AI-ELTS: Cost-Effective Essay Generation with Advanced GPT Architecture

An Dinh Ngoc Phuc Phan Van Hoa Dam Nguyen Quynh Van Nguyen Phuc Thanh Nguyen Phuoc

Supervisors
Hieu Tang Quang

FPT University

June 25, 2023

Overview

Introduction: IELTS Overview

IELTS is a widely recognized English examination used to evaluate a student's English proficiency.

[1] IELTS. *IELTS grows to 3.5 million a year*. Available: https://www.ielts.org/news/2019/ielts-grows-to-three-and-a-half-million-a-year.

[2] Nhon H Nguyen and Khoi D Nguyen. "Vietnamese Learners' Performance in The IELTS Writing Task 2". In: *Nguyen, HN, & Nguyen, DK (2022). Vietnamese Learners' Performance in The IELTS Writing Task 2 (2022), pp. 170–189*.

Introduction: IELTS Overview

IELTS is a widely recognized English examination used to evaluate a student's English proficiency.

As of 2018, over 3.5 million people are taking the IELTS test every year.[1]

[1] IELTS. *IELTS grows to 3.5 million a year*. Available: https://www.ielts.org/news/2019/ielts-grows-to-three-and-a-half-million-a-year.

[2] Nhon H Nguyen and Khoi D Nguyen. "Vietnamese Learners' Performance in The IELTS Writing Task 2". In: *Nguyen, HN, & Nguyen, DK (2022). Vietnamese Learners' Performance in The IELTS Writing Task 2 (2022), pp. 170–189*.

Introduction: IELTS Overview

IELTS is a widely recognized English examination used to evaluate a student's English proficiency.

As of 2018, over 3.5 million people are taking the IELTS test every year.[1]

Many universities in Vietnam are using the IELTS Score as the primary evaluation of English for enrollment.

[1] IELTS. *IELTS grows to 3.5 million a year*. Available: https://www.ielts.org/news/2019/ielts-grows-to-three-and-a-half-million-a-year.

[2] Nhon H Nguyen and Khoi D Nguyen. "Vietnamese Learners' Performance in The IELTS Writing Task 2". In: *Nguyen, HN, & Nguyen, DK (2022). Vietnamese Learners' Performance in The IELTS Writing Task 2 (2022), pp. 170–189*.

Introduction: IELTS Overview

IELTS is a widely recognized English examination used to evaluate a student's English proficiency.

As of 2018, over 3.5 million people are taking the IELTS test every year.[1]

Many universities in Vietnam are using the IELTS Score as the primary evaluation of English for enrollment.

Among the four Sections in IELTS: Reading, Listening, Speaking, and Writing, research has shown that Writing is the most difficult section to master.[2]

[1] IELTS. *IELTS grows to 3.5 million a year*. Available: https://www.ielts.org/news/2019/ielts-grows-to-three-and-a-half-million-a-year.

[2] Nhon H Nguyen and Khoi D Nguyen. "Vietnamese Learners' Performance in The IELTS Writing Task 2". In: *Nguyen, HN, & Nguyen, DK (2022). Vietnamese Learners' Performance in The IELTS Writing Task 2 (2022), pp. 170–189*.

Introduction: Task

Task

Develop a fast, reliable AI system based on Natural Language Processing, without the need of GPU, to assist students in writing a perfect essay for IELTS Writing Task 2.

Introduction: Task

Task

Develop a fast, reliable AI system based on Natural Language Processing, without the need of GPU, to assist students in writing a perfect essay for IELTS Writing Task 2.

Input and Output

1. **Input:** A string of sentence you currently have.
2. **Output:** A piece of text likely to come after the input text based on its context.

Introduction: Task

Task

Develop a fast, reliable AI system based on Natural Language Processing, without the need of GPU, to assist students in writing a perfect essay for IELTS Writing Task 2.

Input and Output

1. **Input:** A string of sentence you currently have.
2. **Output:** A piece of text likely to come after the input text based on its context.

Goal

The end goal is to help student develop an optimal idea to write in an essay when they do not know what to write next.

Introduction: Demo

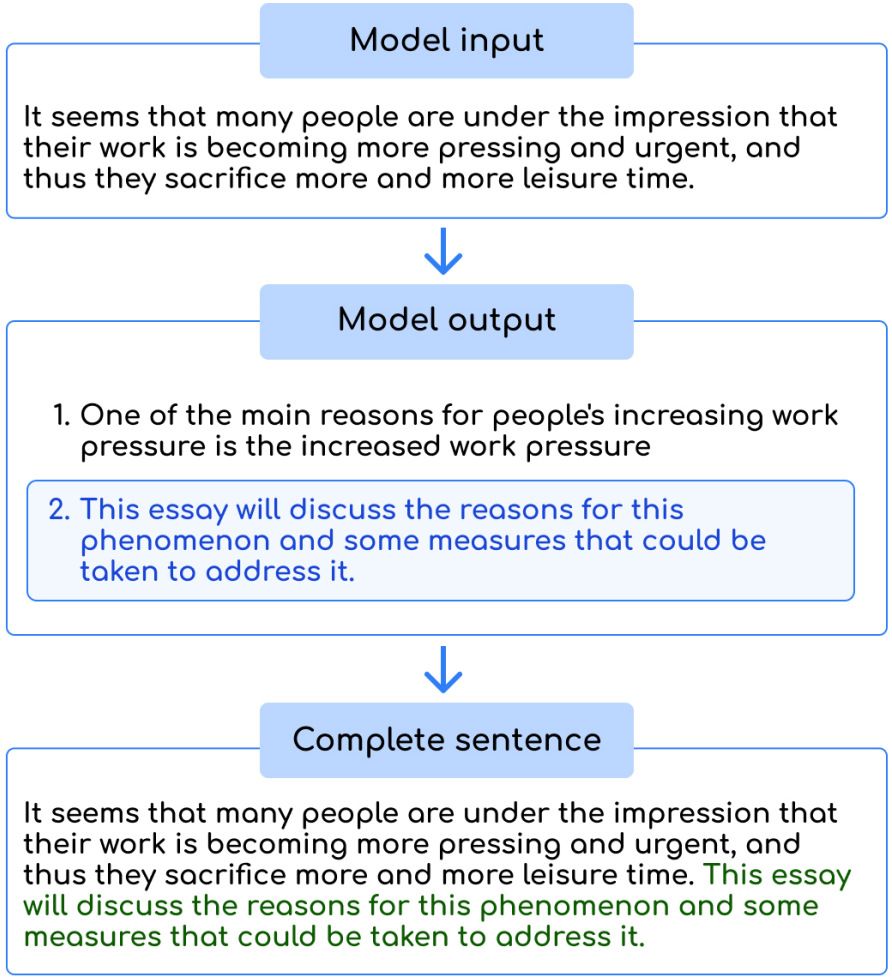

Figure 1: Model Visualization.

Related Work

GPT-3 (OpenAI)

- Demonstrates remarkable performance across various language tasks.

GPT-Neo (Eleuther AI)

- Computationally efficient alternative to GPT-3.
- Maintains strong language generation capabilities.

T5 (Google)

- Text-to-text framework, achieves remarkable results in different NLP tasks

LaMDA (Google)

- focuses on conversational abilities

Motivation

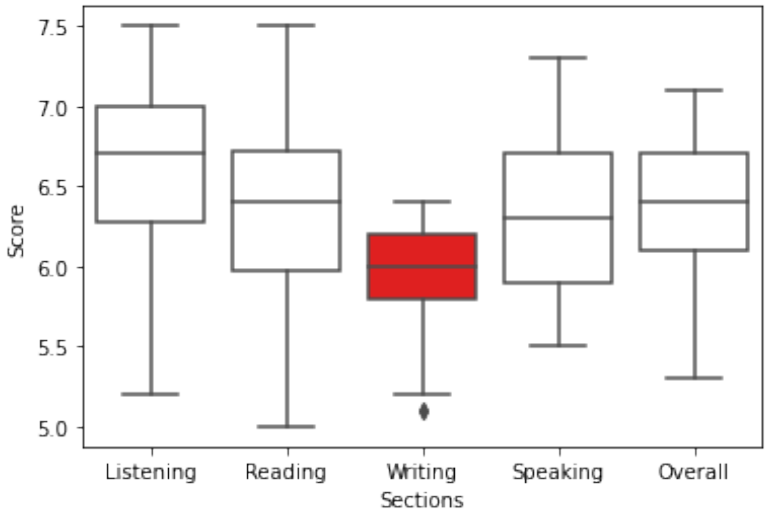

Figure 2: IELTS Academic mean performance by Nationality[a].

[a] **IELTS**. ***Test taker performance 2022***. **Available**: https://www.ielts.org/for-researchers/test-statistics/test-taker-performance. **(Visited on 2022)**.

Motivation

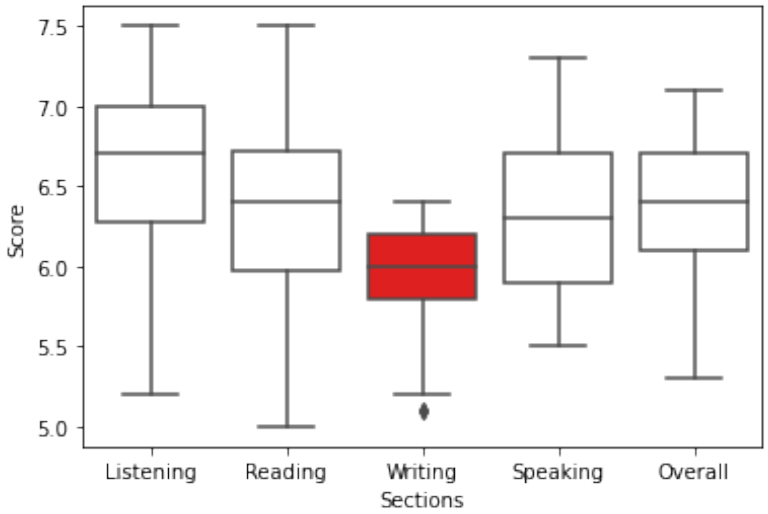

Figure 2: IELTS Academic mean performance by Nationality[a].

[a] **IELTS**. ***Test taker performance 2022***. **Available:** https://www.ielts.org/for-researchers/test-statistics/test-taker-performance. **(Visited on 2022)**.

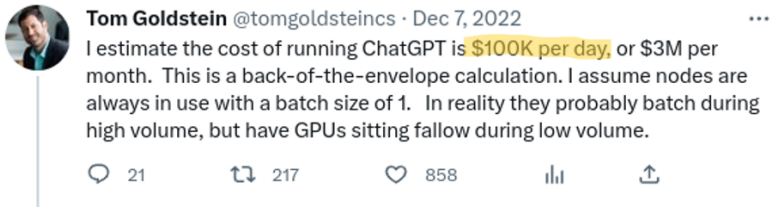

Figure 3: Costs of running ChatGPT

Motivation

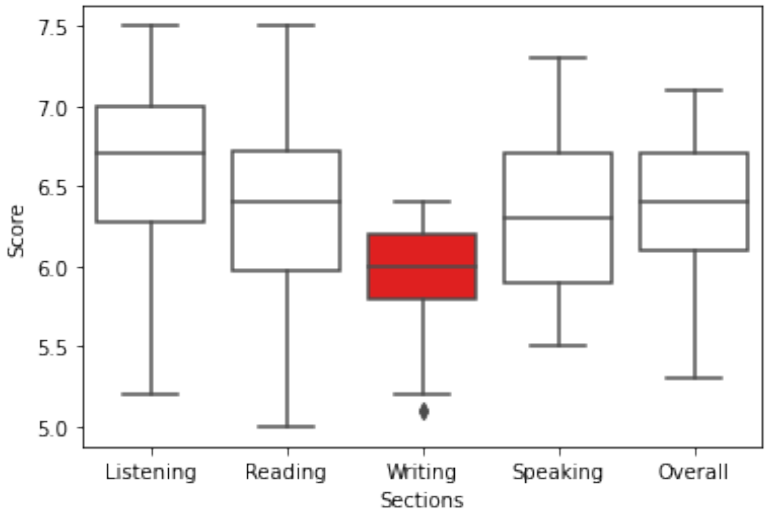

Figure 2: IELTS Academic mean performance by Nationality[a].

[a] **IELTS**. ***Test taker performance 2022***. **Available:** https://www.ielts.org/for-researchers/test-statistics/test-taker-performance. **(Visited on 2022)**.

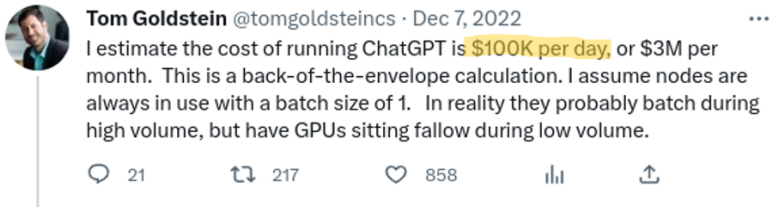

Figure 3: Costs of running ChatGPT

Conclusion

The model needs to be academic and reliable to be trusted by students, and using as little resources as possible (such as 1.3B parameters).

Practical Use

Use Cases

1. Suggest next sentences in an IELTS Writing essay.
2. Help write a good thesis essay for university enrollment.
3. Prepare good cover letter and Curriculum Vitae (CV).
4. ...

System Overview

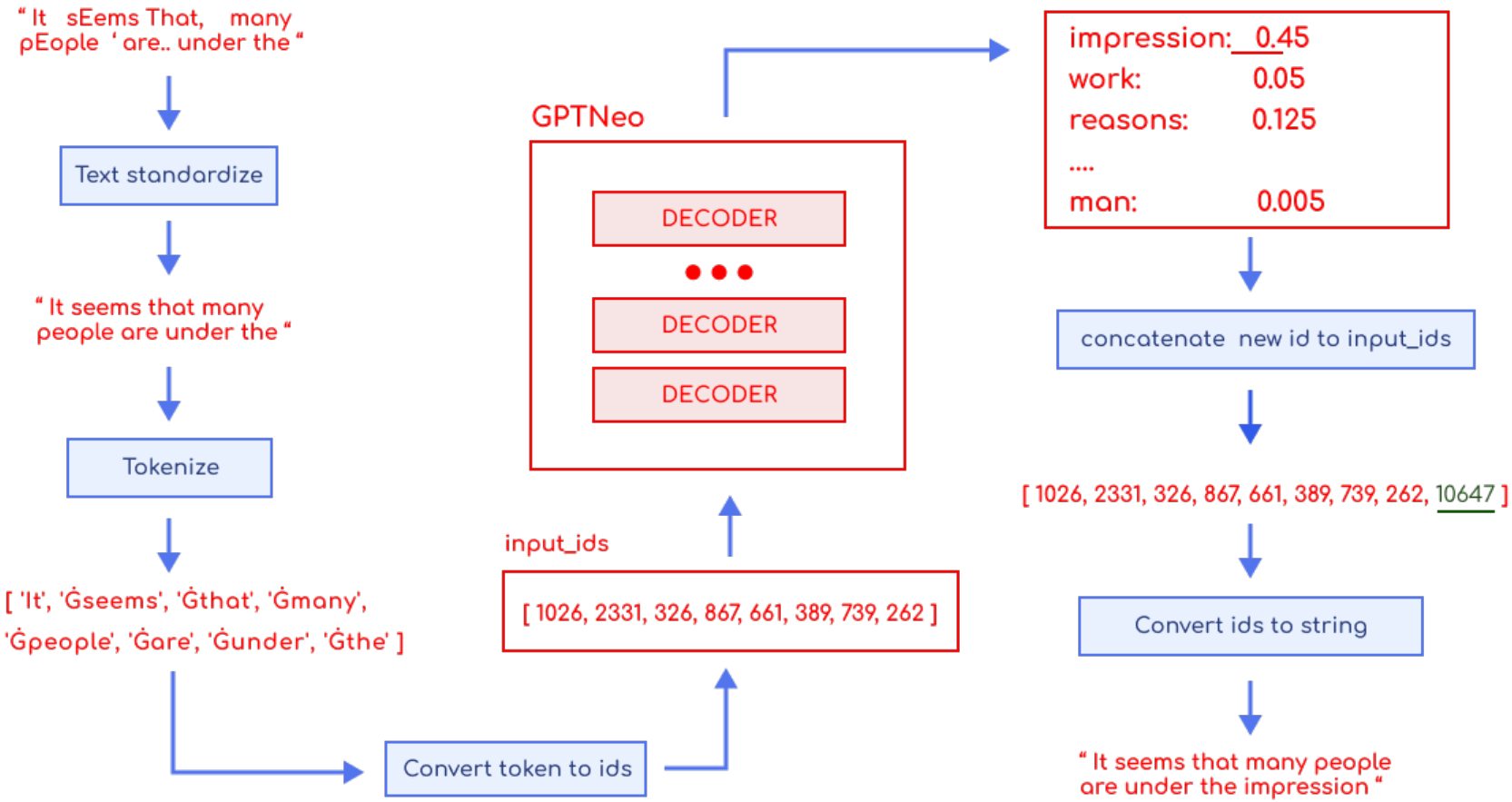

Figure 4: Full Model Architecture

Comparision of LLM Models: Data Usage

Most 3 common data types:

Webpages

- LLMs leverage webpages to acquire diverse linguistic knowledge and enhance performance.
- They can contain both high-quality and low-quality content, so filtering is necessary.

Comparision of LLM Models: Data Usage

Most 3 common data types:

Webpages

- LLMs leverage webpages to acquire diverse linguistic knowledge and enhance performance.
- They can contain both high-quality and low-quality content, so filtering is necessary.

Conversation Text

- Includes conversation data, improves conversational competence and question-answering performance.
- Can mistake instructions for conversation starters.

Comparision of LLM Models: Data Usage

Most 3 common data types:

Webpages

- LLMs leverage webpages to acquire diverse linguistic knowledge and enhance performance.
- They can contain both high-quality and low-quality content, so filtering is necessary.

Conversation Text

- Includes conversation data, improves conversational competence and question-answering performance.
- Can mistake instructions for conversation starters.

Books

- Books offer valuable linguistic knowledge, model long-term dependencies, and support coherent narrative generation.

Comparison of LLM Models: Data Usage

- We have chosen GPT-Neo because it has been pre-trained on a diverse range of data sources, with a significant emphasis on scientific data.

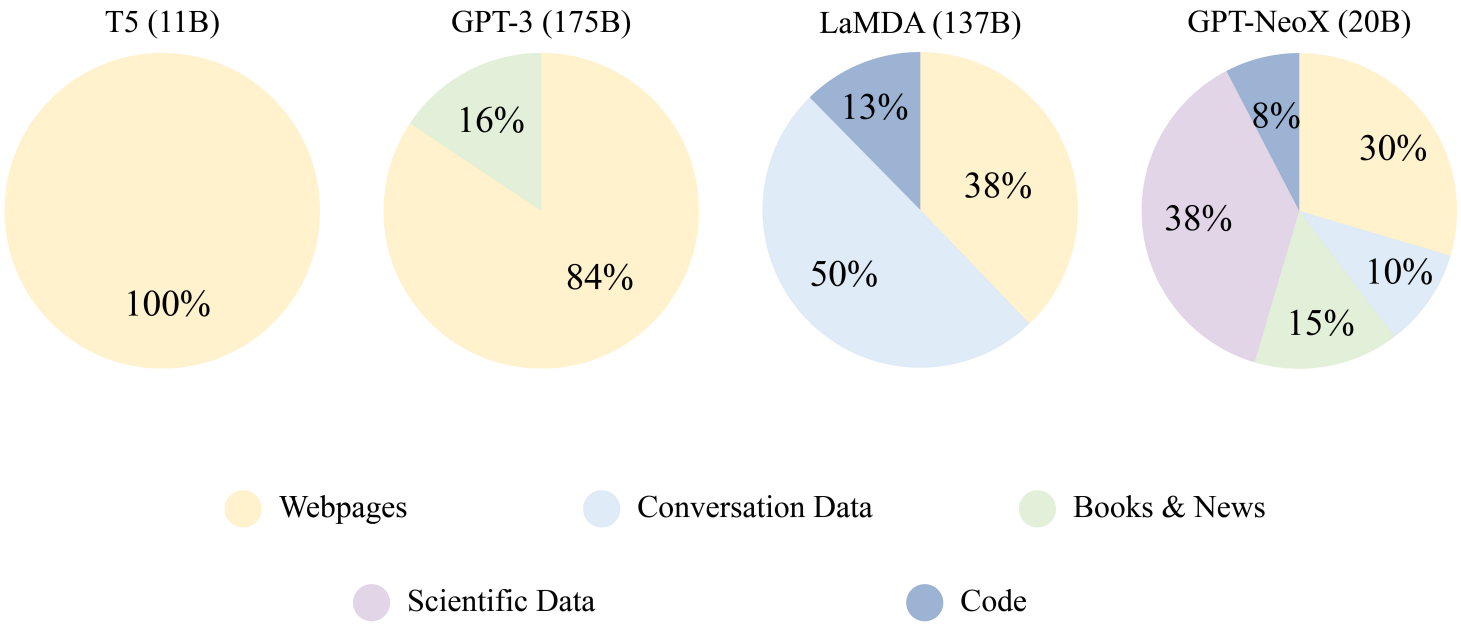

Figure 5: Data usage of each model.

Model Development: Data Preparation

- Gathered over 5000 passages from primary sources, including ChatGPT, IELTS Writing samples, and news articles. The training set - test set is 80% - 20% respectively.

Model Development: Data Preparation

- Gathered over 5000 passages from primary sources, including ChatGPT, IELTS Writing samples, and news articles. The training set - test set is 80% - 20% respectively.

- Balanced the data to ensure equal representation of each category, avoiding issues with imbalanced datasets, removed instances containing toxic or racist content.

Model Development: Data Preparation

- Gathered over 5000 passages from primary sources, including ChatGPT, IELTS Writing samples, and news articles. The training set - test set is 80% - 20% respectively.
- Balanced the data to ensure equal representation of each category, avoiding issues with imbalanced datasets, removed instances containing toxic or racist content.
- Set a maximum input length of 256 tokens for smoother model training and optimized memory usage using sliding window.

Model Development: Data Preparation

- Gathered over 5000 passages from primary sources, including ChatGPT, IELTS Writing samples, and news articles. The training set - test set is 80% - 20% respectively.
- Balanced the data to ensure equal representation of each category, avoiding issues with imbalanced datasets, removed instances containing toxic or racist content.
- Set a maximum input length of 256 tokens for smoother model training and optimized memory usage using sliding window.
- One IELTS essay divided into parts, input to model is a sentence and the label is next parts of input sentence.

Model Development: Training details

Hyperparameter	Value
Update Steps	7020
Batch Size	32
Warmup Steps	50
Optimizer	AdamW
β_1	0.9
β_2	0.999
ϵ	1×10^{-6}
Learning Rate	3×10^{-4}
Learning Rate Scheduler	Linear Decay
Loss	Cross Entropy
Weight Decay	0

Table 1: Hyperparameters

- Hardware: GPU A40.
- Using mixed precision training to increase the batch size, conserve memory and speed up training process.
- Almost all knowledge in LLM is learned during pre-training[a], fine-tune is conformed to a specific style or format. So that we just trained on few epochs (10 epochs).
- In experiment, training more than 10 epochs with batch size 32 makes model forget knowledge that has been learned in pre-training.

[a] Chunting Zhou et al. *LIMA: Less Is More for Alignment*. 2023. arXiv: 2305.11206 [cs.CL].

Model Development: Evaluation

There is no official metric for IELTS Writing evaluation, so alternatives are used.

[3] Lower loss is better, higher score is better.

Model Development: Evaluation

There is no official metric for IELTS Writing evaluation, so alternatives are used.

Model name	Train loss	Validation loss	BLEU score	ROUGH score
GPT2-124M	1.7742	2.004801	0.231548	0.497854
GPT2-355M	0.6788	1.792658	0.370100	0.607069
GPT2-774M	0.2429	**1.336823**	0.567848	0.708003
GPT-Neo-125M	1.3602	2.307180	0.288826	0.516954
GPT-Neo-1.3B	**0.1776**	1.743037	**0.600481**	**0.720475**

Table 2: Evaluate score on BLEU and ROUGE[3].

[3] Lower loss is better, higher score is better.

Model Development: Evaluation

There is no official metric for IELTS Writing evaluation, so alternatives are used.

Model name	Train loss	Validation loss	BLEU score	ROUGH score
GPT2-124M	1.7742	2.004801	0.231548	0.497854
GPT2-355M	0.6788	1.792658	0.370100	0.607069
GPT2-774M	0.2429	**1.336823**	0.567848	0.708003
GPT-Neo-125M	1.3602	2.307180	0.288826	0.516954
GPT-Neo-1.3B	**0.1776**	1.743037	**0.600481**	**0.720475**

Table 2: Evaluate score on BLEU and ROUGE[3].

BLEU Equation

$$BLEU = \min\left(1, \frac{\|output\|}{\|reference\|}\right)\left(\Pi_{i=1}^{4} precision_i\right)^{0.25}$$

[3] Lower loss is better, higher score is better.

Model Development: Evaluation

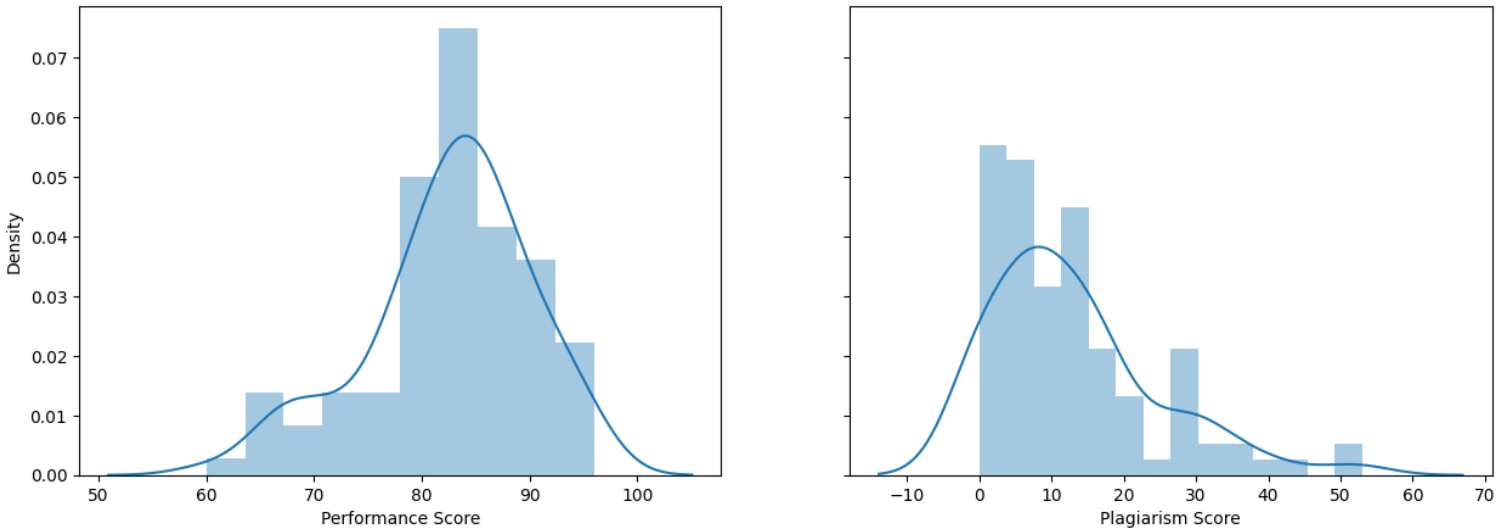

Figure 6: Evaluation on Grammarly tool.

On average, the performance score is 82.53% and plagiarism score is 12.95%

Deployment

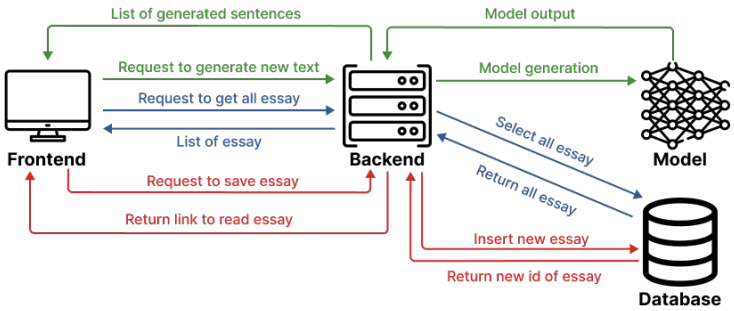

Figure 7: Deployment Process

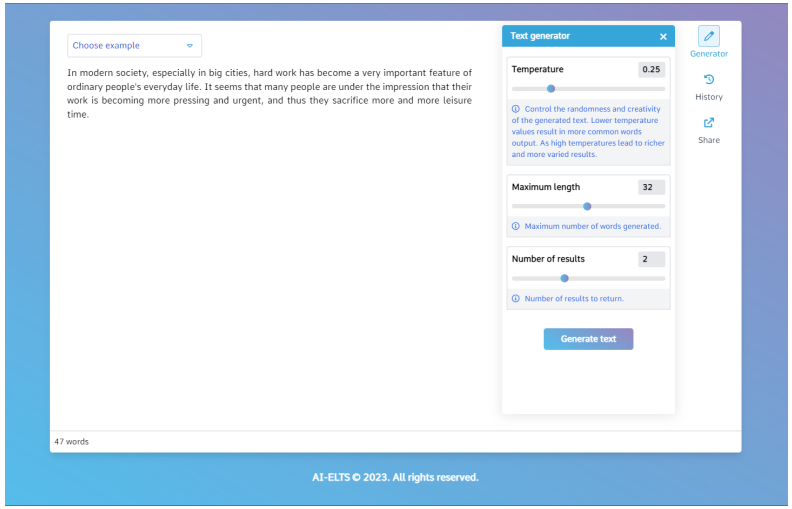

Figure 8: Web App

Deployment

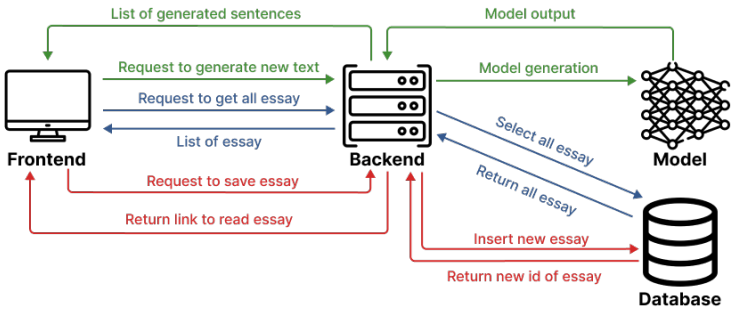

Figure 7: Deployment Process

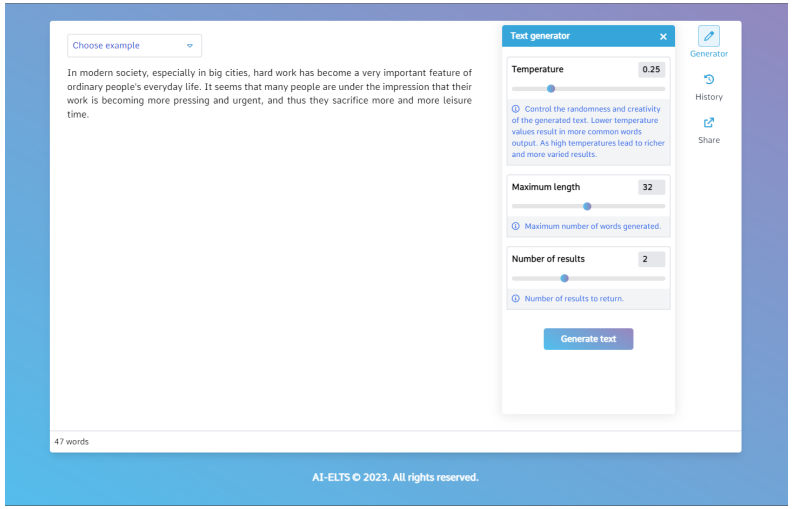

Figure 8: Web App

The output will be generated using CPU power for reduced cost.

Deployment

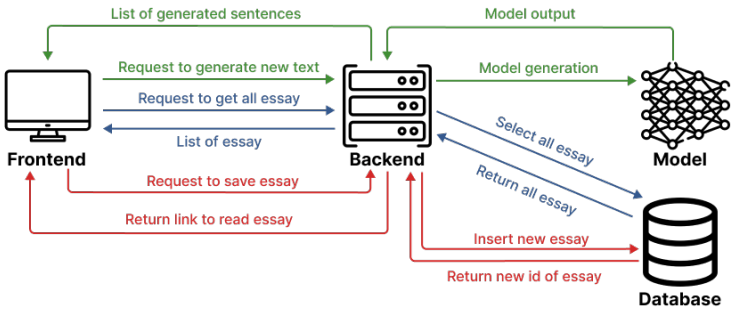

Figure 7: Deployment Process

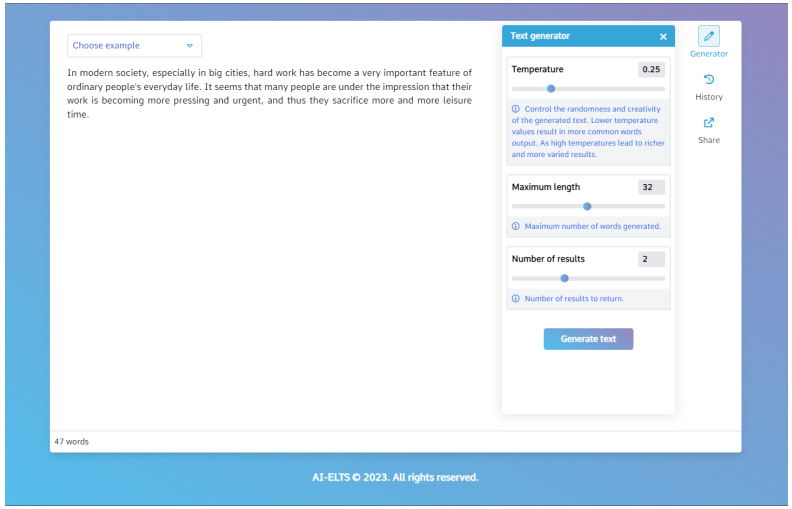

Figure 8: Web App

The output will be generated using CPU power for reduced cost.

The user can tweak many options on the Web: Temperature, Maximum length, Number of results.

Limitation and Future Work

Lack of Official Metric

- Focus on developing a evaluation metric to accurately the appropriate band

Limitation and Future Work

Lack of Official Metric

- Focus on developing a evaluation metric to accurately the appropriate band

Data Quantity and Quality

- Increase the size of dataset, more quality and introduce more diverse prompts.

Limitation and Future Work

Lack of Official Metric

- Focus on developing a evaluation metric to accurately the appropriate band

Data Quantity and Quality

- Increase the size of dataset, more quality and introduce more diverse prompts.

Model still not answering well on suitable style

- Fine-tune the model using reinforcement learning techniques, leveraging human ranking feedback to further enhance its accuracy and responsiveness.

Limitation and Future Work

Lack of Official Metric

- Focus on developing a evaluation metric to accurately the appropriate band

Data Quantity and Quality

- Increase the size of dataset, more quality and introduce more diverse prompts.

Model still not answering well on suitable style

- Fine-tune the model using reinforcement learning techniques, leveraging human ranking feedback to further enhance its accuracy and responsiveness.

Can not generate the ending sentence by user request

- We acknowledge the need to improve upon this aspect.

Conclusion

- Conducted experiments on GPT-Neo to generate the next sentence in an IELTS text.

Conclusion

- Conducted experiments on GPT-Neo to generate the next sentence in an IELTS text.
- The model can perform inference without the needs of strong hardware.

Conclusion

- Conducted experiments on GPT-Neo to generate the next sentence in an IELTS text.
- The model can perform inference without the needs of strong hardware.
- The generated text can avoid plagiarism and achieves high score on Grammarly check.

Conclusion

- Conducted experiments on GPT-Neo to generate the next sentence in an IELTS text.
- The model can perform inference without the needs of strong hardware.
- The generated text can avoid plagiarism and achieves high score on Grammarly check.
- With more research and experiments, the model can develop to be the next big thing in the world of IELTS and Education.

References

[1] IELTS. *IELTS grows to 3.5 million a year*. Available: https://www.ielts.org/news/2019/ielts-grows-to-three-and-a-half-million-a-year.

[2] IELTS. *Test taker performance 2022*. Available: https://www.ielts.org/for-researchers/test-statistics/test-taker-performance. (Visited on 2022).

[3] Nhon H Nguyen and Khoi D Nguyen. "Vietnamese Learners' Performance in The IELTS Writing Task 2". In: *Nguyen, HN, & Nguyen, DK (2022). Vietnamese Learners' Performance in The IELTS Writing Task 2* (2022), pp. 170–189.

[4] Chunting Zhou et al. *LIMA: Less Is More for Alignment*. 2023. arXiv: 2305.11206 [cs.CL].

The End