ECE 1786 - AutoGradeMl

Project Proposal

Word Count: 1192   Penalty: 0%

Eric Zheng 1005751027

Yihang Lin 1006130447

Bokai Shang 1006031928

Yunchuan Zhou 1005771737

1.0 Introduction

As the world becomes more connected, travel becomes increasingly convenient, more people are choosing to study, work or immigrate to other countries, making standardized language tests necessary when assessing one's language proficiency. While attempting to make grading processing objective, the writing sections, which are evaluated by human examiners [1], are prone to subjectivity and inconsistencies. Moreover, many people spend considerable time and money [2] on private tutors just to receive feedback and insights. Therefore, this has created a need for an automated scoring system to provide objective and economic assessments for the writing components.

In this project, the design team decided to focus on the writing section of the academic IELTS exam, since it is the most commonly accepted English exam in Canada. Considering the time constraints, the design will focus on writing task 2, where students are presented with a problem and asked to write an essay in response. This system will generate an overall band score along with comments and suggestions based on the writing submission. This project aims to deliver accessible, consistent feedback to support learners as they prepare for exams.

2.0 Background

Paper [3] explores how large language models (LLMs) can generate both scores and feedback for student essays by optimizing prompt strategies for effective performance. Various prompting approaches, shown in Figure 1, are investigated to enhance feedback quality, illustrating useful techniques for structured feedback generation using LLMs. Findings show that scoring's impact on feedback is low, the approach introduces the feasibility of generating band scores and feedback in parallel, with prompt engineering tailored for specific evaluation tasks.

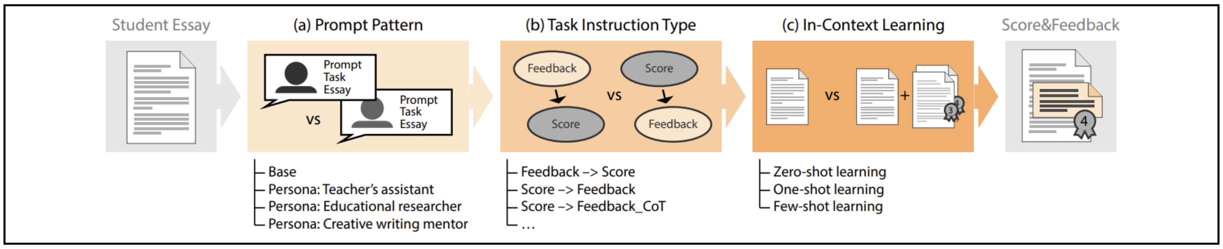

Figure 1: Prompting Strategies [1]

Paper [4] evaluates the reliability and validity of using LLMs to grade EFL essays with a rubric across five domains. High intraclass correlation (ICC) scores, particularly from a fine-tuned GPT model, indicate that LLMs can provide consistent, objective grading comparable to human graders. The study involves 15 EFL instructors for comparison, confirming that both GPT and Bard perform reasonably well, with fine-tuning further enhancing GPT's accuracy. These findings support the use of LLMs for consistent, rubric-based assessments, relevant to the IELTS writing component focus of the project.

3.0 Source of Data and Processing

Kaggle and HuggingFace are our main sources of data, where the design team has found three collections of high-quality IELTS essays from the past, where each sample has prompt, essay, and scores. Despite the conveniences, the samples from different collections are organized and labeled differently, which require extra cleaning procedures.

The processing procedure begins with filtering, removing those samples that belong to the first task through matching regular expressions and removing those that are explicitly labeled under task 1. Then, the remaining samples will be merged together, where exact replicas and undefined samples (i.e. unmeaningful or ill-structured data) will be removed. The scores of each essay will then be standardized to an agreed format.

The last step of data processing requires an evaluation regarding the distribution of essays based on their scores. If the dataset is unbalanced, then sampling or data augmentation (i.e. synonym replacement) techniques will be performed to assure the access of a balanced, standardized, and centralized dataset that is ready to be used for training, fine-tuning, testing, and other application scenarios.

4.0 System Architecture

The project is a Class 2 project that employs a multi-agent system where each agent is powered by a LLM. They perform specific tasks in evaluating IELTS essays. The system processes the input essay to generate scores, produce feedback, and review that feedback.

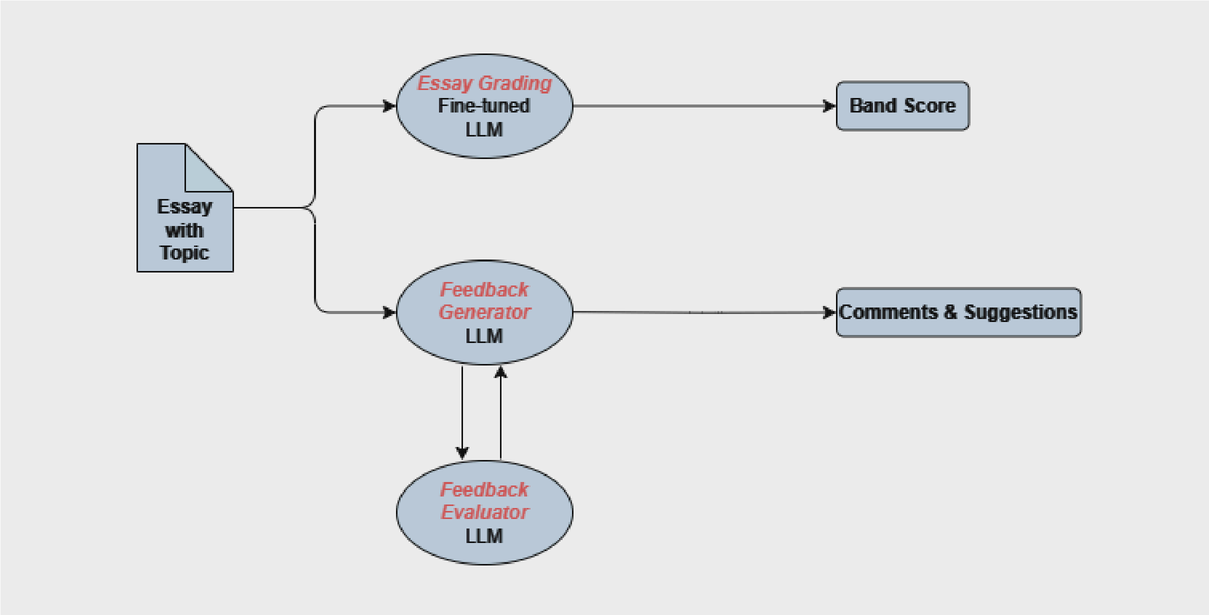

Figure 2: Architecture of the system

4.1 Essay Grading Agent

This agent is a fine-tuned LLM specifically adapted for grading IELTS essays. The model is trained on IELTS essay samples labeled with their band scores to indicate the quality and patterns associated with different score levels.

4.2 Feedback Generator & Feedback Evaluator Agents

The two agents are collaborative LLMs who work together to generate comments and suggestions for the essays. The generator uses prompt engineering to create feedback based on the essay content. It analyzes specific areas and provides suggestions. The evaluator acts as a quality control agent; this LLM uses advanced techniques to evaluate and refine the feedback generated by the generator. It ensures that the feedback is relevant, accurate, and consistent with IELTS standards.

4.3 System Workflow

The main workflow is to simultaneously input the essay with the topic into the Essay Grading Agent to predict a band score and the Feedback Generator to generate comments and suggestions.

5.0 Baseline Model and Comparison

This project consists of two main tasks: predicting overall band scores and generating feedback for IELTS writing.

For the task of predicting overall band score, the baseline model will be a fine-tuned version of a pre-trained LLM, such as GPT-2. The model will be fine-tuned on a collected dataset to function as a score predictor. It will serve as a benchmark for further improvements.

For the task of generating feedback, due to the absence of standardized labels, the system's performance will be evaluated through the following validation method. An evaluation rubric will be established to assess the effectiveness of generated feedback. Based on this rubric, prompt engineering will be applied to a LLM (GPT-4). This model will serve as an evaluator to assess the quality of the generated feedback. During the prompt engineering process, human annotators will label the evaluator’s outputs, and this feedback will be used to continuously improve the evaluator. Ultimately, the success of the generated content will be determined by the evaluator’s assessments.

6.0 Plan

This section describes a table of tasks the design team has determined along with the work assignment to different team members. The design team will follow the responsibilities in the below table, but will also work together if any member needs more help on any particular tasks.

Sub-tasks	**Start**	**End**	**Responsibility**
Data collection and processing	25-Oct	3-Nov	Yihang
Grader agent development	3-Nov	15-Nov	Bokai
Baseline model development	3-Nov	15-Nov	Yunchuan
Feedback generation agent development	3-Nov	15-Nov	Yihang
Feedback evaluation agent development	3-Nov	15-Nov	Eric
Feedback system integration and potential debugging	15-Nov	23-Nov	Team
UI development	23-Nov	30-Nov	Team
Final deliverables	30-Nov	3-Dec	Team

7.0 Risks

The risks of this project will be discussed with respect to different agents proposed in architecture.

The scoring agent will be fine-tuned based on GPT4, the risk of a failure in this classification task of the model is low. However, GPT4 is a very advanced transformer and it contains billions of parameters. There is a very high risk of running out of computational power in fine-tuning the model. If this happens, the design team will adopt LoRA methodology to reduce the trainable parameters in the model. If still not working, the design team will explore the older versions of GPT, like GPT2.

Both feedback agents are developed by prompt engineering. The design team is able to fine tune the two agents individually, but there is a moderate risk of the two agents not working well together. In this case, the design team will intervene in this process by providing more sophisticated prompts and also the IELTS rubrics to guide the agents.

8.0 References

[1] "Ielts writing band scores explained Australia," IDP IELTS Australia. Available at: https://ielts.com.au/australia/results/ielts-band-scores/writing-band-score

[2] "Pricing for services: IELTS tests, Tutoring, & Prep Courses," Coast English Testing. Available at: https://englishtests.ca/pricing/#:~:text=IELTS%20Test%20%24359.99%3B%20IELTS%20Tutoring%20%28per,hour%29%20%2449.50%3B%20IELTS%20Prep%20Course%20%24149.99

[3] Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics. Available at: https://aclanthology.org/2024.bea-1.23/

[4] F. Yavuz, Ö. Çelik, and G. Yavaş Çelik, "Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric‐based assessments," *British Journal of Educational Technology*, Jun. 2024. doi:10.1111/bjet.13494. Available at: https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13494