ECE 1786 Final Report

SimpliText

Written by

Peier Chen (1009223966)
Zirui Xu (1003031532)

Word Count: 1996, Penalty: 0%

Submitted to

Instructor: Jonathan Rose
Teaching Assistants: Mohamed Abdelwahab, Jiading Zhu
University of Toronto
Toronto, Ontario, Canada
Dec 12th, 2023

◆ Introduction

With the current iteration of the internet, the acquisition of knowledge has become easier, but for English beginners, there is currently a huge amount of online content that is too obscure to read. So, we came up with an idea: build a model based on GPT-4 which was developed by OpenAI. This model includes 3 functions to help those non-native English speakers to understand obscure English sentences: Simplification, Summarization, and Explanation. The reason to use GPT-4 is because we believe that the powerful capabilities of large language models for processing language should be able to meet our goal.

◆ Background & Related Work

We found 2 documents to serve as references. The first is a 2010 paper named 'Text Simplification for Children' written by J De Belder and MF Moens. The authors propose a method that includes syntactic simplification, such as splitting sentences, and lexical simplification, such as replacing difficult words with easier synonyms. We also use this method in the simplification function.

The second article is titled 'Controllable Text Simplification with Lexical Constraint Loss' written by D Nishihara, T Kajiwara, and Y Arase and published in 2019. The paper discusses a method for controlling the level of a sentence in a text simplification task.

◆ Data Collection & Processing

In terms of data collection, we picked 100 complex sentences from various aspects. Some of the sentences are from academic papers and articles, and some are extracted from English tests, such as IELTS, TOEFL, etc. These sentences may contain obscure words, complex structures, and excessive sentence length. Here is an example of sentences we collected:

" If successful, virus therapy could eventually form a third pillar alongside radiotherapy and chemotherapy in the standard arsenal cancer, while avoiding some of the debilitating side-effects. "

Based on our model, there are three different outputs from three instances, so our input and output data structures are as follows:

Table 1: Input and output data structure

Input	Output		
100 complex sentences extracted from the internet	100 simplified sentences	100 summarization of input sentences	100 explanations of input sentences

◆ Architecture & Software

Our system architecture is structured as three independent pipelines, each dedicated to a specific NLP task: Simplification, Summarization, and Explanation. Each pipeline is composed of a series of GPT-4 instances, each configured to perform a step in the process as outlined in the flowchart.

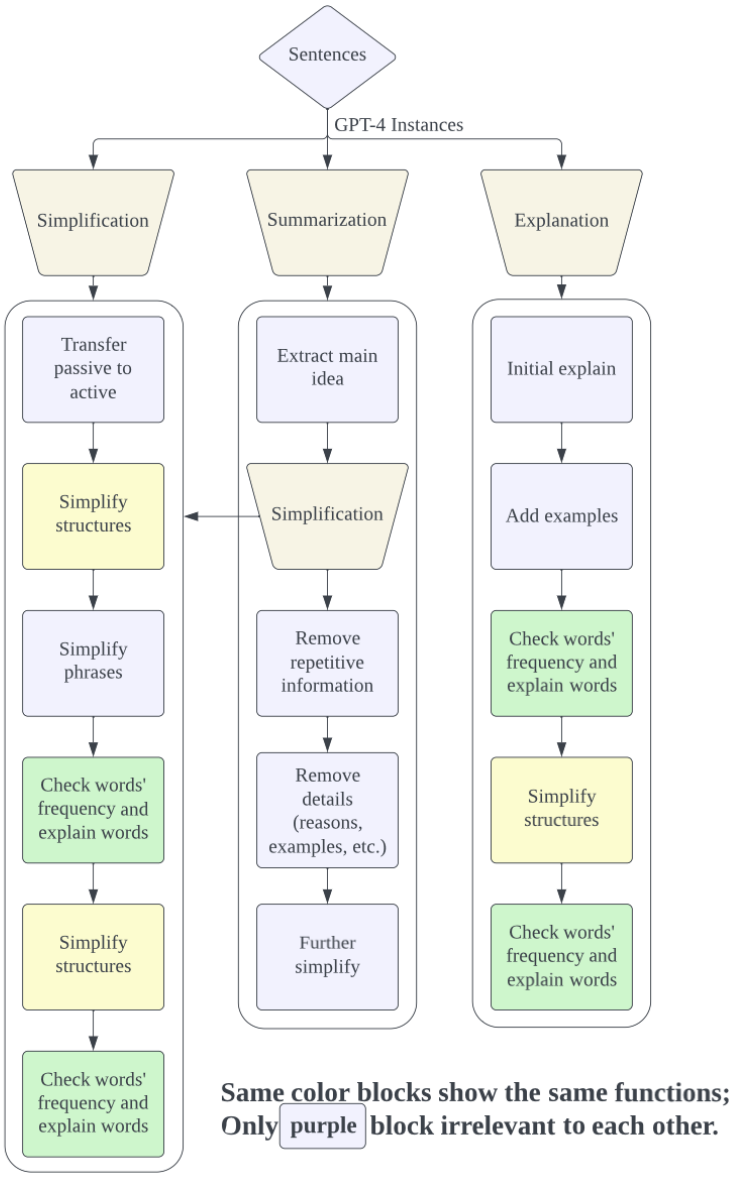


Figure 1: System Architecture

1. Pipeline for Simplification

The Simplification Pipeline is made up of several GPT-4 instances that are intended to gradually transform text into a comprehensible format. The process begins by transforming passive structures into active voices to enhance clarity. It then simplifies complex sentence structures into more straightforward ones by breaking complex sentences into simple sentences. Next, it focuses on simplifying phrases and employing the spaCy library for word frequency analysis to identify and replace fewer common words with a frequency threshold below 1e-4. This results in the substitution of more widely used synonyms for these words. Simplifying sentence structures and words are iterated several times to guarantee that the readability of the text is maintained through phrase simplification and subsequent word substitutions. This iterative refinement ensures that complexity is not unintentionally introduced in a different form throughout the reduction process. The result is text that is both simple and faithful to the original meaning, optimized for easy comprehension.

2. Summarization Pipeline

The first step of the Summarization Pipeline is a GPT-4 instance that is used to identify the primary themes and assertions in the text and extract their major ideas. After identifying these key elements, the pipeline employs our previously developed simplification model to further distill the main ideas into more comprehensible language. Then our model removes any repetitive elements and unnecessary details, ensuring that the summary remains focused and relevant. The final step involves a further simplification pass to refine the summary into its most accessible form. These steps result in a summary that is concise and understandable while preserving the original text's coherence and meaning.

3. Explanation Pipeline

The Explanation Pipeline begins by generating an initial explanation with a dedicated GPT-4 instance providing the foundational understanding. The next step is to add relevant examples to the explanation to help further illustrate and make the concepts clear. Next, to make sure that even the more obscure terms are understandable, the pipeline integrates a word frequency analysis using spaCy to find and then explain any words with a frequency lower than 1e-4. The subsequent GPT-4 instances then work on simplifying the sentence structures of these explanations, making them easier to follow. Explain words are repeated several times because simplifying sentence structures may produce difficult words. The result is an explanation that uses simple language and adds more details to make users more accessible.

◆ Baseline Model & Comparison

Our baseline model uses a single GPT-4 instance for each NLP task by simply and directly asking GPT to complete the task for us. For instance, we say 'Simplify/Summarize/Explain below sentences for English learner'. The outputs of the baseline model have some disadvantages such as long sentences, hard words, redundancy in summarization, and simple explanations, which can be fixed through further prompts design. That is why we're using GPT-4 as our baseline and refining it in our main model.

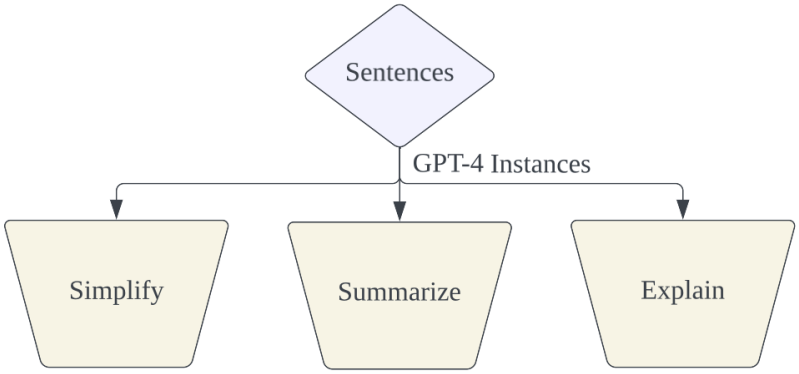


Figure 2: Baseline Model

◆ Quantitative Results

Every sentence we collected in our data collection part will be used for evaluation. The same number of results will be produced by the baseline model and our model, and we will compare each result from both models in the below metrics:

1) Numbers of words with a frequency lower than 1e-4: Words with a frequency lower than 1e-4 are typically hard words for English beginners—more numbers of words with a frequency less than 1e-4 mean that the sentence is harder to understand.

2) Minimum word frequency: Minimum word frequency indicates the hardest word in the sentence. So, the sentence with a lower minimum word frequency means that its hardest word is more difficult than another sentence.

3) Average sentence length: Normally, if a sentence is not that long, its sentence structure will be simple because if a sentence contains clauses, its length will increase. So average sentence length is shorter means that the sentence structure is simpler.

4) Total sentence length: A larger total sentence length indicates that the sentence may contain redundancies and does not summarize effectively.

5) Readability: We use Flesch Kincaid grade level to represent readability. Flesch Kincaid's grade level estimates the U.S. school grade level needed to understand the text. For example, a score of 8.0 means that an eighth-grader would be able to understand the text. So, the score is higher means that the readability is worse for English beginners.

6) Lexical Diversity: TTR is the ratio of the number of types to the number of tokens. A higher TTR indicates a greater variety of vocabulary, suggesting that the text uses a wider range of different words which means the sentence is more difficult than another.

7) HDD usually involves drawing a fixed number of word samples from the text and calculating the likelihood of encountering different types in these draws. A higher HDD means that the sentence is more complicated.

Table 2: Simplification evaluation results

Simplification Metrics	Better than the baseline model
Low frequency (1e-4) words number	86.87%
Minimum word frequency	86.87%
Average sentence length	96.97%
Readability	100%
TTR	81.82%
HDD	79.8%

Table 3: Summarization evaluation results

Summarization Metrics	Better than the baseline model
Low frequency (1e-4) words number	93.94%
Minimum word frequency	85.86%
Average sentence length	78.79%
Total sentence length	100%
Readability	61.62%

Table 4: Explanation evaluation results

Explanation Metrics	Better than the baseline model
Low frequency (1e-4) words number	92.93%
Minimum word frequency	83.84%
Average sentence length	58.59%
Readability	88.89%
TTR	55.56%

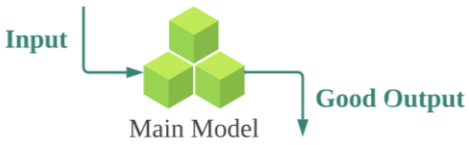
◆ Qualitative Results

The words in the following tables have been included in the total word count. We use those tables to express our results more clearly.

1) Simplification

Good output example

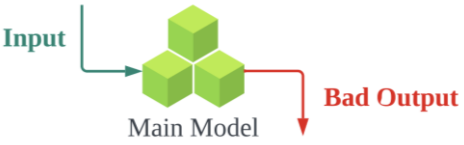
" Forest stands of red oaks, with their tannin-rich foliage, may even provide caterpillars with safe havens from disease."



" Red oak trees have a special thing in their leaves. This thing helps keep bugs safe from getting sick in the woods. "

Bad output example

" A constant failure rate (CFR) is indicative of externally induced failures. A CFR is also typical of complex systems subject to repair and overhaul, where different parts exhibit different patterns of failure with time and parts have different ages since repair or replacement. "



" Things often get damaged on the outside. Stuff might break and sometimes we can't repair it. Even when we repair things, they can still break because their pieces get old and worn out. The pieces don't all get old at the same time. "

The good example shows that our model produces a simpler sentence's structure and uses easy words which is the same as the quantitative result. But the bad example sentence is overly simplified by our model and the meaning changes. So, we evaluated the meaning produced manually to avoid oversimplification. And our model reaches an 88% success rate. Our model is not good at simplifying super academic sentences.

2) Summarization

Good output example

" Managing down operating costs within such a complex technology environment is extremely difficult. Retiring applications that are tightly coupled within the technical and data architecture results in nearly insurmountable dependency management issues and change risk. Retiring one application inevitably requires upstream and downstream changes and this added cost erodes the business case and often leads to a decision to leave things as they are. "

" Removing old software can be hard and might cause problems with other parts. It can be expensive, so people often don't change the system. "

Bad output example

" If Dickens's prose fiction has "defects"—excesses of melodrama, sentimentality, contrived plots, and manufactured happy endings—which for all his greatness Dickens had not the rebellious spirit to resist; he was at heart a crowd-pleaser, a theatrical entertainer, with no interest in subverting the conventions of the novel as his great successors D.H. Lawrence, James Joyce, and Virginia Woolf would have. "

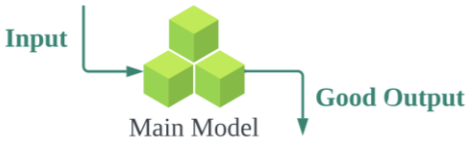
" Charles Dickens frequently wrote emotional scenes and happy endings. "

The good example uses simple language which is the same as the quantitative result. It extracts the main ideas and removes redundancies. But the bad example sentence doesn't contain all the important ideas. Both models do well in extracting key points and fidelity to the original intent. Our model is 100% better in non-redundancy. Overall, evaluated by us manually, we think our model reaches a success rate of 92% because sometimes our model may not be able to extract all the main ideas.

3) Explanation

Good output example

" However, Edwards believes the small changes in solar heating produced by Milankovitch cycles are then amplified by feedback mechanisms on Earth. "



" Edwards thinks that slight shifts in warmth from the sun can change our planet. The Earth's weather system, which includes loops that can make these shifts stronger or weaker, intensifies these changes. These loops are called feedback loops. They can make the sun's warmth changes powerful, causing big weather changes. For instance, when there's less ice, the sea soaks up more warmth. This soaking up can cause even more ice to melt. "

Bad output example

" Now recall that the underlying LLM's task, given the dialogue prompt followed by a piece of user-supplied text, is to generate a continuation that conforms to the distribution of the training data, which are the vast corpus of human-generated text on the Internet. "

" LLM: An LLM is a type of advanced law degree called a Master of Laws.

Contextually: This word means to look at something based on the situation or background it's in.

Datum: A datum is a single piece of information... "

The good example uses simple language which is the same as the quantitative result. It also adds more details such as an example of the red font. But the other example output is so bad. Firstly, it does not generate a coherent sentence, but several points and it misunderstood the meaning of 'LLM'. But in general, evaluated by us manually, we think our model reaches a success rate of 97%. Our explanation model performs well in academic sentences.

◆ Discussion & Learning

Based on the results, the project is a success. All the functions show a better performance than the baseline model. Among the 3 aspects, GPT-4 can do the best in the Summarization part. We think this reason is mainly due to GPT-4's attention mechanism and contextual understanding capabilities. The attention mechanism enables the model to weigh the importance of different words in a sentence concerning each other, which helps capture long-range dependencies and relationships within the text, crucial for understanding and summarizing content. For contextual understanding capabilities, the model can consider the entire context of a sentence or passage, allowing it to generate summaries that are contextually relevant and coherent.

But when dealing with simplification and explanation, it is easy to fail to produce a coherent sentence. We think this is because we attempt to simplify sentence structure by splitting sentences. Problems frequently occur when explaining obscure words. The sentence's meaning sometimes changed due to the GPT-4 replacing a wrong word with the original one. Overall, this project lets us know the GPT-4 advantages and disadvantages, which makes us understand deeply about this large language model.

Individual Contributions:

Peier Chen	Zirui Xu
Collected half of the dataset (50 rows).	Collected half of the dataset (50 rows).
Built simplification, summarization, and explanation models and created several versions each for optimization.	Built simplification, summarization, and explanation models and created several versions each for optimization.
Built and ran the baseline model.	Built user interface by using Gradio
Evaluated all model's quantitative and qualitative results.	Ran all models to generate output data and adjusted output data

Permissions:

	Peier Chen	Zirui Xu
Permission to post video	Yes	Yes
Permission to post final report	Yes	Yes
Permission to post source code	Yes	Yes

Link to project code on GitHub repository:

https://github.com/ece1786-2023/SimpliText/tree/main

References

Nishihara D, Kajiwara T, Arase Y. Controllable text simplification with lexical constraint loss[C]//Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop. 2019: 260-266.

Shardlow M. A survey of automated text simplification[J]. International Journal of Advanced Computer Science and Applications, 2014, 4(1): 58-70.