Fine-Tuned LLM based approach to Scientific Text Simplification

Authors: Syed Muhammad Ali, Hammad Sajid, Owais Aijaz, Owais Waheed

Supervisors: Dr Faisal Alvi, Dr Abdul Samad

Habib University

Background and Motivation

- Scientific texts are often hard to understand by the general audience since they use complex and technical language.

- The CLEF 2024 SimpleText Lab aims to enhance accessibility by simplifying scientific texts and producing easier comprehension for a wider audience.

- The problem is to simplify scientific texts to make it easier for individuals outside specialized fields to understand it.

Objective and Goals

- Large Language models have successfully showed great results in text generation, summarization and manipulation.
- Models like GPT-3.5 are publicly available and can be used for tasks like text simplification and elaboration.
- Our goal is to use state-of-the-art language models for simple yet accurate explanations of scientific texts for the general public.

Division of Tasks

- Task1: What is in (or out)? Selecting passages to include in a simplified summary [1].
- Task 2: What is unclear? Difficult concept identification and explanation (definitions, abbreviation deciphering, context, applications,..) [2].
    - Task 2.1: Extract difficult keywords from the selected paragraph.
    - Task 2.2: Provide a brief definition of the extracted keywords.
- Task 3: Rewrite this! Given a query, simplify passages from scientific abstracts [3].

Method

Task 01:

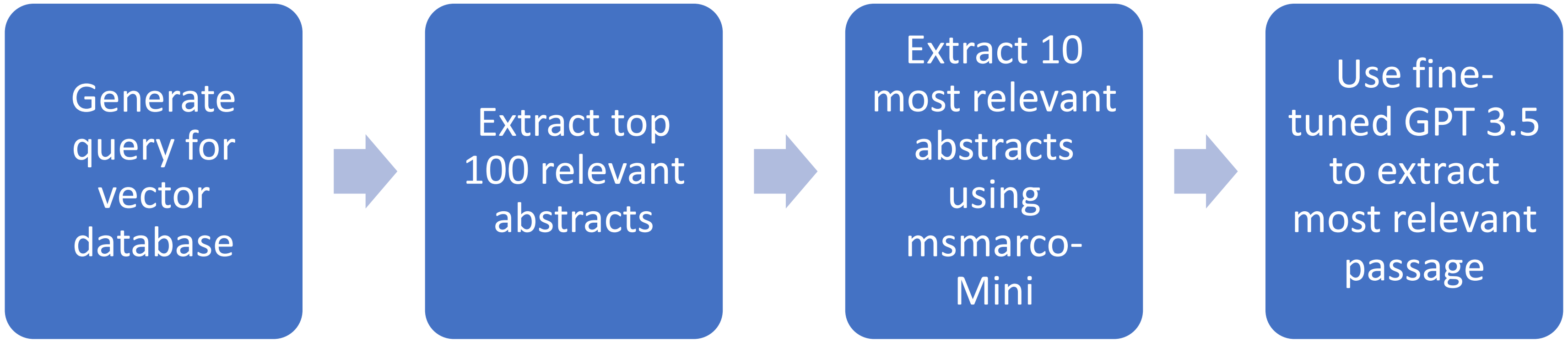


Table 1
Examples of queries generated for vector database based on the length of query text

Sentence/Phrase	Corpus Parameter	Query
Digital Assistant	title	https://guacamole.univ-avignon.fr/stvir_test?corpus=title&phrase=Digitalassistant&length=100
how AI systems, especially virtual assistants, can perpetuate gender stereotypes	abstract	https://guacamole.univ-avignon.fr/stvir_test?corpus=abstract&phrase=howAIsystems,especiallyvirtualassistants,canperpetuategenderstereotypes&length=100

Table 2
Prompts used for the two-step process to select the most relevant passage from the re-ranked abstracts

Step	Prompt
Selecting the abstract	Select the abstract which gives the most relevant definition/explanation for the following term/phrase: *(list of 10 abstracts)*
Extracting the passage	Extract the most relevant part of abstract explaining the given term/phrase in light of the topic *(topic). (abstract)*

Method

Task 02:

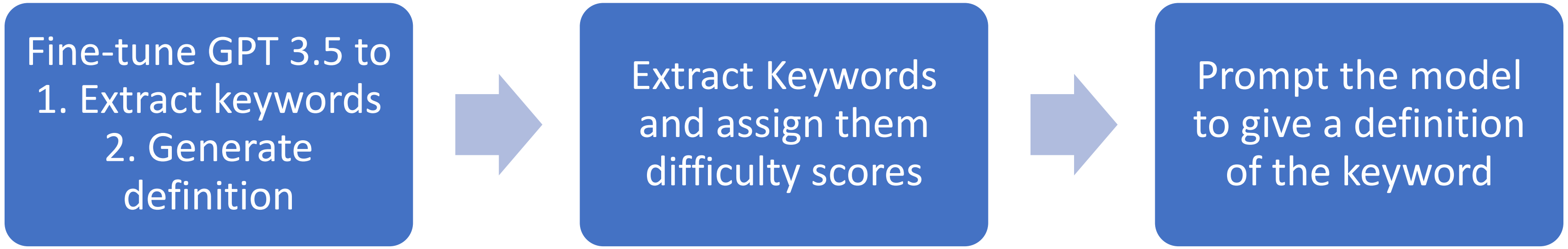
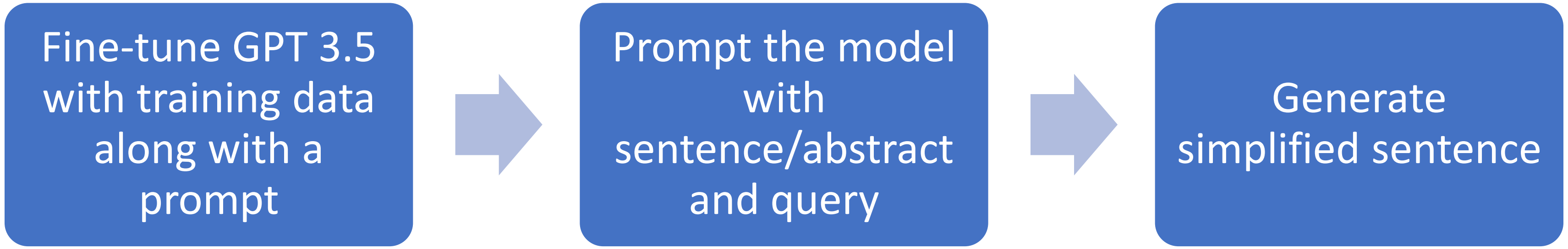


Table 5
Sample prompt to generate definition and explanation of an extracted term

Term	Difficulty	Query
Digital Assistant	m	Generate a definition of the term: "Digital Assistant" having the difficulty score: "m" and provide an explanation.

Method

Task 03:



Other Fine-tuned models:

- BART Sequence to Sequence model
- Pegasus Sequence to Sequence model

Results

Task 01

Table 8
Run scores for Task 01

runid	MRR	Precision 10	Precision 20	NDCG10	NDCG20	Bpref	MAP
Sharingans_Task1 _marco-GPT3	0.6667	0.0667	0.0333	0.1149	0.0797	0.0107	0.0107

- For task 01, our model did not give satisfactory results.
- The most relevant document is best ranked only 66% of the time
- Precision values are low indicating that there is significant irrelevance in the retrieved documents.
- This could be due to manual curation of training data for fine-tuning.
- This could also be due to the inability of GPT3.5 to work on such task.

Results

Task 02

Table 9
Run scores for Task 02

runid	recall			precision	BLEU			
	overall	average	difficult_terms		n1	n2	n3	n4
Sharingans _Task2.2_GPT	0.472222	0.530246	0.544811	0.595361	0.225719	0.103904	0.0300	0.0160

- For task 02, our model gave fairly good results
- The model give comparatively good results for recall and precision but the BLEU score is low.
- Low BLEU score indicates that the word used by our model in the definition were not quite in line with the reference definitions.
- This could also be due to wrong extraction of keywords which would in turn result in complete definition mismatch.

Results

Task 03

Table 10
Run scores for Task 3.1

runid	Count	FKGL	SARI	BLEU	Lexical Complexity	Compression ratio	Levenshtein Similarity
Sharingans_task3.1 _finetuned	578	11.39	38.61	18.18	8.70	0.83	0.77

Table 11
Run scores for Task 3.2

runid	Count	FKGL	SARI	BLEU	Lexical Complexity	Compression ratio	Levenshtein Similarity
Sharingans_task3.2 _finetuned	103	11.53	40.96	18.29	8.80	1.2	0.65

- Our model gave fairly good results for task 03.
- The model has a fair SARI and FKGL score.
- The sentences could have been further simplified but at the cost of losing details.

Conclusion and Future Work

- We found that out of all approaches, GPT 3.5 model gave the best results for task 2 and 3.
- For task 01, our pipeline utilizing GPT 3.5 did not give good results. Further research is needed improve the approach.
- For task 02, we hypothesize that keyword extraction plays an important part. Improvement in keyword extraction is needed for better results.
- For task 03, research is needed to further simplify the text without losing the details.
- Achieve performance of GPT using open-sourced models

Acknowledgements

- Supervisors: Dr Faisal Alvi, Dr Abdul Samad
- Office Of Research (OoR) at Habib University, Karachi, Pakistan for funding this project through the internal research grant IRG-2235.
- SimpleText@CLEF-2024 chairs for their guidance and organization.

Thank You

Train Parameters

Table 3
Experimental setup for GPT-3.5 Turbo for Task 1

Model Name	Examples	Epochs	Batch Size	learning_rate_multiplier
GPT-3.5 Turbo	30	3	1	2

Table 4
Experimental setup for GPT-3.5 Turbo for Task 2

Model Name	Queries	Epochs	Batch Size	learning_rate_multiplier
GPT-3.5 Turbo	501	3	1	2

Table 6
Experimental setup for GPT-3.5 Turbo for Task 3.1

Model Name	Queries	Epochs	Batch Size	learning_rate_multiplier
GPT-3.5 Turbo	958	3	4	2

Table 7
Experimental setup for GPT-3.5 Turbo for Task 3.2

Model Name	Queries	Epochs	Batch Size	learning_rate_multiplier
GPT-3.5 Turbo	175	3	1	2