# Capstone Project Weekly Report

## Date: 20/10/2024

## Project Details:

1. Sponsor Company: Prepshala.Pvt.LTd

2. Project Title: An adaptive screening tool using Generative AI and LLMs for the assessment of language and personality traits

**Note:** All the fields in the form are required.

## Project Milestones:

Progress made in the current week and contribution from individual team members:

This week, we focused on finalizing enhancements to the writing module and developing a prototype for the listening model. These updates were implemented based on the valuable feedback provided by both the sponsor and faculty mentors.

**Writing Module Enhancements:** We introduced several key updates to the writing module as follows:

1. **Integration of Evaluation Columns:** Following discussions with the faculty mentor, we integrated the **essay question** and **evaluation columns** into the model. These evaluation columns encompass the four key components used in IELTS scoring:
    - Task Achievement
    - Coherence and Cohesion
    - Lexical Resource
    - Grammatical Range and Accuracy

Incorporating these evaluations has significantly enhanced the model's ability to train on these metrics and predict more accurate band scores. Previously, the model did not account for these evaluations, which limited its precision. With this integration, we observed a marked improvement in the model's performance.

Despite the added complexity, the model achieved **strong RMSE and MAE scores**, demonstrating its effectiveness. We developed two distinct models: one trained on full-length essays and the other on summarized essays. Consistent with earlier observations, the **summarized model** outperformed the full-length essay model.

**Performance Metrics for the Essay Model:**

- Final Test MAE: 1.1196
- Final Test RMSE: 1.3398

**Performance Metrics for the Summarized Essay Model:**

- Final Test MAE: 0.4904
- Final Test RMSE: 0.6517

**2. Incorporating Relevancy Scores:** After integrating the evaluation columns, we observed that the model still did not account for **relevancy**—whether the answer provided was contextually appropriate for the question. This limitation allowed irrelevant responses to receive high scores. To address this issue, we explored several approaches for incorporating relevancy into the scoring mechanism:

- **a) SBERT Model for Relevancy Calculation:** We utilized SBERT to calculate the relevancy between the answer and the question.
- **b) Penalization Approach Using SBERT:** We implemented various penalty functions, as suggested by the faculty mentor, to reduce scores for irrelevant answers. Two penalty equations were tested:
  1. **Band score with penalty = A * B**, where **A** is the band score and **B** is the relevancy score. This is a dynamic approach, as it adjusts the band score based on relevancy, with a set threshold (tested range 0.05 to 0.5).

  2. **Band score with penalty = A * B / ($A^2$ + $B^2$).** This equation, however, disrupted the entire IELTS band scoring system, as it produced inconsistent results and did not align well with the desired scoring scale.
- **c) Relevancy Scoring Using LLMs:** Finally, we explored using large language models (LLMs) to assess relevancy and apply appropriate penalties.

While the SBERT-based penalization approach (Approach B) provided some improvement, it proved to be too **raw** for our specific use case. The SBERT model, not being fine-tuned for our task, even penalized **relevant answers**, which reduced its overall effectiveness. In contrast, **Approach C**—utilizing LLMs—proved to be the most effective. The LLM was able to assess relevancy more accurately, leading to better model performance when applying the penalty function **Band score with penalty = A * B**, where **A** is the band score and **B** is the relevancy score.

As suggested by the sponsor mentor, we conducted **Exploratory Data Analysis (EDA)** on the final results to compare the predicted scores for **summarized** and **full-length essays** versus actual band scores. The **predicted scores for summarized essays** showed a smaller deviation from the actual scores, ranging from **-0.5 to 0.5**, whereas the deviations for **full-length essays** were higher, indicating less accuracy.

```
Prompt (Not Evaluated): The ocean is ____. (Please fill in the blank with **one word**.)
Your response: blue
Prompt: Some people say that playing computer games is bad for children in every aspect. Others say that playing computer games can have positive effects on the
Your response: Some people believe that pursuing higher education at a university or college is the best path to a successful career. They argue that obtaining a
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the model
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the model
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 320ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 175ms/step
Model 1 - Response 1 (Dataset Question) predicted band score: 5.5
Model 2 - Response 1 (Dataset Question) predicted band score: 6.5
```

*Model with no relevancy scores and penalty*

```
Prompt (Not Evaluated): The ocean is ____. (Please fill in the blank with **one word**.)
Your response: blue
Prompt: Some people think that art is an essential subject for children at school while others think it is a waste of time. (Answer in 100-150 words)
Your response: The debate surrounding the inclusion of art as a subject in school curricula is ongoing. While some argue that art is essential for children's dev
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the model
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the model
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 127ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 133ms/step
Model 1 - Response 2 predicted band score: 5.5, Relevancy Score: 1.00, Final Adjusted Band Score: 5.5
Model 2 - Response 2 predicted band score: 6.0, Relevancy Score: 1.00, Final Adjusted Band Score: 6.0
```

*Model with relevancy scores (Answer is relevant).*

```
Prompt (Not Evaluated): The ocean is ____. (Please fill in the blank with **one word**.)
Your response: blue.
Prompt: Some people think that art is an essential subject for children at school while others think it is a waste of time. (Answer in 100-150 words)
Your response: Celebrities often find themselves under intense scrutiny from the media, leading to complaints about the ways in which their lives are portrayed.
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the model
WARNING:absl:Compiled the loaded model, but the compiled metrics have yet to be built. `model.compile_metrics` will be empty until you train or evaluate the model
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 83ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 78ms/step
Model 1 - Response 2 predicted band score: 5.5, Relevancy Score: 0.00, Final Adjusted Band Score: 1.0
Model 2 - Response 2 predicted band score: 5.5, Relevancy Score: 0.00, Final Adjusted Band Score: 1.0
```

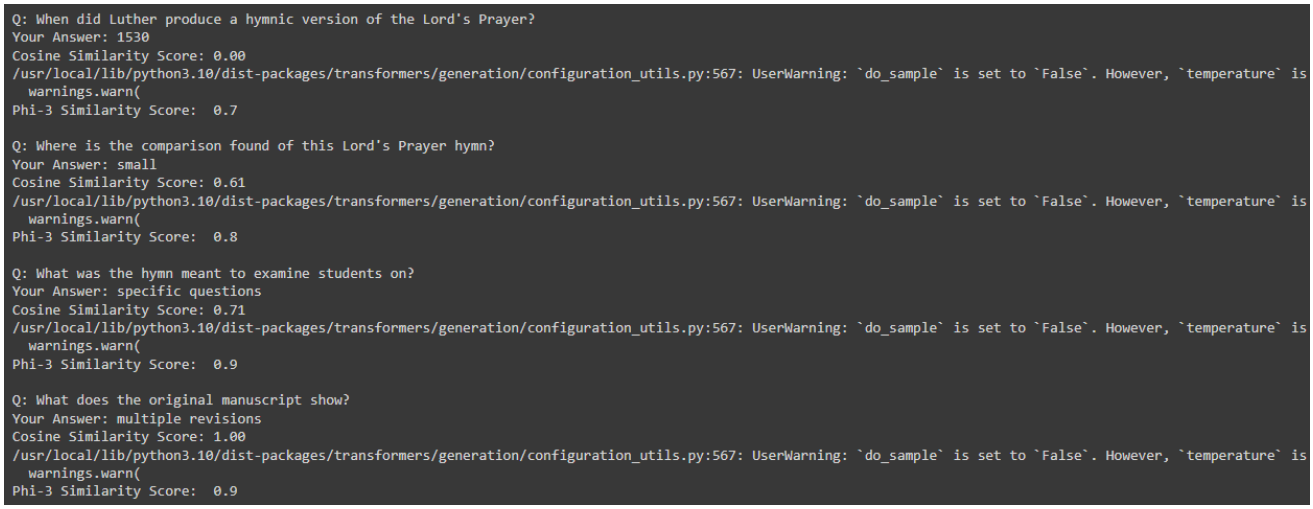
*Model with relevancy scores (Answer is not relevant to the question).*

**Listening Module:**
After conducting thorough research on the IELTS listening module, we identified that it consists of 40 questions, encompassing various question types such as question-and-answer, fill-in-the-blanks, and others. We have decided to proceed with the question-and-answer format, where a paragraph will be read aloud using a text-to-speech module, and questions will be posed based on the content of the paragraph.

We have sourced a relevant question-answering dataset that includes both the questions and corresponding answers. In our initial approach, we developed a model that accepts user input for the questions and compares it to the dataset. To assess whether the response is correct, we implemented phi similarity score and cosine similarity score. If the similarity score between the user's answer and the dataset answer is greater than 0.7, the answer is marked correct and a score of +1 is awarded; otherwise, it is marked as incorrect with a score of 0.

The band score will be mapped according to the IELTS listening module's standard mapping, based on the total number of correct answers. We plan to explore additional models and approaches to further enhance accuracy and alignment with IELTS standards.


*Text to speech module.*




*Evaluation module.*

## Tasks to finish in next week:

1. **Demonstrate the Enhancements in the Writing Module:** Present the improvements made to the writing module to both the sponsor mentor and faculty mentor, incorporating their feedback for further refinement.
2. **Complete the Listening Module:** Explore and implement different models to enhance the accuracy and performance of the listening module, ensuring it aligns with IELTS standards.
3. **Prepare for Mid-Review:** Ensure the writing and listening modules are fully functional and ready for the upcoming mid-review, with both modules refined and demonstrated.
4. **Research on the Speaking Module:** Begin researching various approaches and datasets for developing the speaking module, focusing on how to evaluate speaking skills based on IELTS standards.

## Updates/MoM from Sponsor and Faculty Mentor:

**Note: It is expected that you have at least one weekly connect with the faculty mentor and sponsor.** If you were not able to schedule meetings with the sponsor or faculty mentor in the current week, please mention the reason for your inability to meet with the Sponsor or Faculty Mentor.

### Updates/MoM from Sponsor:

We had a 45-minute meeting with the sponsor mentor on **17th October 2024**, during which the following points were discussed:

1. **Demonstration of the Writing Module:** The enhancements made to the writing module were presented, and we received appreciation for achieving lower **RMSE** and **MAE** values, reflecting improved model accuracy.
2. **Discussion on Summarization and Model Performance:** We discussed the summarization process and the challenges encountered with models trained on full-length essays, which did not perform as well on the test set.
3. **Relevancy in the Writing Module:** The idea of introducing relevancy scoring and penalization in the writing module was proposed. The sponsor mentor expressed interest in this approach and encouraged further exploration.
4. **Demonstration of the Listening Module:** We provided a demonstration of the listening module, showcasing the evaluation approach using **similarity scores**.
5. **Action Item:** The sponsor mentor requested that we discuss the relevancy and evaluation methods for the listening module with the professor mentor for further refinement.

### Updates/MoM from Faculty Mentor:

We had a 45-minute call with the faculty mentor on **19th October 2024**, during which we demonstrated the current model and shared the feedback received from the sponsor mentor. Below are the key takeaways:

1. **Demonstration of the Model:** We presented the current model and explained the steps taken to improve the **RMSE** and **MAE** over the past week.
2. **Relevancy Problem Discussion:** We highlighted the relevancy issue in the existing model, where it was not effectively assessing the relevancy of answers.
   **Equations for Relevancy:** The professor provided us with additional equations to try out for improving relevancy scoring, which we have implemented and will demo in the coming week.
3. **Methods to Address the Relevancy Problem:** We discussed three potential approaches to address the relevancy issue:
    1. Integrate the **evaluation column** into the model for better relevancy scoring.
    2. If the results are unsatisfactory, consider adding a **separate module** for relevancy using a **dynamic relevancy** approach.
    3. Explore the use of **LLMs** to improve relevancy detection.
4. **Threshold Concerns:** We discussed potential issues related to manipulating **thresholds** in relevancy scoring and the impact this might have on overall model performance.
5. **Demonstration of the Listening Module:** We showcased the approach for the listening module. The professor confirmed that the approach was sound and aligned with the project objectives.

## Challenges:

Mention any technical and non-technical challenges that you faced during the current week that hindered your project progress. Enter "NA" if you didn't face any challenges.

### Technical Challenges:

1. Difficulty in finding high-quality IELTS-level datasets for each module, particularly for the listening and speaking modules.
2. Difficulty in Finding an IELTS Dataset for the Listening Module: We faced challenges in locating a suitable open-source dataset specifically designed for the IELTS listening module.
3. **Working with Open-Source Models:** Implementing and fine-tuning open-source models presented certain complexities, particularly in adapting them to meet the specific requirements of the listening and writing modules.
4. **Addressing the Relevancy Problem:** We successfully tackled the relevancy issue in the writing module by introducing the IELTS evaluation column into the model, which improved the prediction accuracy and overall performance.

### Non-Technical Challenges:

NA

### Mention any other queries/challenges regarding the project that you want to highlight:

NA