Enhancing Automated Grading Systems in College Educational Environments through GPT Models for Text Summarization

Omar Basheer
Ashesi University

1. Introduction

The use of open-ended questions in college assessments poses a significant challenge due to the complexity involved in grading them accurately, yet efficiently. Manual grading is often time-consuming, repetitive, and prone to subjectivity. While Computer Assisted Assessment (CAA) systems have attempted to automate this process, they are designed for multiple choice questions, limiting their effectiveness in evaluating critical thinking skills associated with open-ended responses. This limitation is particularly pronounced in college institutions where emphasis is placed on the development of independent, critical thinking skills necessary for future professional and social interactions.

2. Background

2.1 Learning Assessment Overview

- The assessment is integral to the learning process, aiding educators in gauging the effectiveness of their teaching methods (Hasanah et al., 2018).
- Manual grading is marred by drawbacks such as monotony, repetitiveness, and time inefficiency (Hasanah et al., 2018).
- Computer Assisted Assessment (CAA) systems offer automated assessment but are limited to multiple-choice questions (Hasanah et al., 2018).
- Multiple-choice questions promote memorization and factual recall (Nicol, 2007).
- A broader range of assessment tools is needed to capture important learning goals and to more directly connect assessment to ongoing instruction" (Shepard et al., 2001).
- Open-ended questions foster critical thinking skills necessary for future contexts, challenging to assess automatically (Ku, 2009).

2.2 Proposal: AI-driven Text Summarization

- AI-driven text summarization, specifically using GPT models, presents a promising solution for efficient and accurate grading (Widyassari et al., 2022).
- GPT models, like Generative Pre-trained Transformers, offer high-quality text generation, impacting various applications (Nilsson & Tuvstedt, n.d.).

3. Objective and Research Questions

Objective

- To examine the implementation of GPT-driven text summarization in the automated grading of answers to open-ended college assessment questions. The primary goal is to provide empirical insights into the efficiency and accuracy of this innovative approach

Research Questions

- To what extent does GPT-driven text summarization contribute to the efficiency of automated grading in comparison to traditional grading methods?
- How does the accuracy of automated grading using GPT-driven text summarization compare to human grading standards in evaluating open-ended college assessment answers?

4. Literature Review

4.1 Definition of Concepts

- Text summarization is the process of producing a summary of a particular text that contains important sentences and all relevant information from the original document (Widyassari et al., 2022).

4.2 Role of Artificial Intelligence (AI)

- Teaching a computer the same content as students allows for the computer to assess the knowledge of the students (Gardner et al., 2021).
- The biggest challenge faced by AI in grading open-ended questions is the existence of complex trins in human writing (Gardner et al., 2021).

4.3 Focus on Generative Pretrained Transformer (GPT) Models

- The major applications of GPT include language translation, text summarization, responding to questions, and essay writing (Tajik & Tajik, 2023), but its functionalities can be fine-tuned to specific applications (Rudolph et al., 2023).

4.4 Text Summarization Techniques

- Abstractive summarization is more efficient than extractive summarization due to its ability to develop new sentences to tell the important information from the text (Moratanch & Chitrakala, 2016).
- Humans are unable to distinguish between real summaries and those produced by ChatGPT (Soni & Wade, 2023).

4.5 Evaluation of Existing Automated Systems

- One system provided suggestions to instructors by re-using previous comments as well as scores (Bernius et al., 2020), but was limited by the population of tutors and their experience levels used in the study.
- Another system sought to provide feedback to software professionals in training, by comparing their evaluated answers to open-ended questions (Pinto et al., 2023). The answers were evaluated by by expert developers and ChatGPT, but the study was limited by the number of topics and professionals employed.

4.6 Gap Analysis

- There is a need to investigate the scalability and generalizability of these models in the context of grading open-ended questions in various academic disciplines.

5. Methodology

Research Method:
Quantitative Approach

1. Participants:
   - Use purposive sampling to target students from diverse academic backgrounds within the chosen institution.
   - Identify instructors who have experience in grading open-ended questions.

2. Data Collection:
   - Collect a set of open-ended assessment questions from multiple courses across various subjects to ensure representativeness.
   - Choose questions covering different levels of complexity and diverse content areas.
   - Anonymously collect responses to the selected open-ended questions from students.
   - Acquire rubrics used by instructors for grading the selected open-ended questions

GPT-Driven Text Summarization:
- Select a GPT model tailored for text summarization tasks, considering compatibility with research objectives (e.g., GPT-3.5).
- Input the selected open-ended questions into the GPT model for abstractive text summarization.
- Generate summaries and ensure that the GPT model is focused on summarizing key details from the open-ended responses.
- Compare the generated summaries with the rubric, assigning scores based on the inclusion of relevant details.

Comparative Analysis:
- Compare the grades assigned by the GPT model with the original human-graded scores provided by instructors.
- Conduct a question-by-question analysis to assess the model's competence in grading specific open-ended questions.
- Use similarity metrics (e.g., percentage agreement, weighted kappa statistics) to quantify agreement between GPT-generated scores and human-graded scores.
- Employ correlational analyses to determine the overall consistency between GPT-generated scores and human-graded scores across the sampled questions.

Acknowledgements

I express my sincere gratitude to my esteemed lecturers for their invaluable guidance and to my peers for their unwavering support throughout this research journey

References

Bernius, J. P., Kovaleva, A., Krusche, S., & Bruegge, B. (2020). Towards the Automation of Grading Textual Student Submissions to Open-ended Questions. Proceedings of the 4th European Conference on Software Engineering Education, 61–70. https://doi.org/10.1145/3396802.3396805

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?' Journal of Computer Assisted Learning, 37(5), 1207–1216. https://doi.org/10.1111/jcal.12577

Hasanah, U., Astuti, T., Wahyudi, R., Rifai, Z., & Pambudi, R. A. (2018). An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian. 2018 3rd International Conference on Information Technology, Information System and Electrical Engineering (ICITISEE), 230–234. https://doi.org/10.1109/ICITISEE.2018.8720957

Ku, K. Y. L. (2009). Assessing students' critical thinking performance: Urging for measurements using multi-response format. Thinking Skills and Creativity, 4(1), 70–76. https://doi.org/10.1016/j.tsc.2009.02.001