جامعة الملك عبدالله للعلوم والتقنية
King Abdullah University of Science and Technology

أكاديمية كاوست
KAUST ACADEMY

Project Title: GenIELTS: Complete AI Simulation & Evaluation for IELTS

Group Members & Contributions

- **Abdulkareem Abdulgadir Omer** – Writing Evaluation Module.
- **Mahmoud Omar**– Listening and reading Modules.
- **Abdalla Abbas**– Speaking Module.

Project Mentor

Dr. Salah Abdeljabar

Project Objectives and Goals

The project aims to develop an end-to-end AI system to assist users in preparing for the Academic IELTS exam. The system will:

- Evaluate **Speaking** using AI transcription and pronunciation analysis.
- Score **Writing Task 2** based on IELTS writing band descriptors and IELTS writing key assessment criteria and give a detailed report for the user that has sub-scores, estimated overall band, examiner's comments, strengths and weaknesses and areas to focus on. The user's input will be the subject of the essay and the essay they have written.
- Generate IELTS-style **Listening** tests using speech synthesis and topic modeling, the user will choose the part they want to have a test on and the topic of the part.
- Generate **Reading** exams on specific topics and parts based on user's choice.
- Score **Writing Task 1** based on IELTS writing band descriptors and IELTS writing key assessment criteria and give a detailed report for the user that has

sub-scores, estimated overall band, examiner's comments, strengths and weaknesses and areas to focus on. The user's input will be the subject of the report image, the report image and the essay they have written.

Current Progress and Milestones Achieved

- **Writing Task 1:** The writing task 1 module works perfectly. The user's input will be the image, the subject of the image and the essay they have written. The module then generates detailed evaluation for the essay.
- **Speaking:** All objectives have been achieved. One limitation of the speaking module is that it doesn't cover all English accents and dialects and that it has some edge cases in spontaneous speech.
- **Writing Task 2:** The writing task 1 module works perfectly. The user's input will be the essay question, and the output will be the detailed evaluation report.
- **Listening:** All objectives have been achieved, and the module works as expected, the user will choose an exam part and the topic they want, then the module will generate a test that satisfies the user's choice.
- **Reading:** All objectives have been achieved, and the module works as expected, the user will choose an exam part and the topic they want, then the module will generate a test that satisfies the user's choice.

Resources and Tools Used

- **Models**: OpenAI Whisper, XTTS v2, CEFR analyzer, Transformers, Qwen 2.5 vl 7B, glm 4.5 air.
- **Libraries**: HuggingFace, torchaudio, NLTK, pyannote, G2P
- **Datasets**: YouTube scraped transcripts, IELTS essays samples, Kaggle audio sets, SQuAD (Stanford Question Answering Dataset)

Methodology and approach

- **Writing Task 1:** The user will enter the image, image subject and the essay. The image will be passed to a vision language model (Qwen2.5-vl-32B in this case) that will write a paragraph which perfectly describes what is in the image, the user's essay, image subject and image description extracted by the vision language model will all be entered to another LLM (GLM 4.5 air)

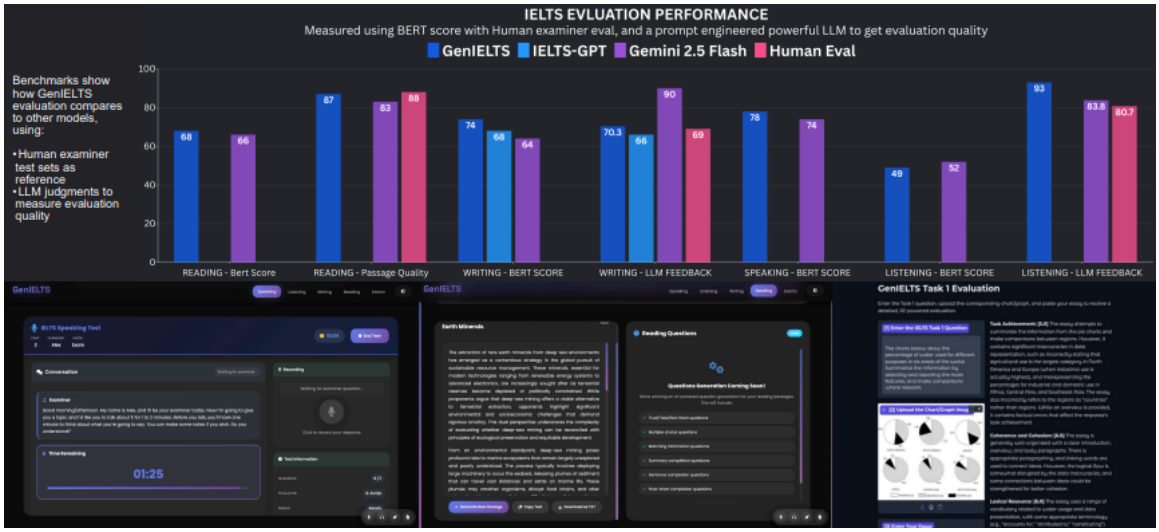
that will then write a detailed report for the essay. The report contains the sub-scores for each one of the criteria, examiner's feedback, strengths and weaknesses and areas of improvement.

- **Writing Task 2:** The user will enter the essay and the essay question. The essay will then be evaluated by a finetuned LLM (Qwen3-14B) that will be giving detailed evaluation report. The model was finetuned on a combination of many IELTS writing task 2 datasets from Huggingface, the problems that have been faced in this part of the project are that the model's generated sub-scores for each one of the criteria were highly inaccurate and inconsistent and that the data was not clean. To overcome the first problem a GRPO model was finetuned to predict the sub-scores and the overall score for the essay. Data has been cleaned and preprocessed so well which improved performance.
- **Listening:** We have collected the data using YTDLP library and downloaded videos from IELTS YouTube channels, then transcript the audio using WhisperX for speakers' distillation, then pass the whole data to LLM to generate the expected prompt that will generate the corresponding part of the exam. Then the model (Qwen3-14B) was finetuned on this data. The user will put the topic and part, it will pass them to LLM to generate the conversation, and it will take the conversation and pass it to bigger model to generate the questions and answers. The problem faced with this part of the exam was the lack of data and to reformat the available data.
- **Reading:** Data has been taken from Kaggle SQuAD2.0 dataset and the owner of it. Data has been reformatted into IELTS reading format using an LLM. This dataset was specifically chosen for this task because of its high level and advanced vocabulary which perfectly satisfies the IELTS reading passages. An LLM (Qwen3-14B) was then finetuned to generate passages. Another LLM was used to generate questions and their answers for the passages generated. One obstacle for this part was the lack of the data and novelty of the idea which made it hard to compare the model with other models doing the same task**.**
- **Speaking:** The system is designed to take a user's spoken response as input (in audio format) and produce a comprehensive IELTS-specific feedback package as output. The output includes:
    - Pronunciation and grammar feedback, aligned with IELTS descriptors.
    - Correct answers to the questions given for reference.
    - Evaluation and predicted band score, based on IELTS marking criteria.
    - Audio-based feedback, enabling users to listen to their evaluation.

To achieve this, the user's audio is processed through an IPA transcription model, an Automatic Speech Recognition (ASR) model, a phonemizer, and finally a finetuned Large Language Model (LLM) trained specifically for IELTS evaluation and feedback. The model was finetuned on IELTS speaking task data that has the taker's answer and examiner's feedback. A major challenge was the lack of publicly available datasets for this task. No open-source implementation existed that simulated the IELTS speaking evaluation in the way required for this project, which makes the idea novel. As a result, we collected our own dataset and adapted existing resources creatively to build a robust evaluation system. References used in this part were:

- The paper titled **Assessment of Non-Native Speech Intelligibility using Wav2vec2-based Mispronunciation Detection and Multi-level Goodness of Pronunciation Transformer.**
- The GitHub repository titled **crazycloud/mispronunciation-detection-diagnosis-wav2vec2-and-llm**: Mispronunciation Detection using a pretrained and finetuned wav2vec2 model for phoneme recognition and diagnosis and feedback using large language model
- IPA Transcription Leaderboard **-** a Hugging Face Space by KoelLabs.

Evaluation of all the models



Contributions

- We collected three new datasets (Reading dataset, listening dataset and speaking dataset)
- Our project is novel. It is by far the first complete IELTS simulation using AI.
- Our framework is verified against different models which makes it trustworthy.
- We used very efficient models and achieved excellent performance from them.

Conclusions

- GenIELTS is the first complete end-to-end AI IELTS simulation with novel capabilities. Compact modular system with models ≤ 14B parameters.
- GenIELTS delivers actionable feedback for measurable learner progress
- It Achieves near human-level evaluation, paving the way for more effective IELTS preparation.