# ECE 1786 - AutoGradeMl

*Progress Report*

Word Count: 998   Penalty: 0%

Eric Zheng 1005751027

Yihang Lin 1006130447

Bokai Shang 1006031928

Yunchuan Zhou 1005771737

## 1.0 Introduction

The goal of this project is to create an automated scoring system that provides objective and consistent feedback for the students practicing and preparing for IELTS Writing Task-2.

## 2.0 Input/Output

The system input includes a writing prompt and the corresponding essay response from the user.

The output of the system consists of a band score and feedback on the essay. The band score ranges from 4 to 9 with 0.5 intervals, with scores below 4 labeled as “<4” for standardization. The feedback consists of a review of the essay, with improvement suggestions aligned to the IELTS rubric.

### 2.1 Example:

**Input:**

[Writing Question]
Not enough students choose… What are the effects on society?
[Essay]
In recent time, it is observed that the number of students …

**Output:**

[Score]
7
[Feedback]
**Comment:**

Your essay provides …

…

**Suggestions:**

Consider discussing …

# 3.0 Data

Our data, sourced from Hugging-Face and Kaggle, was standardized into three attributes: writing question, essay, and band score. The original datasets remain unchanged, as some include feedback examples that can be utilized for few-shot learning in the feedback generator and evaluator.

A balanced subset of the cleaned dataset was created for fine-tuning GPT-4 grading agent. Both datasets have the same attributes, with distributions and a training example shown below. Part of the subset was used to test prompts and integrate feedback generator and evaluator into an interactive agent system. The team uses official IELTS writing rubrics for prompt engineering.

### 3.1 Datasets Visualization:

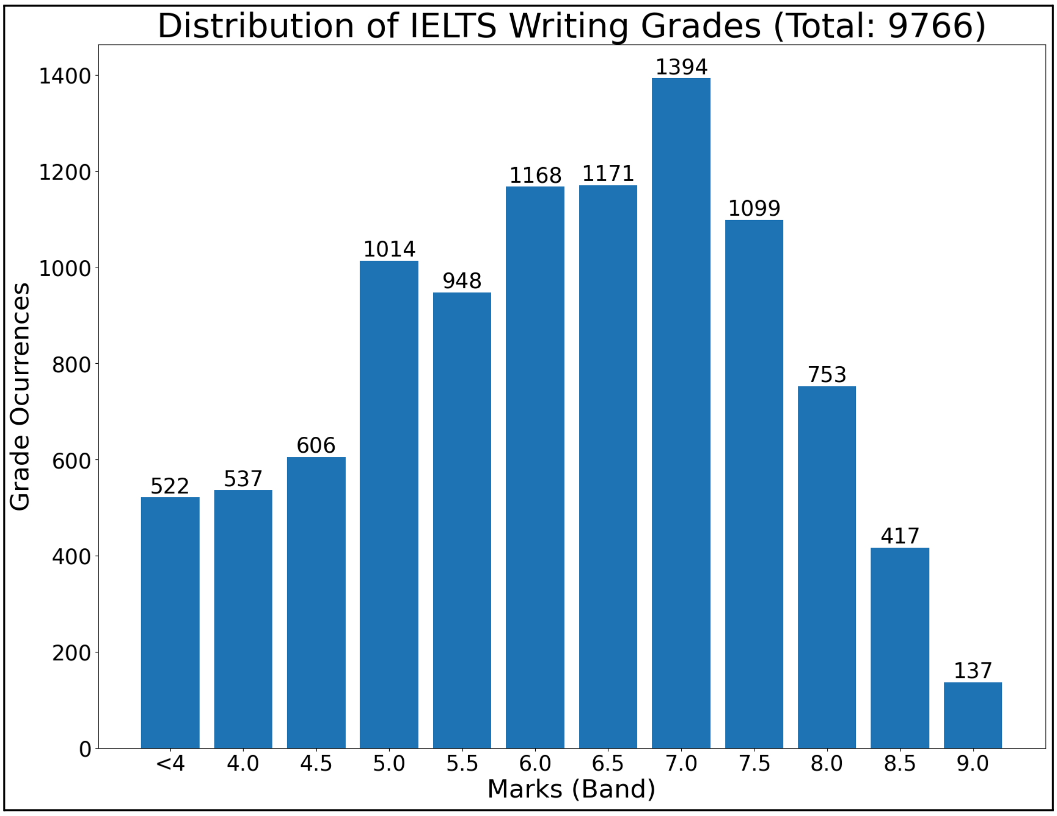


Figure 1: Cleaned Dataset Distribution

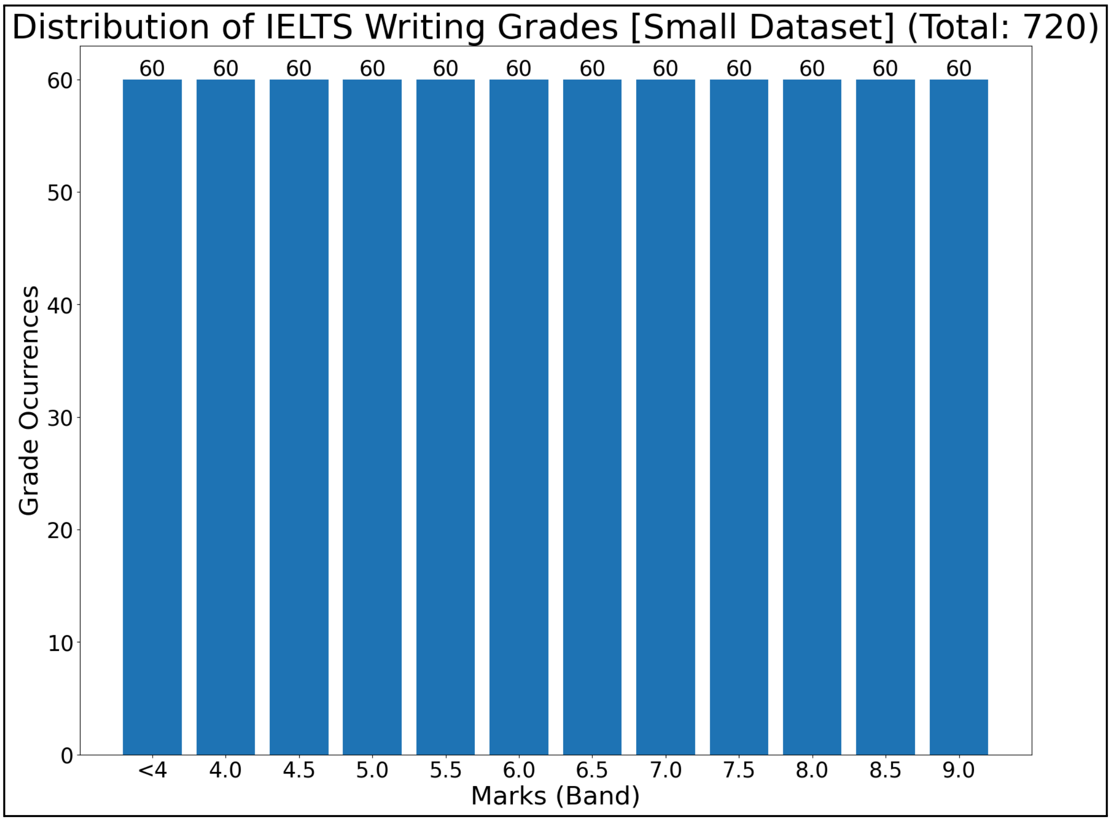


Figure 2: Small Dataset Distribution

**Example:**

[Writing Question]

The increase in the production… What can be done to solve this problem?

[Essay]

Nowadays, the production of consumer goods results**…**

[Band]

7.5

## 4.0 Baseline Model and Comparison Method

This project consists of two main tasks: Task 1 - Predicting overall band scores and Task 2 - Generating feedback for IELTS writing.

For Task 1, we use GPT-4o-mini as the baseline model. It is the benchmark for further developments in fine-tuning other LLMs.

For Task 2, we developed a detailed rubric to evaluate the quality of feedback generated. This rubric consists of the following criteria:

1. **Clarity:** Feedback should be easy to understand, using simple and precise wording to ensure the student knows what is meant, avoiding ambiguities.
2. **Relevance:** The feedback must address the four main components of the official IELTS rubric.
3. **Specificity:** Feedback should include detailed comments, avoiding vague suggestions. Instead, it should pinpoint exact issues and, if possible, provide examples from the essay.
4. **Actionable Suggestions:** Feedback should offer clear, constructive advice that the writer can apply to improve their writing.
5. **Tone:** The feedback should maintain an encouraging and respectful tone to motivate the writer to make improvements.

Human annotators will use this rubric to determine the success of the evaluator. The evaluator will determine the success of the generator.

## 5.0 Architecture

The project employs a multi-agent system powered by LLMs.

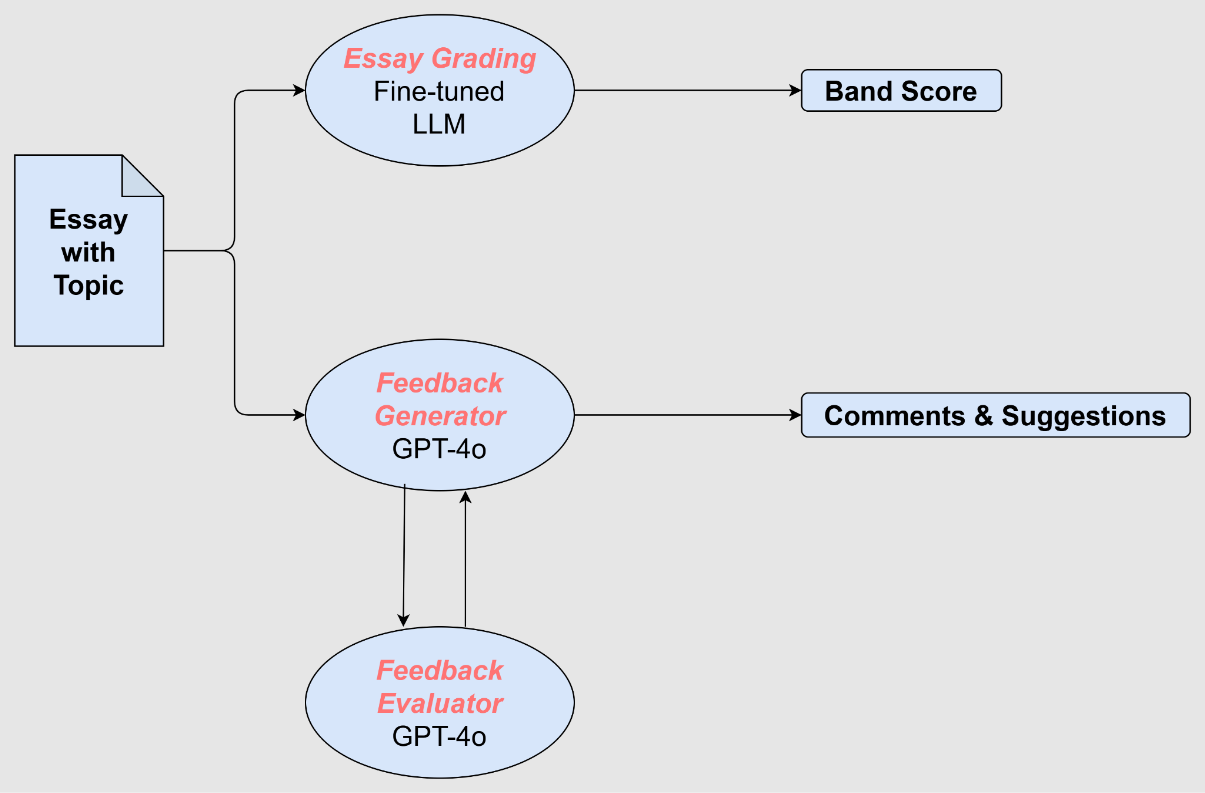


Figure 3: Architecture

### 5.1 Essay Grading Agent

It’s a fine-tuned LLM predicting the band score given the input. We have fine-tuned four LLMs, but require further investigation for better performance.

### 5.2 Feedback Generator & Evaluator Agents

Feedback Generator: Analyze the input and generate detailed comments and improvement suggestions.

Feedback Evaluator: Review the feedback, if improvements are needed, provide revision suggestions.

Two GPT-4o agents interact conversationally: the generator provides feedback and refines it based on suggestions from the evaluator. This iterative process continues until the feedback is sufficiently refined.

The models have been built, we are in the process of human evaluating the feedback for further improvement.

### 5.3 System Workflow

Input: Essay topic and text.

Processing: The Grading Agent predicts the band score, simultaneously the Feedback Generator generates feedback which is supervised by the Evaluator.

Output: Band score and feedback.

## 6.0 Result

### 6.1 Band Score Classifier

The band score classifier was tested with various LLMs. Initial experiments used GPT4o-mini with limited data and epochs due to high cost, followed by full-scale training on free models. Table 1 shows improvements over the baseline for all models except Llama2, and DistillBERT performs best.

| Model | Test Accuracy (no tolerance) | Test Accuracy (±0.5 tolerance) |
|---|---|---|
| Baseline | 9% | 31% |
| GPT4o-mini | 14% | 41% |
| GPT2 | 20% | 47% |
| Llama2 | 9% | 12% |
| DistillBERT | 23% | 52% |

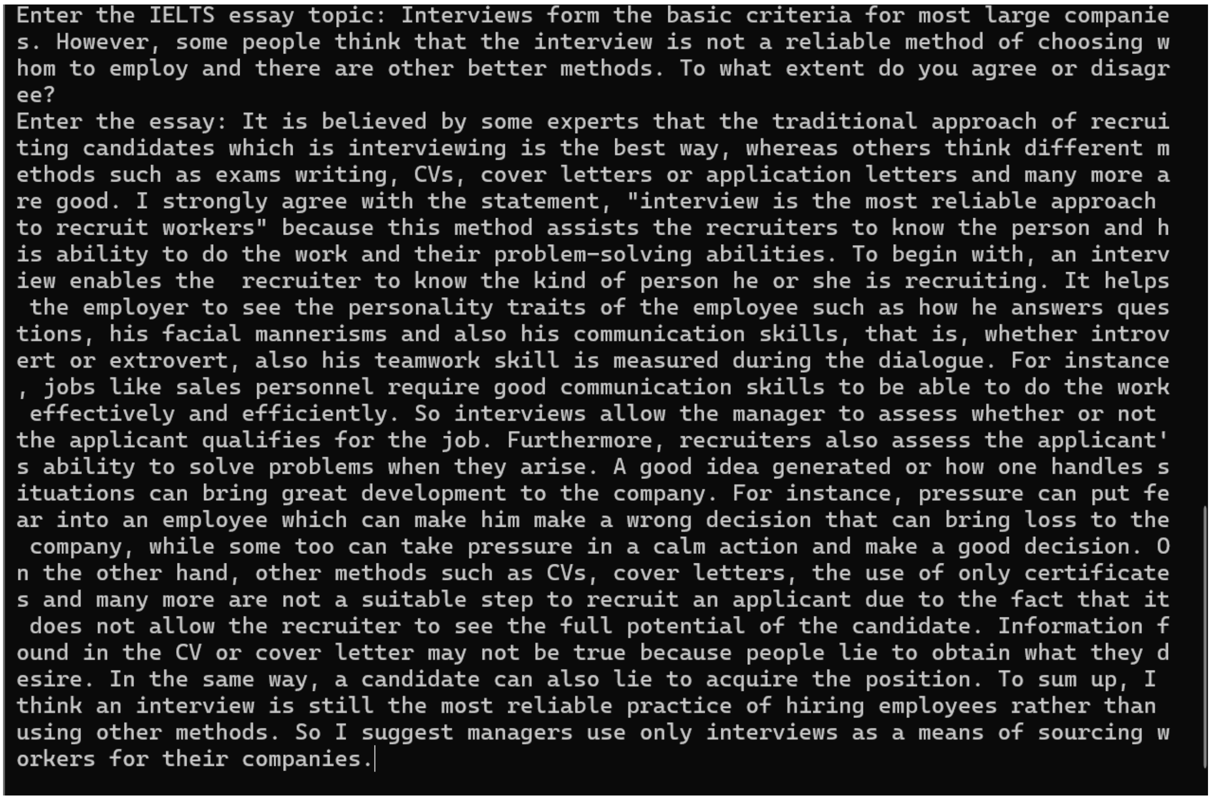
Table 1. Band Score Result

### 6.2 Feedback Generator & Evaluator

The generator’s success is evaluated by the evaluator shown in 6.3, and feedback is evaluated by humans based on the IELTS rubric for initial screening. The evaluator's success is verified by human annotators using the evaluation rubric in section 4.0.

### 6.3 Example Result

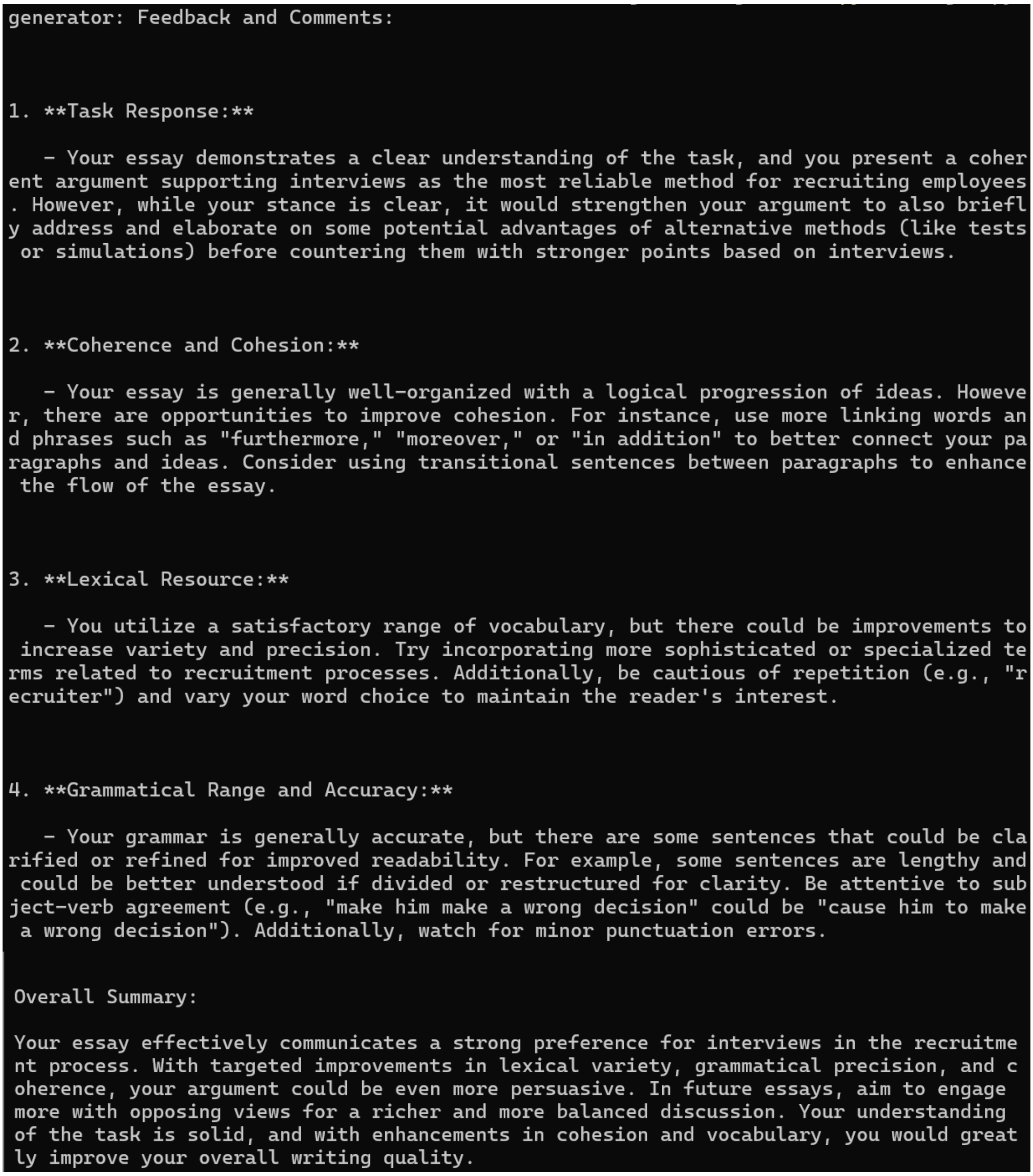
**<u>Generator Input:</u>**



```
Enter the IELTS essay topic: Interviews form the basic criteria for most large companie
s. However, some people think that the interview is not a reliable method of choosing w
hom to employ and there are other better methods. To what extent do you agree or disagr
ee?
Enter the essay: It is believed by some experts that the traditional approach of recrui
ting candidates which is interviewing is the best way, whereas others think different m
ethods such as exams writing, CVs, cover letters or application letters and many more a
re good. I strongly agree with the statement, "interview is the most reliable approach
to recruit workers" because this method assists the recruiters to know the person and h
is ability to do the work and their problem-solving abilities. To begin with, an interv
iew enables the  recruiter to know the kind of person he or she is recruiting. It helps
 the employer to see the personality traits of the employee such as how he answers ques
tions, his facial mannerisms and also his communication skills, that is, whether introv
ert or extrovert, also his teamwork skill is measured during the dialogue. For instance
, jobs like sales personnel require good communication skills to be able to do the work
 effectively and efficiently. So interviews allow the manager to assess whether or not
the applicant qualifies for the job. Furthermore, recruiters also assess the applicant'
s ability to solve problems when they arise. A good idea generated or how one handles s
ituations can bring great development to the company. For instance, pressure can put fe
ar into an employee which can make him make a wrong decision that can bring loss to the
 company, while some too can take pressure in a calm action and make a good decision. O
n the other hand, other methods such as CVs, cover letters, the use of only certificate
s and many more are not a suitable step to recruit an applicant due to the fact that it
 does not allow the recruiter to see the full potential of the candidate. Information f
ound in the CV or cover letter may not be true because people lie to obtain what they d
esire. In the same way, a candidate can also lie to acquire the position. To sum up, I
think an interview is still the most reliable practice of hiring employees rather than
using other methods. So I suggest managers use only interviews as a means of sourcing w
orkers for their companies.
```

Figure 4: Generator Sample Input

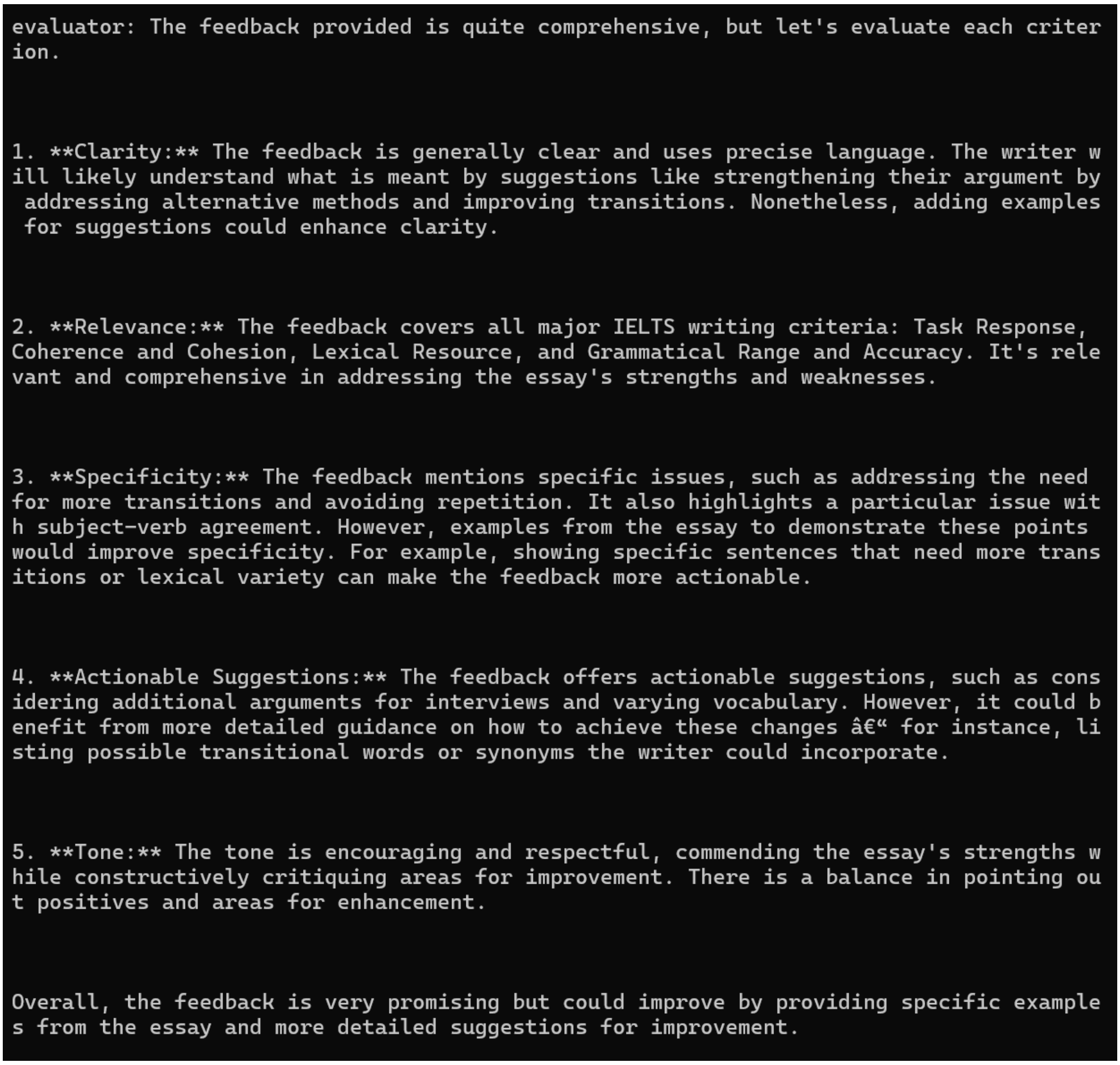
**<u>Generator Output / Evaluator Input:</u>**



```
generator: Feedback and Comments:



1. **Task Response:**

   - Your essay demonstrates a clear understanding of the task, and you present a coher
ent argument supporting interviews as the most reliable method for recruiting employees
. However, while your stance is clear, it would strengthen your argument to also briefl
y address and elaborate on some potential advantages of alternative methods (like tests
 or simulations) before countering them with stronger points based on interviews.



2. **Coherence and Cohesion:**

   - Your essay is generally well-organized with a logical progression of ideas. Howeve
r, there are opportunities to improve cohesion. For instance, use more linking words an
d phrases such as "furthermore," "moreover," or "in addition" to better connect your pa
ragraphs and ideas. Consider using transitional sentences between paragraphs to enhance
 the flow of the essay.



3. **Lexical Resource:**

   - You utilize a satisfactory range of vocabulary, but there could be improvements to
 increase variety and precision. Try incorporating more sophisticated or specialized te
rms related to recruitment processes. Additionally, be cautious of repetition (e.g., "r
ecruiter") and vary your word choice to maintain the reader's interest.



4. **Grammatical Range and Accuracy:**

   - Your grammar is generally accurate, but there are some sentences that could be cla
rified or refined for improved readability. For example, some sentences are lengthy and
 could be better understood if divided or restructured for clarity. Be attentive to sub
ject-verb agreement (e.g., "make him make a wrong decision" could be "cause him to make
 a wrong decision"). Additionally, watch for minor punctuation errors.

Overall Summary:

Your essay effectively communicates a strong preference for interviews in the recruitme
nt process. With targeted improvements in lexical variety, grammatical precision, and c
oherence, your argument could be even more persuasive. In future essays, aim to engage
more with opposing views for a richer and more balanced discussion. Your understanding
of the task is solid, and with enhancements in cohesion and vocabulary, you would great
ly improve your overall writing quality.
```

Figure 5: Generator Sample Output

**Evaluator Output:**:



```
evaluator: The feedback provided is quite comprehensive, but let's evaluate each criter
ion.



1. **Clarity:** The feedback is generally clear and uses precise language. The writer w
ill likely understand what is meant by suggestions like strengthening their argument by
 addressing alternative methods and improving transitions. Nonetheless, adding examples
 for suggestions could enhance clarity.



2. **Relevance:** The feedback covers all major IELTS writing criteria: Task Response,
Coherence and Cohesion, Lexical Resource, and Grammatical Range and Accuracy. It's rele
vant and comprehensive in addressing the essay's strengths and weaknesses.



3. **Specificity:** The feedback mentions specific issues, such as addressing the need
for more transitions and avoiding repetition. It also highlights a particular issue wit
h subject-verb agreement. However, examples from the essay to demonstrate these points
would improve specificity. For example, showing specific sentences that need more trans
itions or lexical variety can make the feedback more actionable.



4. **Actionable Suggestions:** The feedback offers actionable suggestions, such as cons
idering additional arguments for interviews and varying vocabulary. However, it could b
enefit from more detailed guidance on how to achieve these changes â€“ for instance, li
sting possible transitional words or synonyms the writer could incorporate.



5. **Tone:** The tone is encouraging and respectful, commending the essay's strengths w
hile constructively critiquing areas for improvement. There is a balance in pointing ou
t positives and areas for enhancement.



Overall, the feedback is very promising but could improve by providing specific example
s from the essay and more detailed suggestions for improvement.
```

Figure 6: Evaluator Sample Output

# 7.0 Discussion

## 7.1 Band Score Classifier

The results show improvement compared to the baseline but still remain insufficient. The task involves 12-class classification, where DistillBERT outperforms generative models due to its specialization in classification tasks.  However, in the training process, the design team noticed an overfitting problem, leaving room for improvement with generalization techniques. Considering the overall low performance, this task is a very difficult classification task. The team is also open to switching to an agentic system if accuracy issues persist.

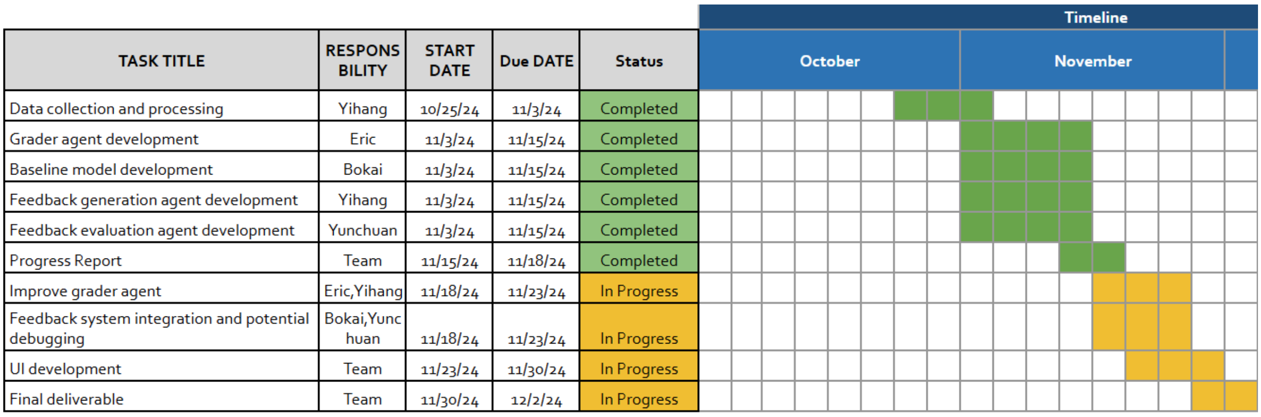
### 7.2 Feedback Generator & Evaluator

From the example in part 6.3, the generated results from the generator and the evaluator were overall satisfactory, with both successfully passing the initial human evaluation screening. The evaluator has also successfully provided suggestions for the generator to improve.

The final architecture requires both agents to interact with each other, so finding a plausible method of linking both agents effectively will be one of the main focuses.

## 8.0 Team Work and Progress

The team works together very well. All team members have completed all the tasks assigned on time, but there is a small change in the job assignment. All the tasks planned to be completed by this report are completed. A detailed Gantt Chart is provided below.



| | | | | | Timeline | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| TASK TITLE | RESPONSBILITY | START DATE | Due DATE | Status | October | | | | | | | | November | | | | | | | | |
| Data collection and processing | Yihang | 10/25/24 | 11/3/24 | Completed | | | | | | | 🟩 | 🟩 | 🟩 | | | | | | | | |
| Grader agent development | Eric | 11/3/24 | 11/15/24 | Completed | | | | | | | | | 🟩 | 🟩 | 🟩 | 🟩 | | | | | |
| Baseline model development | Bokai | 11/3/24 | 11/15/24 | Completed | | | | | | | | | 🟩 | 🟩 | 🟩 | 🟩 | | | | | |
| Feedback generation agent development | Yihang | 11/3/24 | 11/15/24 | Completed | | | | | | | | | 🟩 | 🟩 | 🟩 | 🟩 | | | | | |
| Feedback evaluation agent development | Yunchuan | 11/3/24 | 11/15/24 | Completed | | | | | | | | | 🟩 | 🟩 | 🟩 | 🟩 | | | | | |
| Progress Report | Team | 11/15/24 | 11/18/24 | Completed | | | | | | | | | | | | 🟩 | 🟩 | | | | |
| Improve grader agent | Eric,Yihang | 11/18/24 | 11/23/24 | In Progress | | | | | | | | | | | | | 🟨 | 🟨 | 🟨 | | |
| Feedback system integration and potential debugging | Bokai,Yunchuan | 11/18/24 | 11/23/24 | In Progress | | | | | | | | | | | | | 🟨 | 🟨 | 🟨 | | |
| UI development | Team | 11/23/24 | 11/30/24 | In Progress | | | | | | | | | | | | | | 🟨 | 🟨 | 🟨 | |
| Final deliverable | Team | 11/30/24 | 12/2/24 | In Progress | | | | | | | | | | | | | | | | 🟨 | 🟨 |

Figure 7: Gantt Chart