

MASTER THESIS

by

Andrei Modiga

A THESIS

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

MASTER THESIS

Andrei Modiga, M.S.

Southern Adventist University, 2024

Adviser: Scot Anderson, Ph.D.

In the educational landscape, grading written student work is a task of high importance, directly influencing the learning experience and providing crucial feedback to both students and educators. This thesis explores the use of Large Language Models (LLMs) in automating and enhancing the grading process.

Contents

Contents	5
1 Introduction	7
2 Methodology	9
3 Results	11
4 Conclusion	13

Chapter 1

Introduction

In the educational landscape, grading written student work is a task of high importance, directly influencing the learning experience and providing crucial feedback to both students and educators. While traditional grading methods are generally effective, they can be time-consuming and labor-intensive, particularly for teachers managing large volumes of student responses. The need for efficient and scalable grading solutions has become increasingly evident as educators seek to streamline their workloads without compromising the quality of their assessments.

Recent advancements in artificial intelligence, particularly in the development of Large Language Models (LLMs), offer a promising approach to this challenge. LLMs, such as GPT-4, have shown remarkable proficiency in understanding and generating human-like text by analyzing vast datasets and recognizing patterns in language. This research explores the potential of LLMs not only to recognize correct answers but also 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, LLMs can assist educators in grading more efficiently while maintaining the integrity of their evaluations.

The primary goal of this research is to test and analyze two different LLM models, OpenAI's GPT-4 and Meta's LLaMA, to determine their effectiveness in simplifying the grading process for teachers, including the ability to recognize handwritten text. By comparing these models, the study aims to identify which one better categorizes student responses based on semantic similarity, providing insights into their strengths and limitations. This approach seeks 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, the integration of LLMs into the grading process presents certain challenges that must be carefully addressed. One of the key considerations is the models' ability to accurately differentiate between nuanced meanings and appropriately group responses that, while semantically similar, may vary in their correctness. Additionally, the effectiveness of these models in handling a diverse range of student expressions and potential errors needs to be thoroughly evaluated.

Chapter 2

Methodology

Chapter 3

Results

Chapter 4

Conclusion