A COMPARATIVE STUDY OF TEXT COMPREHENSION IN IELTS READING EXAM USING GPT-3

Christopher Le[1], Tauheed Khan Mohd[2]

[1]Dept. of Math and Computer Science, Augustana College, IL USA
[2]School of Information Security and Applied Computing, Eastern Michigan University, Ypsilanti, MI

ABSTRACT

This paper discusses the capabilities and limitations of GPT-3 (0), a state-of-the-art language model, in the context of text understanding. We begin by describing the architecture and training process of GPT-3, and provide an overview of its impressive performance across a wide range of natural language processing tasks, such as language translation, question-answering, and text completion. Throughout this research project, a summarizing tool was also created to help us retrieve content from any types of document, specifically IELTS (0) Reading Test data in this project. We also aimed to improve the accuracy of the summarizing, as well as question-answering capabilities of GPT-3 (0) via long text.

KEYWORDS

GPT, Artificial Intelligence, Natural language processing.

1. INTRODUCTION

In recent years, natural language processing (NLP) has seen significant advances, thanks to the development of large-scale language models such as BERT (Bidirectional Encoder Representations from Transformers) from Google, RoBERTa (Robustly Optimized BERT approach) from Meta, and GPT (Generative Pre-trained Transformer) from OpenAI. These models have been a breakthrough in enabling machines to understand, interpret, and generate natural languages in a way that is similar to human communication. Thus, they brought benefits in a variety of fields, including healthcare, finance, marketing, and education.

However, the success of NLP also raises important ethical and societal concerns, including issues related to bias, privacy, and automation of tasks that were previously the domain of human expertise. Therefore, there is a need to explore the capabilities and limitations of GPT-3 and critically examine its ethical implications.

In this paper, we aim to provide a comprehensive overview of GPT-3's capabilities and limitations in the context of text understanding. We will use the dataset of IELTS (0) Reading test to test GPT-3's capabilities of summarizing long text, text understanding, and finally question-answering. In the end, a summarizing tool that we created will also be discussed to test and improve the current model on text understanding. In order to do this, we will need to understand the architecture and training process of GPT-3 and related models, thus figuring out its challenges or limitations of text understanding.We will also analyze the ethical implications of such models like GPT-3 and provide a roadmap for future research in the field of NLP.

The ability to summarize a lengthy text is crucial for anyone who works in the media, business, or education since we live in a fast-paced world where we must constantly ingest vast amounts of information. GPT-3 (Generative Pre-trained Transformer 3) is one such model that has the potential to replace humans in this time-consuming and challenging work. Fortunately, the development of natural language processing (NLP) models has given the ability to automate text summarization. One of the tasks that GPT-3 has shown particular proficiency in is text summarization. GPT-3 can summarize lengthy texts by identifying the most necessary information while presenting it concisely and coherently.

The study will involve comparing the performance of summaries generated by GPT-3 to those performed by human summarizers using a sample of long documents. The main target of this research is the IELTS Reading Test, one of the four sections of the International English Language Testing System (IELTS) exam that assesses non-native speakers' ability to understand and analyze written texts.

The reason IELTS is an ideal research target is that GPT-3 has been shown to perform remarkably well on question answering and text completion tasks, which are equivalent to multiple choice, yes/no/not given, and filling in the blank tasks in the IELTS Reading test. Furthermore, IELTS Reading tests cover a wide range of topics, including natural science, social studies, business, and global issues. Lastly, IELTS is a widely recognized English language proficiency test used for several purposes, including applying to study abroad or finding employment.

Analyzing the strengths and limitations of GPT-3 in this context can help us to gain valuable insight into the potential implications of using NLP models like GPT-3 to automate language proficiency testing and assessment. If GPT-3 proves to be an effective and reliable tool for summarizing and comprehending lengthy texts, it could significantly make the collecting information process more efficient and accessible. Moreover, this research could also inform the development of more sophisticated language testing tools that incorporate NLP models and other advanced technologies.

2. RELATED WORKS

In comparison to GPT-3, which is the main focus of this research, there have been other great LLM models from Google, and Meta: BERT and Roberta respectively. It's also important to understand transformer and attention, the foundation behind GPT model.

2.1. Large Language Model

Large Language Model (LLM) (0) is a type of artificial intelligence that uses deep learning techniques to analyze and understand natural language. LLMs are designed to learn from massive amounts of text data and use that knowledge to generate human-like responses to questions or statements. These models have been trained on vast corpus of text data to store world knowledge within their neural network weights. LLMs have many practical applications in natural language processing, such as machine translation, speech recognition, and text generation. Despite their potential, there are also concerns about their ethical and social implications, including issues related to bias, privacy, and misinformation.

Top LLMs known at this moment are: ChatGPT (OpenAI) (?), and BARD (Google), and LLaMA(0) (Meta).

2.2. Transformer

In 2017, Vaswani et al. introduced the Transformer (0), a potent deep-learning model. Natural language processing (NLP) has undergone a revolution as a result of its ability to create models to analyze collections of inputs or outputs (text, speech, etc.). The application of the Transformer (0) was seen in many stateof-the-art NLP tasks, including language translation, language modeling, and question answering, with unprecedented levels of accuracy.

What sets the Transformer (0) apart from previous NLP models is its use of attention mechanisms. The Transformer (0) model uses self-attention to enable the processing of sequences of variable length without relying on sequential processing like traditional recurrent neural networks. This allows the model to be much more efficient and effective at processing long sequences of text or speech. Since then, The Transformer (0) has become a foundational model in the field of NLP and impacts several popular deep learning frameworks, including TensorFlow (0), PyTorch (0), and Keras(0). The Transformer (0) is undeniably an active area of research and development because it has created a wide range of opportunities for NLP.

2.3. Attention

The attention mechanism is a powerful concept in deep learning that has gained widespread use in various natural language processing (NLP) applications. It enables the model to focus on important features in the input sequence while processing it, which can significantly improve the model's performance (0). The fundamental principle of attention mechanisms is to emphasize the pertinent portions of the input sequence while dismissing the unimportant ones. It enables the model to develop the ability to selectively attend to various input components, which is very advantageous when processing lengthy sequences.

Many NLP tasks, including machine translation, text categorization, question answering, and text summarization, have proven to perform much better when attention mechanisms are applied. More specifically, compared to conventional models like recurrent neural networks, attention mechanisms have made it possible for models to handle lengthy input sequences more successfully (RNNs (0)). Attention mechanisms have evolved with different types and variations, including self-attention and multi-head attention. These processes have allowed for the creation of cutting-edge models and have grown to be a necessary component of the NLP toolkit in the field.

2.4. BERT

Bidirectional Encoder Representations from Transformers (BERT (0)) is an effective deep learning model for natural language processing (NLP) that was introduced by Google in 2018. This pre-trained language model can be fine-tuned for several NLP tasks, such as text summarization and text comprehension. BERT's capacity to comprehend context by taking into account the words that surround a word in a phrase in both directions sets it apart from other language models. This is accomplished via a novel pre-training method that trains the model on a large corpus of text data using a masked language model and a task for predicting the following phrase. For a variety of NLP tasks, such as question answering, named entity identification, and sentiment analysis, BERT has produced state-of-the-art results. While it can properly capture the meaning of a sentence or document, it has also demonstrated significant potential for text summarization and text interpretation. BERT (0) can be fine-tuned for text summarizing to provide accurate and informative summaries of lengthy articles or documents. It can accomplish this by learning to identify and summarize the text's key points. BERT may be used in text comprehension to examine and extract the meaning from a sentence or document, allowing the

model to more accurately respond to queries, categorize text, or carry out other NLP activities. Overall, BERT (0) is a formidable NLP tool that might revolutionize how we interpret and analyze textual data. Text summarization, text interpretation, and other NLP tasks now have more options thanks to their capacity to capture the meaning and context of language.

2.5. RoBERTa (0)

RoBERTa (0) (Robustly Optimized BERT Approach) is a state-of-the-art natural language processing (NLP) model that was introduced by Facebook AI in 2019. RoBERTa (0) is based on the same architecture as BERT, but it has been trained with improved pre-training techniques and larger amounts of data, resulting in a significant improvement in performance on a range of NLP tasks. One of the key differences between RoBERT (0) and BERT (0) is the training data. RoBERTa (0)(0)(0) was trained on a much larger corpus of text, with up to 160 GB of training data, compared to 16 GB for BERT. This larger corpus of text allowed RoBERTa (0)(0) to develop a more nuanced understanding of language and improve its performance on a range of NLP tasks. In addition to the larger corpus, RoBERTa (0) also uses a more advanced pre-training approach. It removes the next sentence prediction task used in BERT and instead focuses on optimizing the masked language modeling objective. It also uses dynamic masking, which randomly masks different tokens during training rather than masking the same tokens in every epoch. These improvements lead to better generalization and improved performance on a range of NLP tasks. RoBERTa (0) has achieved state-of-the-art results on a range of NLP benchmarks, including the GLUE benchmark, which evaluates a model's performance on a range of natural language understanding tasks. It has also shown impressive results in text classification, named entity recognition, and question-answering tasks. Overall, RoBERTa (0) represents a significant improvement in the performance of NLP models, and its success has led to the development of many other models based on its architecture and pre-training techniques. It has opened up new possibilities for NLP tasks, and its impact can be seen in many popular NLP applications today.

3. RESEARCH

3.1. How GPT-3 was Created

GPT-3 (Generative Pre-trained Transformer 3) is a language model created by OpenAI that uses deep learning techniques to generate human-like text. GPT-3 bases on 3 main stages: Pre-training: GPT-3 first goes through unsupervised training with its massive, internet-harvested set of text data (such as Wikipedia, websites, articles, etc.) using a transformer-based neural network architecture. Transformer-based (0) is a type of neural network architecture used in natural language processing (NLP) applications (language translation, sentiment analysis, and text creation). The concept of self-attention is the foundation of this model where each word in a sequence is compared to all the other words in the same sequence to calculate a weighted representation of the input sequence. This makes NLP be able to capture context and word relationships more effectively than typical recurrent neural network (RNN) architectures. Moreover, the Transformer (0) model consists of an encoder and a decoder, each includes several layers of multi-head self-attention and feedforward neural networks. During training, the model is optimized to predict the correct output given an input sequence. Fine-tuning: After being pre-trained, GPT-3 may be fine-tuned for certain tasks, including text completion, translation, summarization, and question-answering. Fine-tuning involves adapting the pre-trained model to a specific task by providing it with labeled examples and adjusting the model's parameters to optimize performance. ChatGPT, for example, was fine-tuned with 175B weights.

Furthermore, ChatGPT used Transformers (0) and attention layer with tokens made up from a piece of text.

Inference: After fine-tuning, the model is ready to generate text. Given a prompt (such as a sentence or paragraph), GPT-3 uses its knowledge of language patterns and context to generate a plausible continuation of the prompt. GPT-3 can generate text in various styles and tones, from technical writing to creative writing, and can even mimic the writing style of a particular author.

3.2. Building a Summarizing Tool

In this project, we built a summarizing tool based on GPT. This tool was created with Streamlit, open-source app framework for Machine Learning and Data Science, and Python programming language. Firstly, we accessed the OpenAI website to obtain the API key for API calls and training. We also gather 100 different prompts from sample research paper, Youtube video transcript, and article with the responses of their summaries respectively.

After training and obtaining the finetuned model, the model was used as the core of our back end to do the prediction. In our tool, there are 4 types of documents used for summarizing: article, YouTube videos, research paper or PDF, and pure text. With article, we used "newspaper" library to read infomation from URL. The tool used scrapping technique to fetch article's content as well as their thumbnail picture, author(s). The article's content and thumbnail picture were kept for the output of this model. Likewise, with Youtube videos, a tool called "youtube_transcript_api" was used to fetch the content of the video. In this research, 100 videos with various length from 5 to 40 minutes were used to to be summarized by the tool. Next, with PDF/Research paper, we used "fitz" module to extract text from PDF.

In this research, we tested the model's ability to do IELTS reading test. Our IELTS (0) Reading exam dataset came from British Council with answer key. We have collected over 25 exams with more than 75 reading tasks. In the exam, there are multiple types of questions: Yes/No/Not Given question, Fill in the blank, one choice, multiple choice, select correct corresponding paragraphs.

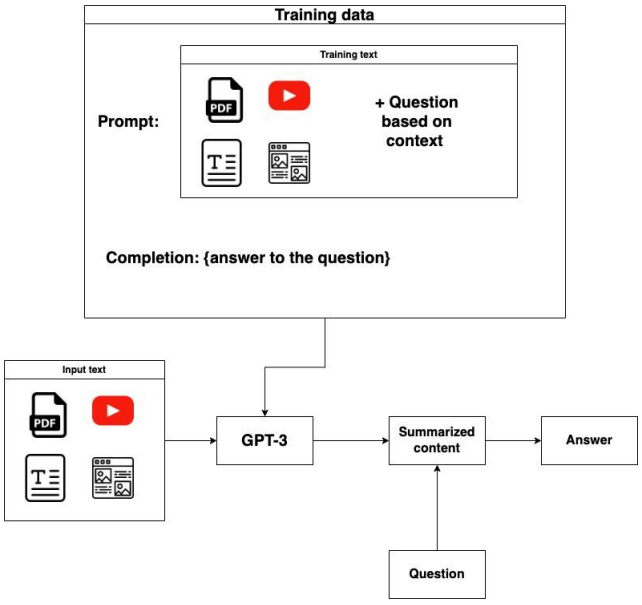

Figure 1: How the summarizing tool works

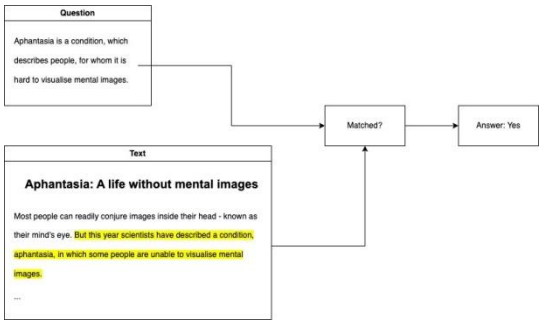

Figure 2: Example of question and answer process

3.3. Improving GPT's Text Understanding on Long Text

In other to summarize a long chunk of text, the token limit of ChatGPT is 4096. We divided the long text into multiple chunks of shorter texts of a maximum of 4096 tokens. Each short chunk of text went through summarizing to retrieve the key idea from the text. The key ideas were then connected to gain the big picture of the big idea. We also stored these short chunks of key ideas in a database, which then were used to answer questions from the text.

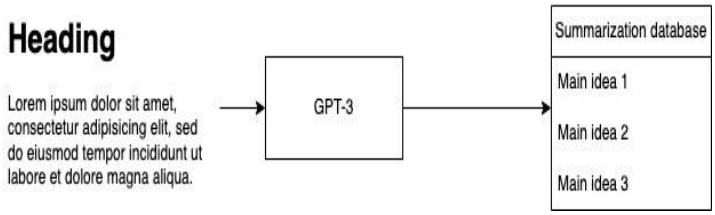

Figure 3: Summarizing database

In terms of answering questions based on the text, we used get embedding module from OpenAI to encode the question we wanted to ask from the text, each sentence, and each summarized main idea into word embeddings. Then, a comparison of the similarity of the embeddings was done. By doing the similarity comparison, it was able to answer the questions, which is similar to the way human do IELTS Reading Test by scan and skimming techniques.

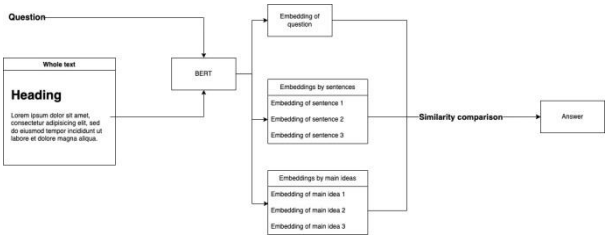

Figure 4: Similarity comparison

For visual inputs such as diagrams, flowcharts with questions such as fill in the blank. It was challenging to understand the context of the blank's location.

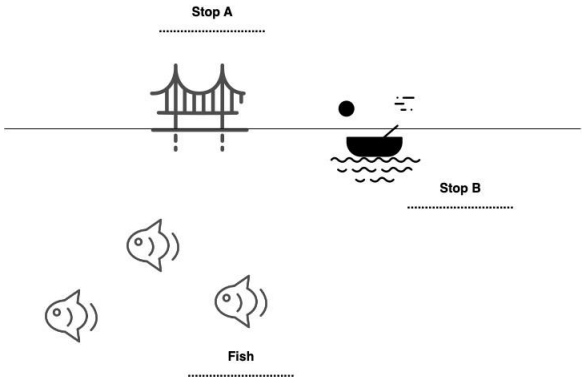

Figure 5: Example of visual fill-in-the-blank question

4. RESULTS

After testing GPT-3 on 50 IELTS Readings tests, we found that the average total accuracy of GPT-3 on one test is 65.38%, and the total accuracy of all 75 tasks is 58.70% based on the actual answers. This is equivalent to band 5.0-6.0 in IELTS. Our analysis of four main types of IELTS reading tests includes: Yes/ No/ Not Given, Fill in the blank, Matching paragraph, and Multiple choice. The result also suggested that GPT-3 performs best on Multiple choice questions.

Table 1: IELTS (0) Reading Result from GPT-3

Type of question	Accuracy
Yes/No/Not Given	75.85%
Fill in the blank	90.65%
Correct letter	6.18%
One choice	85.03%
Multiple choice	18.84%

5. FUTURE WORK

For further research, we would love to test the capability of the GPT-4 model from OpenAI. By taking visual inputs, some IELTS questions such as fill in the blanks in diagrams, flowcharts, and visuals would be possible. One way we thought of was to use OCR technology and imagine a captioning technique to describe the characteristics of a chart. In terms of imagine captioning, the AI model usually employs a neural network architecture, specifically a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). The CNN extracts visual features from the image, which are then fed into an RNN that generates the corresponding textual description. We are looking forward to training the model on the IELTS Reading Chart images.

REFERENCES

[1] Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.
[2] Council, "Ielts," 2016.
[3] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large language models in machine translation," 2007.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] M. Abadi, "Tensorflow: learning functions at scale," in *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*, pp. 1–1, 2016.

[7] S. Imambi, K. B. Prakash, and G. Kanagachidambaresan, "Pytorch," *Programming with TensorFlow: Solution for Edge Computing Applications*, pp. 87–104, 2021.

[8] N. Ketkar and N. Ketkar, "Introduction to keras," *Deep learning with python: a hands-on introduction*, pp. 97–111, 2017.

[9] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, pp. 64–67, 2001.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.