Automated Grading and Feedback System for Assignments using NLP

Neel Singh
MS CS
Purdue University
singn03@pfw.edu

Ajinkya More
MS CS
Purdue University
morea04@pfw.edu

Arjun Malik
MS CS
Purdue University
malia01@pfw.edu

Abstract

In college departments, grading assignments can be a time-consuming and labor-intensive process for instructors. This project aims to develop an automated grading and feedback system using Large Language Models to streamline the grading process, provide timely feedback to students, and improve the efficiency of assessment workflows.

1 Introduction

Grading assignments in college departments can be a labor-intensive and time-consuming task for instructors. The proposed Automated Grading and Feedback System aims to address this issue by automating the grading process, thereby reducing grading time and providing students with immediate feedback on their assignments. This system leverages pretrained LLM from Langchain which will be fine tuned using real time custom dataset.

2 Motivation

The motivation behind this project stems from the need to enhance the learning experience for students and reduce the burden on instructors within our college department. Manual grading often leads to delayed feedback, which can hinder students' progress and cause anxiety. By automating the grading process, we can provide students with rapid and constructive feedback, ultimately improving their learning outcomes.

3 Problem Statement

The primary problem to address is the inefficiency and time consumption of manual assignment grading. Students often experience anxiety while waiting for their assignments to be graded, affecting their overall learning experience. The proposed system aims to provide a solution by automating the grading process enabling students to receive rapid feedback and improving instructors' workflow.

4 Model/Algorithm to Address Problem

The system will consist of several key components:

4.1 Dataset

Our custom dataset was created by using real time information of students. The dataset includes the task for homework, homework expectations, student's code, student's report, grade points received for the homework(out of 50) and feedback for the student. The missing values in feedback have been interpolated using feedbacks corresponding to those homeworks where the student received the same grade points.

4.2 Modelling

We used multiple models to generate refined feedback generation including Deepset/tinyRoberta, google/FlanT5xxl, intel/bert. However, the performance on google/Palm2 exceeded that on all other models. This model has been fine tuned on tasks such as evaluating programming assignments and question answering which were primarily the reasons for its better performance.

4.3 Approach

There are different ways in which power of transfer learning can be leveraged for custom tasks. This includes but is not limited to fine tuning, instruction fine tuning, reinforcement learning with human feedback and retrieval augmented generation(RAG). We chose RAG for our task primarily because our custom dataset was too small(around 15 training examples). Fine tuning a pretrained model on such a small dataset would have led to overfitting. Therefore the power of RAG was leveraged to enable model to perform well even with small yet meaningful training data.

4.4 Challenges of Context based response generation

Perhaps the easiest way to harness power of LLM is to feed it some context and ask the model questions based on that context. Although this gives good results but the biggest limitation of this approach is that for any LLM the context length is restricted due to fixed token limit. Usually the token limits is around 1024 for open source models like tiny-Roberta. For GPT3.5 which is not openly available the token limit is 4096. Contrary to our belief, these 4096 tokens translate to roughly 300-500 words which is the limit of context these models can usually take without throwing an error. Thus, if we want to provide our huge data to the model then context has to be given differently.

5 Retrieval Augmented Generation (RAG)

The data was divided into smaller documents with fixed chunk size using textsplitter from Langchain. These documents were converted to vector embeddings. The new documents and embeddings were given to FAISS(Facebook AI for Similarity Search) for storage. These documents were fed to our model. The model we chose was Google Palm2 which is an open source pretrained LLM. Once the data is stored in model in vectorized format, the user query is sought in such a way that student can upload the homework task, their code and report. All the 3 files are converted to vectorized format before the model launches a similarity search operation using FAISS. Since the model has identified features of good and poor homework examples when the data was given, it is able to provide the grade along with appropriate feedback for the student's homework.

6 Measurement of Success

Since the dataset is not too big, we are evaluating the model manually. For instance, for a given task if the student received 50 points and positive feedback based on their code and report then we expect the model to assign lower grade and critical feedback in cases where code had mistakes or the report was poorly drafted without proper analysis.

7 Automated Grading and Feedback System for Assignments: Project Update 2

7.1 Current Results

The model is returning near appropriate score based on the given task, code and report. The quality of feedback is fairly alright and is in harmony with the grade received by the student for that homework.

7.2 Upcoming Results

We have two options. First, if we are able to find an openly available dataset where students code or reports have been graded with appropriate feedbacks then we will proceed with fine tuning our existing model. Second, go with the existing approach of RAG and find a way to compress the user query so that it remains within the model's token limit. Further we will feed our model more examples of what a poor report looks like so as to create a balanced dataset.We will also want to explore few shot learning which has shown to give better results with fewer data points. In this way better results can be achieved.

7.3 Analysis Done

Few critical learnings have been made. First, bigger model is not necessarily better. As we saw tiny-Roberta and Palm2 both were smaller models with far lesser parameters than their base models but performed surprisingly better than their counterparts. Second, when it comes to data- size and relevance both matter. We can have publicly available dataset of huge size but it may not be relevant to out task. On the other hand, we may have relevant data but if it is too small in size it will create problems during fine tuning as the pretrained model will tend to overfit on such data. Therefore data should be decently sized and relevant to the task at hand.

7.4 Upcoming Analysis

It remains to be seen how well the model works when fine tuned using our custom dataset. To what extent will few shot learning or instruction fine tuning be relevant to our task. In case we proceed with RAG, how exactly can we avoid the problem of exceeding token limit when user queries are bigger than expected.(We have already resolved token limit errors when feeding data through large context by using FAISS vector databases).

7.5 Problems Encountered

Our initial models proved less-than-successful at accurately predicting results than anticipated, leading us to re-evaluate our approach and modeling strategies We had more data of good reports and code where the homeworks received high points(nearly 50). This posed barrier in training the model on examples of poor homeworks thereby leading to inappropriate predictions. Moreover, less data meant that fine tuning LLM on our custom data would lead to overfitting. Further, token limit issue in LLMs posed barrier in feeding data to model through context

7.6 Conclusion

After many trials, our experiments have shown us promising results not only in terms of meeting the agenda of the project but also in generating appropriate results based on the real time data of students. We intend to take our approach further for better results.