Research Project Proposal

Automating University-Level Grading with LLMs: A Feasibility Study of AI-Enhanced Educational Assessment

By

Hendrik Matthys van Rooyen

University of London

March 2024

CONTENTS

CHAPTER 1 INTRODUCTION AND BACKGROUND

1.1 INTRODUCTION

The evolution of academic evaluation is increasingly intertwined with advances in data science and artificial intelligence (AI), notably through the deployment of Large Language Models (LLMs) like OpenAI's GPT-4 for automated grading at the university level. This novel integration aims to augment the grading efficiency and enrich the feedback provided to students, leveraging data science to analyse intricate datasets such as student submissions and grading criteria. This study endeavours to assess the practicality and effectiveness of AI in educational assessments, focusing on the accuracy, consistency, and scalability of LLMs for grading diverse subject matters.

This exploration is driven by the potential to address prevalent educational challenges, including the time-intensive nature of manual grading, the variability in feedback quality, and the growing demand for personalized, prompt evaluations. Through data science and AI, the initiative seeks to transcend mere automation, aspiring to enhance the educational journey for students and educators by showcasing how AI can revolutionize academic assessments.

1.2 MOTIVATION

The motivation behind this project arises from the persistent challenges in academia, especially in grading and feedback provision within university contexts. The traditional grading process demands substantial time from educators, detracting from their capacity to engage in other critical academic responsibilities. Moreover, the imperative for quality feedback for student development juxtaposes with the inherent subjectivity and variability in grading essays and open-ended questions.

Proposing the use of LLMs, such as GPT-4, this project aims to automate grading and elevate feedback quality across various disciplines at the university level. By inputting essential elements like the grading rubric, questions, student responses, and the maximum score, the LLM is designed to generate both a grade and constructive feedback. This approach not only seeks to streamline grading, freeing educators for other duties, but also to enrich feedback quality with detailed, consistent, and objective evaluations, thereby enhancing students' learning experiences with precise, actionable insights into their academic performance.

CHAPTER 2 PROJECT DEFINITION

2.1 AIM(S) AND OBJECTIVES

This project's primary goal is to explore the viability and benefits of using cutting-edge Large Language Models (LLMs) for automating grading and feedback on university exams across various fields. It aims to assess LLMs' ability to grade text responses accurately and offer constructive feedback, aiming to make educational evaluations more efficient and improve feedback quality for students. The project outlines specific goals:

1. **Evaluate Grading Accuracy and Reliability**: This goal involves comparing grades from an LLM, like OpenAI's GPT-4, with human graders across different subjects, focusing on the LLM's score consistency over multiple assessments to ensure grading reliability.
2. **Incorporate Afrikaans Language Support**: Acknowledging academic linguistic diversity, particularly where Afrikaans is widely used, this goal seeks to implement Afrikaans language understanding and grading capabilities in the LLM, ensuring it can accurately grade and respond in Afrikaans.
3. **Create a Structured Data Exchange System**: To make this solution practically applicable, there's a need to devise a system for inputting structured data (grading rubrics, questions, student answers) and extracting useful outputs (grades, feedback) that can mesh well with current IT systems, addressing the automation challenges of the grading process.
4. **Design a Scalable Infrastructure**: Anticipating widespread adoption, this goal focuses on developing an infrastructure capable of scaling, potentially utilizing cloud technologies like AWS Lambdas and queue systems to manage a large assessment volume efficiently and cost-effectively.

By achieving these goals, the project intends to offer a thorough examination of the feasibility, advantages, and limitations of using LLMs for university test grading and feedback, with an emphasis on accuracy, system integration, linguistic inclusivity, and scalability.

2.2 STAKEHOLDERS

1. **Students** are central to this project, poised to gain from quicker, more uniform, and thorough feedback on their work, potentially boosting their academic journey and outcomes.
2. **Educators**, including faculty, lecturers, and assistants, would directly benefit from a reduced grading burden, enabling more focus on curriculum enhancement, research, and student interaction.
3. **University administration teams** overseeing academic affairs, curriculum, and IT infrastructure would play a pivotal role in adopting and backing the tech solutions for grading.
4. **IT and educational technology specialists** are key for weaving and maintaining the tech framework essential for the smooth operation of the LLM grading system, ensuring its efficiency and security.
5. **Language communities**, particularly those focusing on Afrikaans, due to the project's inclusion of Afrikaans support, affecting students and teachers in Afrikaans-dominant education settings.
6. **Education policy makers**, given their influence over educational standards and AI tech adoption in education, might find the project's results regarding grading equity and accessibility significant.
7. **Educational researchers and innovators** would be keen on the project’s insights, methodologies, and tech progress in the realms of edtech, AI's educational applications, and teaching strategies.

CHAPTER 3 LITERATURE REVIEW

3.1 RELATED WORKS

The limited existing papers examined reveal various insights regarding the employment of large language models (LLMs) for grading purposes, highlighting strengths, and identifying gaps in limited testing samples, scope, the level of questions, and the utilization of multilingual data.

Limited Testing Samples: A common limitation across studies is the reliance on limited testing samples for evaluating LLMs' grading effectiveness. For instance, Nilsson and Tuvstedt (2023) conducted experiments on a select number of student submissions from specific years of programming courses, focusing on a few assignments that could be easily parsed by GPT. This approach restricts the breadth of the evaluation and may not fully represent the LLMs' grading capabilities across a wide range of subjects or levels of education.

Limited Scope: The scope of these studies is often narrowly defined, focusing on grading accuracy without examining the quality of feedback provided by GPT or the broader implications of employing GPT in grading across different courses. For example, Henkel et al. (2023) specifically limited their investigation to the grading accuracy of GPT-4, comparing it to grades previously assigned by teaching assistants, without delving into feedback quality or exploring the feasibility of using LLMs for comprehensive course assessment.

Level of Questions: The studies generally focus on university-level courses, with examples including introductory programming assignments and tasks requiring plain text answers and passing unit tests. While this provides valuable insights into the LLMs' capabilities in higher education settings, it leaves a gap in understanding how well these models can adapt to grading assignments of varying complexity levels, including more advanced or specialized topics.

Limited Multilingual Data: The reviewed papers do not explicitly address the utilization of multilingual data for grading purposes. This omission highlights a gap in research regarding the performance of LLMs in assessing student submissions across different languages, which is crucial for the global applicability of LLM-based grading systems. The importance of exploring this area is underscored by the diverse linguistic backgrounds of students in global educational settings.

Recognizing Satisfactory Approaches: Some studies successfully demonstrate the potential of LLMs, such as GPT-4, to match or even exceed the grading reliability of human raters, including teaching assistants and expert human raters, in certain contexts. These findings suggest that LLMs can be a viable tool for grading, especially for well-defined short answer questions, thereby reducing the grading burden and potentially enabling more frequent assessments in educational environments (Fagbohun et al., 2024; Schneider et al., Towards LLM-based Autograding).

CHAPTER 4 METHODOLOGY

4.1 DATA

4.1.1 DESCRIPTION OF THE DATASET(S) TO BE USED

The datasets will include a range of assessment materials from tertiary education, focusing on institutions where Afrikaans is a primary language to include Afrikaans data. Covering various subjects, these datasets aim to evaluate LLMs' grading and feedback abilities across disciplines. Each dataset will feature:

- **Questions:** Various study area questions demonstrating diverse types and complexities.
- **Student Responses:** Authentic answers from online assessments.
- **Grading Criteria:** Detailed evaluation standards or model answers.
- **Maximum Marks:** Highest possible scores for benchmarking.
- **Analysis Fields:** Additional data like study year and subject for deeper grading trend analysis.

4.1.2 SOURCE OF DATA, ITS QUALITY, AND VALIDATION METHODS

The primary source of these datasets will be direct partnerships with tertiary education institutions willing to participate in this study. These institutions are expected to have robust online assessment systems from which authentic assessment data can be securely and ethically extracted. For the purpose of including Afrikaans data, collaboration with South African universities or colleges, where Afrikaans is a significant medium of instruction, will be sought.

To ensure the quality of the data:

- **Data Quality Assurance Measures**: Initial screenings will be conducted to verify the completeness, accuracy, and relevance of the data collected. This includes checking for proper documentation of questions, answers, rubrics, and other relevant fields.
- **Validation Methods**: A subset of the data will undergo a manual review process by academic staff or subject matter experts to ensure its integrity and the correctness of the grading rubric. This step is crucial for validating the dataset as a reliable standard for assessing LLM performance.
- **Ethical Considerations and Anonymity**: All data will be anonymized to protect student privacy, with personal identifiers removed or obscured. The project will adhere to ethical guidelines for research involving human subjects, ensuring that data usage complies with relevant privacy laws and institutional policies.

4.1.3 DETAILED DATA BREAKDOWN

- **Total Questions with Responses:** Around 9,000-10,000 questions, each with 10 student answers (about 90,000-100,000 responses total), allowing comprehensive LLM evaluation.
- **Distribution:** 5 questions per academic year per subject, with 6 subjects each year, over 3 years of study, spanning 10 fields of study. This structure ensures wide-ranging, interdisciplinary data for thorough LLM testing.

4.2 METHODS

4.2.1 METHODS FOR ACHIEVING PROJECT AIMS

1. **Data Collection and Cleanup**: Educational institutions will be approached for data, which will be cleaned to anonymize personal details, possibly using Name Entity Recognition (NER) to meet privacy norms. Care will be taken to preserve the balance between privacy and the integrity of data, as removing identifiable details might affect response context and study accuracy.
2. **Prompt Structure Testing**: Various prompt configurations will be trialed to optimize LLM interaction, likely using a JSON format for efficient data handling. Different strategies for grading and feedback justification will be explored across LLMs to enhance the evaluation process.
3. **Infrastructure for Scalability**: The project will investigate cloud-based solutions for managing large data volumes, employing AWS queues and Lambda for streamlined data exchange with the OpenAI API. A PostgreSQL database will store LLM responses for subsequent analysis.
4. **Feasibility and Accuracy Testing**: To gauge LLMs' grading viability, the project will:

    - Compare LLM grades with human evaluations for accuracy.
    - Use metrics like RMSE or Quadratic Weighted Kappa to check grading consistency over multiple assessments, assessing reliability.

4.2.2 JUSTIFICATION OF CHOSEN METHODS

1. **Data Privacy Considerations**: Given the sensitivity of educational data, the use of NER and careful data management practices are essential to protect student privacy while maintaining the integrity of the study.
2. **Prompt Structure Experimentation**: Tailoring the prompt structure is critical to ensure that the LLM can understand and respond to the grading task effectively. This iterative approach allows for adjustments based on preliminary results, ensuring optimal data format and interaction with the LLM.
3. **Scalable Infrastructure**: The choice of cloud-based solutions for scalability is informed by the need to process potentially large datasets efficiently and cost-effectively. Using AWS services enables the project to leverage existing cloud infrastructure for reliable and scalable processing.
4. **Feasibility Evaluation Methods**: Employing both comparison to human grades and consistency checks across LLM runs provides a comprehensive assessment of the LLM's grading accuracy and reliability. These methods ensure that the project's findings are grounded in empirical data and relevant statistical measures.

CHAPTER 5 PROJECT MANAGEMENT

5.1 WORK PLAN

5.1.1 WORK PLAN AND TIMELINE

1. **Weeks 1-2: Project Initiation and Planning**
    - Establish project objectives, scope, and stakeholders.
    - Finalize the project team and roles.
    - Develop a detailed project plan, including resources and budget.
2. **Weeks 3-4: Data Collection and Preparation**
    - Collaborate with educational institutions to gather assessment materials.
    - Cleanse data to remove any identifiable information, ensuring compliance with privacy standards.
    - Prepare datasets for analysis, including categorization by field of study, year, and subject.
3. **Weeks 5-6: Development and Testing**
    - Experiment with various prompt structures for optimal LLM interaction.
    - Develop scalable infrastructure using cloud services for efficient data processing.
    - Conduct preliminary tests with LLMs to assess accuracy and consistency in grading.
4. **Weeks 7-8: Evaluation and Refinement**
    - Compare LLM-generated grades with those provided by human evaluators.
    - Evaluate the consistency of LLM grades and refine the model based on feedback.
    - Document findings, including any limitations and areas for future research.

5.1.2 MILESTONES FOR PROJECT PROGRESS EVALUATION

1. **Project Plan Approval (End of Week 2)**
    - A comprehensive project plan is approved by stakeholders, signifying readiness to proceed with data collection and preparation.
2. **Completion of Data Collection and Preparation (End of Week 4)**
    - Datasets are fully collected, cleansed, and prepared for analysis, marking the readiness for development and testing phases.
3. **Preliminary Testing Completed (End of Week 6)**
    - Initial testing with LLMs is completed, providing early insights into the feasibility and accuracy of automating grading processes.
4. **Project Evaluation and Final Report (End of Week 8)**
    - A final evaluation of the project's outcomes is conducted, and a comprehensive report is prepared, highlighting achievements, challenges, and recommendations for future work.

5.2 RISK ASSESSMENT

1. **Data Privacy Concerns:**
   - Issue: Potential exposure or mishandling of student data during collection, analysis, and storage.
   - Countermeasure: Enforce strict anonymization and encryption, consult privacy experts to ensure compliance with laws, and utilize secure storage with access control.
2. **Grading Accuracy and Reliability:**
   - Issue: Possible discrepancies in LLM grading accuracy and reliability versus human graders, risking project credibility.
   - Countermeasure: Validate extensively with human-graded benchmarks, refine LLM parameters, and incorporate expert feedback to adjust grading algorithms.
3. **System Integration Challenges:**
   - Issue: Difficulty in integrating LLM grading with diverse educational IT systems.
   - Countermeasure: Design a flexible system for easy interfacing with various IT setups, collaborate with institutional IT teams for tailored integrations, and ensure the system's adaptability and scalability.
4. **Language and Cultural Inclusivity:**
   - Issue: LLM's potential shortcomings in understanding non-English responses, especially Afrikaans, and cultural contexts.
   - Countermeasure: Focus on multilingual model development, including Afrikaans datasets for comprehensive language support.
5. **Scalability and Performance:**
   - Issue: Risk of scalability constraints or performance issues under heavy loads or in large-scale deployments.
   - Countermeasure: Opt for scalable cloud solutions and efficient processing techniques, perform load testing to resolve bottlenecks, and employ queue systems and load balancers for peak times.

CHAPTER 6 EVALUATION AND IMPACT

6.1 EXPECTED RESULTS

6.1.1 EXPECTED RESULTS

This feasibility study on using Large Language Models (LLMs) for automating university-level grading anticipates several outcomes. It aims to show that LLMs can accurately and consistently grade across disciplines and question types, aligning closely with human educators' assessments. Key expectations include:

1. **Accurate and Consistent Grading**: The study seeks to demonstrate LLMs' ability to match human grading in accuracy and consistency, providing a robust evaluation of complex academic responses across various subjects.
2. **Afrikaans Language Support**: Successful integration of Afrikaans into the LLM grading system, reflecting the project's dedication to linguistic inclusivity and proving the technology's capacity to serve diverse educational environments.
3. **Efficient Data Processing System**: Development of a structured input and output system that integrates smoothly with existing educational IT infrastructures, showing a viable pathway for adopting LLM technology in assessment processes.
4. **Scalable Infrastructure Solutions**: Exploration of scalable solutions for LLM deployment, ensuring cost-effectiveness and efficiency at handling large volumes of assessments, laying the groundwork for practical implementation in educational settings.

6.1.2 GENERALIZABILITY OF THE RESULTS

The study's findings aim to be widely applicable across academic fields and educational contexts, showcasing LLMs' adaptability in grading diverse academic works. This broad applicability promises insights into leveraging LLMs for grading in varied settings and disciplines.

6.2 EVALUATION

6.2.1 METHODS FOR EVALUATING AND VALIDATING PROJECT RESULTS

1. **Comparison with Human Graders**: Evaluating LLM grading against human graders to assess accuracy, consistency, and feedback quality.
2. **Language Accuracy and Inclusivity Tests**: Conducting language-specific assessments to ensure the LLM's effective grading and feedback in English and Afrikaans.
3. **Technical Integration and Usability Testing**: Testing the LLM grading system's integration with educational IT infrastructures to identify its practical applicability.
4. **Scalability and Performance Analysis**: Stress testing the system under high demand to evaluate its scalability and performance.

6.2.2 CRITERIA FOR MEASURING THE ACHIEVEMENT OF PROJECT AIMS

1. **Grading Accuracy and Consistency**: Success will be measured by the alignment of LLM-generated grades with human assessments, using statistical metrics for inter-rater reliability.
2. **Language Support**: The project's success in Afrikaans support will be evaluated through language accuracy tests.
3. **System Integration and Usability**: Positive integration feedback and the resolution of technical issues will indicate successful system implementation.
4. **Scalability and Efficiency**: The system's ability to perform under high demand, maintaining grading speed and stability, will mark its success in scalability and efficiency.

CHAPTER 7 REFERENCES

Fagbohun, O. *et al.* (2024) 'Beyond Traditional Assessment: Exploring the Impact of Large Language Models on Grading Practices', *Journal of Artificial Intelligence, Machine Learning and Data Science*, 2(1), pp. 1–8. Available at: https://doi.org/10.51219/JAIMLD/oluwole-fagbohun/19.

Henkel, O. *et al.* (2023) 'Can LLMs Grade Short-answer Reading Comprehension Questions : Foundational Literacy Assessment in LMICs'. arXiv. Available at: https://doi.org/10.48550/arXiv.2310.18373.

Nilsson, F. and Tuvstedt, J. (2023) *GPT-4 as an Automatic Grader : The accuracy of grades set by GPT-4 on introductory programming assignments*. Available at: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-330993 (Accessed: 8 March 2024).

Schneider, J. *et al.* (no date) 'Towards LLM-based Autograding for Short Textual Answers'.

Using Large Language Models for Automated Grading of Student Writing about Science (2024). Available at: https://doi.org/10.21203/rs.3.rs-3962175/v1.