Olaf Zawacki-Richter, Victoria I. Marín, Melissa Bond, and Franziska Gouverneur. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):39, 2019. 2.1.2, 2.5.3, 2.6.4