ENHANCING THE COMPREHENSION: TEXT SIMPLIFICATION APPROACHES AND THE ROLE OF LARGE LANGUAGE MODELS

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Ziyu Yang
May 2024

Examining committee members:

Dr. Slobodan Vucetic, Computer and Information Sciences
Dr. Eduard Dragut, Computer and Information Sciences
Dr. Stephen MacNeil, Computer and Information Sciences
Dr. Wenpeng Yin, External Member, Penn State University

ABSTRACT

Radiology reports are highly technical documents aimed primarily at doctor-doctor communication. There has been an increasing interest in directly sharing radiology reports with patients to improve doctor-patient communication. However, radiology reports are primarily created for communication of imaging findings among medical professionals and are difficult to comprehend by laypeople. Thus, it would be useful if a patient-friendly version of a radiology report could be provided to patients in addition to the original report. Addressing this gap by creating patient-friendly versions of these reports not only enhances doctor-patient communication but also empowers patients by providing them with accessible information about their health conditions. However, a conundrum arises as requiring radiologists to augment traditional reports with patient-friendly summaries could exert negative influence on their cognitive load and productivity. Recent research worked on automatic simplification of health records, employing both lexical and semantic simplification methods, with recent advancements incorporating deep learning techniques. Yet, training these deep learning models for medical text simplification necessitates the collection of costly labeled data.

To address these challenges, we investigate the roles of Large Language Models (LLMs) in simplifying radiology reports to make them more accessible to patients. Firstly, we assess GPT-3's capacity to generate patient-friendly summaries of radiology sentences, specifically focusing on liver conditions. Our findings reveal that with appropriate prompting and fine-tuning, GPT-3 can produce high-quality simplifications, as evidenced by both automated

metrics and manual evaluations by radiologists. Secondly, addressing the challenge of data scarcity in training models for medical text simplification, we introduce a novel data augmentation strategy. This approach leverages the generative capabilities of pre-trained LLMs to create simplified versions of unlabeled radiology sentences and employs paraphrasing techniques on labeled data, significantly enhancing the accuracy of our fine-tuned simplification model beyond baseline methods. Lastly, we explore the effectiveness of advanced prompting mechanisms, such as chain-of-thought and self-correction, in improving the quality of text simplifications. Through a dual-evaluation protocol involving both radiologists and laypeople, we demonstrate the superiority of self-correction prompting in producing simplifications that are both factually accurate and easier for laypeople to understand. Collectively, these studies underscore the potential of LLMs in bridging the gap between complex medical information and patient comprehension, providing valuable insights into the methodologies and evaluation frameworks that can facilitate the development of more accessible health communication tools.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Text simplification is an emerging area of research aimed at making written content more accessible to diverse audiences Chandrasekar et al. (1996). It involves the process of modifying the language of a text to make it easier to understand, while retaining the original information and meaning. This can include reducing the complexity of sentence structures Kandula et al. (2010), using simpler vocabulary Zeng & Tse (2006), and paraphrasing jargon and technical terms Oh et al. (2016). The need for text simplification spans various domains, including education Rets & Rogaten (2021), law Garimella et al. (2022), and medical health Ondov et al. (2022), addressing the challenge of understanding complex texts for individuals with lower literacy levels, non-native speakers, and people with cognitive disabilities. In the context of medical communication, and specifically radiology reports, the relevance of text simplification becomes particularly acute Qenam et al. (2017). These reports are typically written by and for medical professionals, utilizing a specialized vocabulary that can be complicated for the layperson. The disparity between the professional medical language used in these reports and the health literacy of the general population highlights a critical area for intervention. Text simplification in this domain not only seeks to bridge the gap in understanding but also aims to empower patients in their healthcare journey, ensuring that they have access to, and can comprehend, vital information about their health. This overarching goal of enhancing accessibility and comprehension through text simplification naturally leads into the specific challenges and solutions associated with making radiology reports patient-friendly, a topic of increasing importance as these reports become more readily available to patients.

An increasing number of healthcare providers are interested in sharing health records with patients. That is a positive development because research has shown that sharing medical records with patients might improve patient-doctor communication Ross & Lin (2003), increase patient involvement in care Delbanco et al. (2012), and improve outcomes Rosenkrantz & Flagg (2015). However, the health literacy - defined as “the degree to which individuals have the capacity to make appropriate decisions regarding their health” Kutner et al. (2006) - of most patients is often not sufficient to enable an understanding of their health records Lalor et al. (2018). The health literacy gap is especially severe for some types of medical reports, such as radiology reports, which are unstructured documents written by radiologists to communicate imaging findings to another physician or a qualified medical professional Goldberg-Stein & Chernyak (2019). As a result, radiology reports use particularly complex medical jargon and highly specialized descriptions Delbanco et al. (2012) and present a particular challenge for patients Hong et al. (2017). For instance, a recent study Yi et al. (2019) found that the mean readability grade level of MRI reports was above the 12th-grade reading level. Without adequate counseling with an experienced clinician, the severity of the radiology findings may be misinterpreted by the patients. It could lead to unnecessary stress, improper follow-up, and even to increased patient mortality Sudore et al. (2006).

Given these facts, there is a growing consensus on the need for patient-friendly radiology reporting that communicates results clearly and is understandable by a diverse patient population. Several studies have highlighted the benefits of sharing medical records with patients, including improved patient-doctor communication, increased patient involvement in care, and improved outcomes Ross & Lin (2003); Delbanco et al. (2012); Rosenkrantz & Flagg (2015). Unfortunately, asking radiologists to write patient-friendly reports has several obstacles. First, simplifying terminology and writing patient-friendly sentences would inevitably result in loss of information and reduced quality of communication between radiologists and ordering physicians. Second, asking a radiologist to supplement a traditional report with a patient-friendly summary would negatively impact their cognitive load and productivity. Therefore, recent research considered medical informatics and machine learning approaches for making the radiology reports patient-friendly.

These challenges motivated recent research on the automatic simplification of health records. The proposed approaches include both lexical simplification that paraphrases text Chen et al. (2018); Biran et al. (2011); Weng et al. (2018) and semantic simplification that seeks to simplify grammatically complex text Shardlow (2014); Leroy et al. (2016) which recently included deep learning approaches Lewis et al. (2019); Zhang et al. (2020). However, training deep learning models for medical text simplification requires the collection of costly labeled data. An enticing alternative that has garnered much recent interest Jeblick et al. (2022); Lyu et al. (2023) is to generate patient-friendly simplifications with large language models (LLMs) and ask radiologists to check the generated simplifications before releasing them.

There are several open challenges to the generation of patient-friendly radiology reports. The first is that it needs to be clarified what constitutes a good simplification. The existing research has varying views of the trade-offs between factuality, completeness, simplicity, and brevity Jiang et al. (2020); Cripwell et al. (2022). A proper combination of these measures may depend on individual user preferences. As a result, it would be very challenging to create a widely acceptable parallel corpus for radiology report simplification. Moreover, very different simplifications could be evaluated as equally successful (longer and more detailed versus shorter with only critical information). Thus, even if the parallel corpus was created and used to train and test an LLM, automatic evaluation using measures that rely on sequence similarity Lin (2004); Xu et al. (2016); Zhang et al. (2019) might be misleading. Instead, until there is more clarity about what constitutes a reasonable radiology simplification, we think humans should perform the evaluation.

CHAPTER 2

RELATED WORKS

2.1 Text Simplification

In text simplification, the output text is a linguistically simplified version of the input text Adduru et al. (2018). Previous work on simplification includes lexical and semantic simplification Alva-Manchego et al. (2020).

Lexical simplification by lexical substitution refers to replacing complex words or phrases with simpler synonyms Oh et al. (2016); Zeng & Tse (2006); Chen et al. (2018); Biran et al. (2011); Weng et al. (2018) and has found some practical success Cook et al. (2017). In the health domain, lexical text simplification often relies on medical dictionaries (UMLS Bodenreider (2004) , MeSH Lipscomb (2000), etc.). Lexical simplification approaches also include rule-based methods Chen et al. (2018); Biran et al. (2011) and deep learning Weng et al. (2018, 2019).

Semantic simplifications seek to simplify grammatically complex text by splitting long sentences into shorter ones, changing passive voice to active, resolving ambiguities and anaphora Shardlow (2014), splitting complex noun phrases Leroy et al. (2016), or reducing morphological negations Mukherjee et al. (2017). Recently, transformer encoder-decoder based pre-trained seq-to-seq models Lewis et al. (2019); Zhang et al. (2020) were proved to be robust in solving text simplification problems. However, fine-tuning pre-trained models require large quantities of labeled data, which are costly and difficult to obtain in the health domain. This paper adopts this more novel emphasis. Plain language summarization Guo et al. (2021); Devaraj et al. (2021) is an alternative term that reminds that the main objective is to enhance laypeople's understanding of expert-written texts.

Benchmark data is important in a research field. Most research on text simplification has concentrated on making sentences easier to understand, relying on sources like the Wikipedia-Simple Wikipedia aligned dataset Zhu et al. (2010); Woodsend & Lapata (2011) and the Newsela simplification collection Xu et al. (2015). Studies on simplifying entire documents have been less common, likely due to a shortage of resources Sun et al. (2021); Alva-Manchego et al. (2019). Besides, in plain language summarization, CELLS Guo et al. (2022) is a parallel corpus featuring scientific abstracts and their plain language summaries, authored by the abstract creators or domain experts. The medical field greatly benefits from the process of making complex information easier to understand. Medical texts are extensive and filled with specialized language, highlighting the importance of making these texts understandable to those who are not experts in the field Apfel & Tsouros (2013). Various studies have explored how to simplify medical documents by adjusting the words and structure to make them more accessible Damay et al. (2006); Llanos et al. (2016). The recent availability of the Cochrane dataset, which pairs detailed medical research findings with summaries understandable by the general public, offers a valuable resource for this purpose Devaraj et al. (2021). The PLABA dataset Attal et al. (2023) consists of expertly revised biomedical abstracts. These abstracts have been simplified to enhance comprehension of health-related information. However, non of these data are huge scaled. Therefore, previous researches have explored different methods for text simplification in low-resource domains. To address data scarcity recent studies include unsupervised methods Surya et al. (2018); Sakakini et al. (2020); Enayati et al. (2021) and reinforcement learning Laban et al. (2021).

2.2 Radiology Text Simplification

Radiology text simplification using LLMs has recently drawn significant attention Ondov et al. (2022). A recent work used fine-tuned BART Lewis et al. (2019) to simplify 140 liver-related radiology sentences Yang et al. (2023). In Jeblick et al. (2022); Lyu et al. (2023), researchers explored the use of prompt learning with the GPT family Brown et al. (2020), including ChatGPT-3.5 and ChatGPT-4, to simplify radiology reports. Jeblick et al. (2022) focused on three artificial reports, while Lyu et al. (2023) considered over 100 reports. However, they did not provide fully coherent evaluation of the simplifications.

There are two related NLP problems that have been popular in radiology. Radiology report generation refers to automated creation of reports from X-ray or other radiographic images Liu et al. (2023a). This is an image-to-text task with a different set of objectives from text simplification. Radiology report summarization refers to condensing the detailed ”Findings” section of radiology reports into a succinct ”Impression” section Zhang et al. (2018). This involves creating a shorter version of the report that retains all critical information without a necessity to make it clearer to laypeople Chaves et al. (2022); Liang et al. (2022).

2.3 Evaluation of Text Simplifications

Assessing output of LLMs is integral to text simplification Van den Bercken et al. (2019); Cripwell et al. (2022) and related natural language generation tasks. In text simplification, automatic metrics such as ROUGE Lin (2004) and BERTScore Zhang et al. (2019) have been popular, which compare similarity between gold standard and generated sentences. SARI Xu et al. (2016) compares simplified text both with reference simplifications and the original sentences, thus assessing the operation of adding, deleting, and keeping words. Unfortunately, these metrics often correlate poorly with human evaluation of text simplification Alva-Manchego et al. (2021); Liu et al. (2023b); Guo et al. (2023). For readability assessment, the Flesch-Kincaid Grade Level (FKGL) is a widely recognized metric that estimates the text's reading difficulty. More recently, Guo et al. (2023) proposed to assess readability by using difference in normalized perplexity scores from in-domain and out-of-domain language models.

Using human evaluators has been increasingly popular in text simplification, despite the significant associated costs. Researchers typically evaluate fluency, adequacy, factuality, and simplicity of the simplified texts Jiang et al. (2020); Cripwell et al. (2022). Very often, these measures are vaguely defined and subject to interpretation. Recently, factuality was formalized in terms of addition, substitution, and deletion of information Devaraj et al. (2022).

In radiology report simplifications, there is no clear standard for evaluation. Jeblick et al. (2022) enlisted 15 radiologists to assess simplified reports for factual correctness, completeness, and potential harm. Lyu et al. (2023) invited two radiologists to evaluate the simplified reports based on metrics such as information loss, misinterpretation, and an overall score. Interestingly, these studies did not evaluate the simplicity of the text. Lu et al. (2023) focused on simplicity, fluency, and factual accuracy. The study recruited students to assess factualness and simplicity and two medical experts to examine the factual consistency. However, it is unclear if the participating students possessed any medical expertise to represent laypeople and if they were qualified to assess the factualness.

CHAPTER 3

EVALUATING SUITABILITY OF GPT-3 MODELS FOR RADIOLOGY REPORT SIMPLIFICATION

3.1 Introduction

The appropriate strategy for machine-assisted patient-friendly reporting has been debated Lourenco & Baird (2020); Kalia (2020). Previous attempts for improving patient comprehension have been lexical and syntactic simplification. Lexical simplification refers to replacing complex medical terms with simpler ones Oh et al. (2016); Zeng & Tse (2006) and has found some success Cook et al. (2017). Natural Language Processing (NLP) approaches to aid lexical simplification include rule-based methods Chen et al. (2018); Biran et al. (2011) and deep learning Weng et al. (2018, 2019). A few studies considered syntactic simplifications that seek to simplify grammatically complex text by splitting long sentences into shorter ones, changing passive voice to active, resolving ambiguities and anaphora Shardlow (2014), splitting of complex noun phrases Leroy et al. (2016), or reducing morphological negations Mukherjee et al. (2017). However, lexical and syntactic simplifications cannot guarantee patient comprehension because most patients lack the knowledge to understand the context and meaning of sentences in medical notes despite understanding every individual term Qenam et al. (2017).

In this work, we consider an alternative approach that treats patient-friendly reporting as text simplification. In this case, original text is provided as an input to simplification software, which produces text simplification as output, where it is acceptable to simplify terminology and grammar and remove technical details. Text simplification belongs to a larger class of sequence to sequence (seq2seq) natural language processing (NLP) problems. Recent rapid advances in deep learning resulted in very powerful language models that can rival humans in text simplification on general-purpose text Van et al. (2020). Probably the most powerful language model at the moment is Generative Pre-trained Transformer (GPT) GPT-3 Brown et al. (2020) model. GPT-3 is the third generation of GPT model, which is a transformer based decoder with 175 billion parameters. It is capable of producing human-like text and was trained on large text data from internet with hundreds of billions of words. While the source code for GPT-3 has not been released, OpenAI provides an API interface for practitioners.

The main mechanism language models such as GPT-3 are used for text simplification is prompting Shin et al. (2020); Schick & Schütze (2020). In prompting, the language model is provided an input text and is asked to continue generating words given the input. An impressive and surprising property of large-scale language models is that the text in the input could be used to effectively explain the task or demonstrate the task through a few examples, resembling the way a task would be explained to a human.

In this work, we designed and performed a set of experiments to carefully evaluate ability of GPT-3 to simplify radiology report sentences. We first created a parallel corpus of radiology sentences about liver condition and their simplifications produced by a radiologist and a layman. Then, we explored three strategies for simplification. The first two include 1) prompting: providing task description before the sentence and 2) in-context learning: providing labeled examples in the contexts for few-shot learning. The third strategy consists of fine-tuning GPT-3 (enabled through the API) with training examples from our corpus and using the fine-tuned GPT-3 to provide simplifications.

3.2 Data Set

We are not aware of a parallel corpus of original radiology sentences and their simplifications that would be appropriate for our study. Thus, the first step of this study was to create such corpus. Instead of taking a random sample of radiology sentences, we focused on simplification of sentences talking about liver condition. This allowed us to collect a large number of simplifications about liver and better compare different prompting strategies for GPT-3 based simplification. We obtained liver-related sentences from radiology reports in the publicly available source MIMIC-IIIJohnson et al. (2016). There are 522 thousand radiology reports in MIMIC-III. We extracted the sentences from radiology reports describing Computed Tomography (CT) of abdomen. In particular, we selected reports containing word 'abdomen' in the report category name.

This section will introduce how we create our radiology sentences simplification data. In particular, we focus on liver-related radiology reports in the public source MIMIC III Johnson et al. (2016), which is the most popular public database in the medical domain, contains more than 522 thousand radiology reports. We focus on liver-related sentence simplifications, and we do several pre-processing steps before we ask radiologists to write simplifications. MIMIC radiology reports have clear category names, which indicate the specific body part of the reports. Here, we select all reports about *abdomen* by matching the word 'abdomen' in category names. Next, we parsed the 'Finding' and 'Impression' sections from the selected reports. Then, we applied the sentence tokenizer from the python package NLTK to extract sentences from those sections. We selected sentences contain word *'liver'* and saved 1,000 of them at random.

We manually explored the 1,000 selected sentences and discovered that some of them described multiple organs, that some of them mention liver as a proximal organ, that context of some of them was unclear without looking at the preceding sentences, and that some of them simply state that the liver looks normal. 306 such sentences were removed. We also observed that some sentences use negation to explain that there is nothing wrong with the liver. We decided that an appropriate simplification of such sentences would be *'The liver looks normal.'*. In order to prevent creating a data set with too many sentences of this type, we selected only half of them in our data set. The resulting data set had 356 liver sentences.

We asked a radiologist to provide a simplification for each of the selected sentence. The simplification guideline proposed using simplification *'The liver looks normal.'* instead of explaining what liver condition was determined to be normal and why. A total of 79 sentences were simplified in this way. The guideline allowed removing technical details that might be confusing to patients. The guideline specified for all medical terms to be stated in simple terms that patients should be familiar with. If possible, the grammar should be simple and the sentences should be short.

3.3 Problem Definition

We define our data set as $\mathbf{D} = \{(\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2), ..., (\mathbf{X}_N, \mathbf{Y}_N)\}$, where $\mathbf{X}_i$ is the ith liver-related sentence in $\mathbf{D}$ and $\mathbf{Y}_i$ is its simplification written by a radiologist. Each sentence is a set of tokens, $\mathbf{X}_i = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, ..., \mathbf{w}_{m_i}\}$ and $\mathbf{Y}_i = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, ..., \mathbf{s}_{n_i}\}$, where m_i and n_i are the length of sentence $\mathbf{X}_i$ and $\mathbf{Y}_i$, respectively. The radiology sentence simplification is a seq2seq problem. The objective is to generate simplification $\mathbf{Z}$ given sentence $\mathbf{X}$ that matches the ground truth simplification $\mathbf{Y}$ as closely as possible. The success of the prediction can be measured by objective measures and by human evaluation.

3.4 Experiments

We implement GPT-3 by fine-tuning or few-shot learning with the labeled dataset $\mathbf{D}$ and evaluate the suitability of this state-of-the-art language model. OpenAI publishes four different GPT-3 models with different scales. GPT-ada, GPT-babbage, GPT-curie, and GPT-davinci, the number of parameters start from 340 million to 175 billion. Ada and

Babbage are powerful in classification problems while Curie and Davinci are both applicable to be used for sovling seq2seq tasks. All GPT models can be applied for few-shot learning and fine-tuning via the official API codes. Besides, OpenAI recently released their new approach about fine-tuning a pre-trained GPT-3 model, which becomes more specific and controllable for solving downstream NLP tasks. All GPT-3 models but GPT-davinci can be fine-tuned via API codes. Therefore, we will first introduce two typical few-shot learning processes with Curie and Davinci. Then, we describe the impressive fine-tuning process with the GPT-Curie model.

Prompting Other than the prompt by examples, we also try another intuitive design, prompt by description. As introduced in the previous section, we can add a line of descriptions as the demonstration of the task and let language model understand our expectation easier. We create a prompt named **Patient**. The description is a demonstration sentence: *Simplify the sentence so that patients can understand:*. Moreover, the liver sentence is appended after the indicator. This prompt does not require examples from our training data. Thus, it is a special condition of few-shot learning, zero-shot learning. Similarly, we create another prompt (**Grader**) by changing the description into: *My second grader student ask me to simplify the following sentence:*, which is inspired by the example in the GPT-3 demo.

Table 3.1: Different prompt designs for GPT-3 models. All these prompts are zero-shot prompts.

Name	Prompt
Patient	My patient asks me the meaning of: X
Grader	My second grader student ask me to simplify the following sentence: X
Plain	Please simplify the sentence: X
...	...

In-context learning Few-shot learning is the most popular and basic usage of GPT-3 models because GPT-3 models are relatively large, which is not easy for fine-tuning. To use GPT models via few-shot learning, we need to provide a prompt with some labeled examples, a paragraph of designed free texts, and let models continue writing after the prompt. Therefore, to let GPT models generate simplifications with similar formats from radiologists, we select top K most related pairs $\{(\mathbf{X}_j, \mathbf{Y}_j)|1 \leqslant j \leqslant K\}$ from labeled training data and construct them as the prompt examples for GPT models.

Refer to Shin et al. (2020), selecting related labeled examples in the in-context learning makes the generations better. Thus, we try to find the most related labeled examples in $\mathbf{D}$ for every query sentence. Also, in ablation study, to support our claim, we try most unrelated and random examples. To find the most proper examples for the query sentence, we apply the Clinical BERTscore Zhang et al. (2019) to compute the sequence similarities between our query sentence and all labeled radiology sentences. To be specify, Clinical BERTscore is a variant of BERTscore Zhang et al. (2019), a robust metric for evaluating sequences similarity by calculating tokens' BERT vector cosine similarities and transforming them into one value with TF-IDF and max-pooling techniques, to calculate the similarity score between raw radiology sentences and the query sentence. Specifically, we change the BERT model to the pre-trained Clinical BERT Alsentzer et al. (2019), which is a BERT model Devlin et al. (2018) that was pre-trained on the MIMIC III data. Then, we select top K highest scored labeled pairs as examples of the prompt.

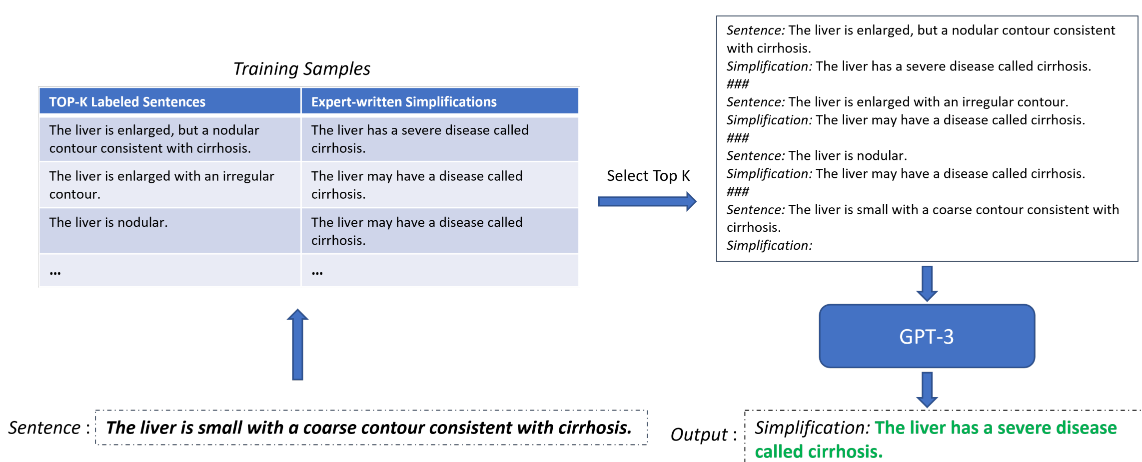

FIGURE 3.1: Architecture of the In-context learning on the query example. Given an original sentence, it generates a prompt and ask GPT-3 to generate the output.

As we know, the prompt is the core of few-shot learning. Therefore, different designs of prompts and choices of few-shot examples may impact the final performances, which has been proved by Shin et al. (2020). We will try different examples and prompts in few-shot learning in our ablation studies.

After selecting few-shot examples, we need to find a way to wrap examples and our query together as the input of GPT models. In language models, text indicators are beneficial to help the model understand the contextual environment. For example, in English to French machine translation, instead of simply passing the original English sentence and the target French sentence side by side to models, we can put some word indicators at the front of sentences: ***English:*** *original sentence* ***; French:*** *target sentence*. **English:** and **French:** help model immediately realize the task we want to solve is a translation problem, and the **;** separate the original and target sentences properly.

Inspired by the prompt design in machine translation, we designed a similar prompt template for GPT models. We choose the template: ***Sentence:*** **X***;* ***Simplification:*** **Y** to store one example pair. A triple pond sign separates all few-shot examples: *###*. Then, we append the unlabeled sentence to the end of the prompt. The final input format is shown at the top of Fig 3.1.

Fine-tuning Besides few-shot learning, we can also fine-tune a pre-trained GPT-3 model released by OpenAI recently. Due to the massive size of models, we can only run the fine-tuning process via API. Following the official documentation, we can fine-tune all GPT-3 models except the most powerful one, GPT-davinci. The fine-tuning process is accomplished by formatting and passing the training data inside. To fine-tune a GPT-3 model, we must create a prompt for each training pair. However, the prompt is much shorter than we used in few-shot learning because we do not need to involve labeled examples in the prompt. Instead, we need to consistently indicate the original sentence, simplification, and some separators or stopper symbols in the prompt. Our paper uses the double break-line sign and trip pond sign as the separator. Many other works prove that the fine-tuned GPT models are more potent than few-shot learning with same-scaled pre-trained GPT models when the training data is enough. The experiment section will evaluate if fine-tuned GPT models have competitive performances in the radiology area.

3.5 Evaluation Methods

Numeric Metrics To better evaluate the output from models, we use multiple automated numeric metrics to evaluate texts simplification performances. As we introduced, we apply clinical BERTscore to evaluate the similarity between the raw and simplified sentence in contextual level. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) Lin (2004) is a set of metrics used for seq2seq tasks. ROUGE-1,2, L calculates the overlap of unigram (word), bigrams, and longest common subsequence between the original and simplification sentences. Another metric is BLEU (bilingual evaluation understudy) Papineni et al. (2002), which calculates the precision between simplification and a set of original sentences. Recently, SARI Xu et al. (2016) is a popular and standard edit-based metric for text simplification task. It calculates the steps of operations (add, delete, keep) between the simplified sentence, the raw sentence, and the human-labeled ground-truth simplifications. Finally, FKGI (Flesch Kincaid Grade Level) is a widely used readability formula that assesses the approximate reading grade level of a text. Lower grade value means the sentence is easier to be understand. Normally, we aim for grade 8 to ensure the content can be understood by 80% of Americans.

Human Evaluation Simply applying automatic evaluation metrics is insufficient to comprehensively distinguish the qualities and capabilities of generated simplifications from different methods. Therefore, we add human evaluation to obtain more versatile feedback and comments from medical experts and normal users. Inspired by the pipeline described by Nisioi et al. (2017) in the normal language domain, we follow and modify this design based on the version from Van den Bercken et al. (2019).

We ask human experts to evaluate our generated simplifications in two aspects: The first one is the **Correctness**. In the medical area, the correctness of simplification is the most important metric. Correctly simplifying the original radiology sentences can reduce hallucinations and confusion for patient customers. We ask radiologists to score the correctness of simplifications on a 1-5 Likert scale. Where 1 is not correct at all and 5 is exactly correct. The second aspect is the **Quality**. We measure how many edits should be done by experts before we send the simplification to patient. It monitors both grammar and readability at the same time. We use the same 1-5 Likert scale. Where 1 means the sentence is awful and 5 means the sentence is exactly how you would say it. Somehow the score in the middle is measuring how many issues should be edited. Higher score means less required edits. Some other works evaluated the simpleness of simplifications as wellVan den Bercken et al. (2019). However, the FKGL score has already provided a reliable measurement for simpleness. Thus, we decide not to let human users evaluate it again in the process.

3.6 Results

We created 356 labeled pairs from MIMIC III abdomen reports, as we introduced previously. To evaluate how GPT models perform in the radiology sentence simplification problem, we randomly select 85 pairs as our testing samples. The remaining 271 pairs are for selecting the few-shot prompt examples and fine-tuning process.

In the process of running few-shot learning with GPT models, some hyper-parameters need to be defined. One is called temperature, a 0-1 ranged numeric value, which controls the randomness of the next word generation from GPT models. The closer to 1 means, the more random texts you may obtain. Usually, high-temperature value is used for generating sequences or continuous writing a story given some starting texts. In contrast, the low-temperature value is suitable for classification or entity extraction tasks. Following the GPT-3 texts summarization examples, we choose the temperature value as 0.6 in our experiment. Another hyper-parameter is the maximum number of tokens to be generated. We choose 30 based on the maximum number of words in our expert-written simplifications.

Table 3.2: Few-shot performance comparison for different GPT models on Top3 prompt. The upper part is the performance by running GPT models once. While, the lower part is the performance by selecting the best answer from five runs for each sentence.

Model(K = 1)	ROUGE 1/2/L	BLEU	SARI	Clinical BERTscore	FKGL↓
GPT-Ada	44.02/27.70/42.11	0.1855	0.4652	0.7748	7.765
GPT-Babbage	50.67/36.79/48.95	0.2803	0.5072	0.8165	7.796
GPT-Curie	57.28/43.17/55.44	0.3314	0.5476	0.8412	7.238
GPT-Davinci	54.11/39.03/52.80	0.2610	0.5296	0.8317	6.554
Model(K=5)	ROUGE 1/2/L	BLEU	SARI	Clinical BERTscore	FKGL↓
GPT-Ada	43.93/29.65/42.08	0.2030	0.4801	0.7709	6.215
GPT-Babbage	52.88/39.10/50.28	0.3066	0.5300	0.8239	7.124
GPT-Curie	**58.85/44.66/56.93**	**0.3562**	**0.5590**	**0.8536**	6.927
GPT-Davinci	54.77/40.43/52.92	0.2853	0.5316	0.8292	6.215

We keep all the hyper-parameter default in the API codes in the fine-tuning process. All the models are fine-tuned with four epochs without early stopping. In the evaluation phase, we directly apply the fine-tuned GPT models to the testing data and obtain the predicted simplifications without any other processing.

Models Effect Now, we evaluate how different GPT models perform on the same prompt. We implement all models with the same prompt and evaluate the generated simplification in the lower part of Table 3.2. We can conclude that the larger scaled pre-trained model we used, the better results can be accomplished, except the most powerful one: GPT-Davinci. Due to the characteristic of few-shot learning, all results are full filled randomness. It might be a coincidence that GPT-Curie can over performed GPT-Davinci in some cases.

As described in the Checker section, we define an algorithm to select the best answer from five runs properly. To demonstrate the importance of our checker, we compare the performances between the first answer and the best answer with the same model and prompt. The comparison can be found in Table 3.2. Clearly, our picking method is always better than selecting the first run under all automated metrics.

Prompts Effect We introduced a BERTscore example selection method and a translation-like template in few-shot learning. To explore the impact of different prompts and examples, we do ablation studies by passing different prompts into the same model. Regarding the example selection, we introduced the default 3-shot translation-like prompt in previous sections (**Top3**). Hence, we first increase 3-shot examples to 5-shot. For the 5-shot, we try both BERTscore related examples(**Top5**) and randomly selected examples (**Rd5**). We implement GPT-Curie model with different prompts. Performances are shown in Table 3.3. Apparently, GPT models cannot write out similar simplifications compared with our expert-written ones without providing examples in the prompt. Comparing the performances on **Top3** and **Top5** prompts, we can conclude that more examples may lead a better performance. Similarly, most related examples are always better than the random selection under the condition of the same number of examples.

Fine-tuning vs Prompting To fairly compare the fine-tuned GPT models and pre-trained GPT models, we select the top-2 best performances from different prompts as the prompting result with the pre-trained GPT model. On the other hand, we apply the fine-tuned GPT models to our testing data and obtain the results for comparison. Table 3.4 shows the comparison between the fine-tuned results and the prompting results, which are obtained from the **Top3** and **Top5** prompts. From the table, we can prove the huge impact between the process with or without learning. Fine-tuned GPT models can easily learn the writing style of our labeled data, while prompting can only guess the pattern from a few examples in the prompts. Therefore, fine-tuned GPTs perform better when we use the measurements that are calculated based on the expert-written simplifications.

We ask radiologists to evaluate the generated results from four typical models: Davinci model few-shot learning with the **Patient** prompt, Davinci and Curie models few-shot learing with the **Top5** prompt, and the fine-tuned Curie model. Table 3.5 shows the mean values of our human evaluation scores from 75 selected testing examples. All these four models present valuable simplifications, except the Davinci-Patient is slightly under performed compared with other models because of the difficulty from zero-shot learning. It is good in correctness but has low quality because it always repeated from the original sentence instead

Table 3.3: Performance comparison for different prompts with GPT-Curie on all five prompts and GPT-Davinci on three selected prompts

Model	ROUGE 1/2/L	BLEU	SARI	Clinical BERTscore	FKGL↓
Curie-Grader	20.79/05.39/19.28	0	0.3540	0.6567	9.487
Curie-Patient	23.92/08.52/21.63	0.0096	0.3777	0.6955	8.908
Curie-Rd5	30.51/11.02/28.37	0.0435	0.4238	0.7443	7.389
Curie-Top3	58.85/44.66/56.93	0.3562	0.5590	0.8536	6.927
Curie-Top5	**59.53/45.73/57.52**	**0.3631**	**0.5654**	**0.8589**	**6.734**
Davinci-Patient	26.82/08.36/23.68	0.0090	0.3892	0.7105	6.89
Davinci-Top3	54.77/40.43/52.92	0.2853	0.5316	0.8292	6.215
Davinci-Top5	57.34/42.21/55.12	0.3056	0.5451	0.8425	6.15

Table 3.4: Performance comparison between the Few-shot learning (FS) and Fine-tuning (FT) with GPT-Ada, GPT-Babbage, and GPT-Curie. The FS results are obtained from the best prompt: Top5

Model	ROUGE 1/2/L	BLEU	SARI	Clinical BERTscore	FKGL↓
GPT-Ada(Top3)	43.93/29.65/42.08	0.2030	0.4801	0.7709	6.215
GPT-Ada(Top5)	49.90/35.42/48.32	0.2531	0.5093	0.8015	6.208
GPT-Ada-FT	70.66/61.67/69.31	0.4983	0.6075	0.8624	5.387
GPT-Babbage(Top3)	52.88/39.10/50.28	0.3066	0.5300	0.8239	7.124
GPT-Babbage(Top5)	56.66/44.67/55.24	0.3635	0.5484	0.8439	7.352
GPT-Babbage-FT	69.61/58.68/68.00	0.4316	0.6015	0.8530	5.452
GPT-Curie(Top3)	58.85/44.66/56.93	0.3562	0.5590	0.8536	6.927
GPT-Curie(Top5)	59.53/45.73/57.52	0.3631	0.5654	0.8589	6.734
GPT-Curie-FT	71.00/61.29/69.17	0.4903	0.6131	0.8687	5.083
Davinci-Top3	54.77/40.43/52.92	0.2853	0.5316	0.8292	6.215
Davinci-Top5	57.34/42.21/55.12	0.3056	0.5451	0.8425	6.15

Table 3.5: Human evaluation results and FKGL score on 75 selected testing data. It involves two human evaluation rubrics.

Model	Correctness↑	Quality↑	FKGL↓
Davinci-Patient	4.29	3.92	6.67
Davinci-Top5	4.46	4.42	6.05
Curie-Top5	4.38	4.41	5.92
Curie-FT	4.33	4.46	5.22

Table 3.6: Example generated simplifications visualization. Bold texts are original liver-related sentences.

Model	Simplification
There is also a wedge-shaped region of low attenuation in the anterior right lobe of the liver.	
Davinci-Patient	There is a wedge-shaped region of low attenuation in the anterior right lobe of the liver.
Davinci-Top5	There is a triangular area of shadowing in the front right side of the liver.
Curie-Top5	There are one or more triangular areas of shadowing in the liver.
Curie-FT	The liver showed one or more triangular areas of shadowing along one or more of its edges.
A tiny amount of intrahepatic air is also identified within the left medial segment of the liver.	
Davinci-Patient	A tiny amount of air is identified within the left medial segment of the liver.
Davinci-Top5	There is some air inside the liver.
Curie-Top5	Some air is found inside a liver blood vessel in the left side of the abdomen.
Curie-FT	The liver contains some air that may be inside a blood vessel.
Focal fatty infiltration in the right lobe of the liver.	
Davinci-Patient	Fatty infiltration in the right lobe of the liver.
Davinci-Top5	There is a small area of the liver where there is an increased amount of fat.
Curie-Top5	There is increased amount of fat in one specific area of the liver.
Curie-FT	There is increased amount of fat in the liver.

of simplifying it. We also conclude that Davinci model is more powerful than Curie model with same few-shot prompt in our testing phase, though Curie has better performances in numeric metrics. Not only from the human evalutation scores, we also gain these insights by visualizing the example simplifications in Table 3.6.

Curie-FT is good at mimicking expert-written simplifications during the fine-tuning process. That's why it leads in the numeric evaluation and has the highest quality score in human evaluation. It is the best model when we have similar-distributed testing samples. However, its generalization ability will be lower than few-shot learning models when we test the out-of-distribution sentences.

3.7 Conclusion

In this paper, we collect and annotate a novel medical-related sentence simplification data. Moreover, we design and implement all GPT models on our data via prompting and fine-tuning process. We evaluated the impact from different prompts and models, respectively. Furthermore, We demonstrated that the method we applied in finding the best answer from multiple GPT outputs is better than the baseline, treating the first run as the answer. From our human evaluation phase and the visualization examples, we successfully evaluate the suitability of pre-trained language models in radiology sentence simplification. In future work, we will consider paragraph-level radiology texts to avoid the incomplete information in the sentence level we face. Also, we plan to invent a more robust method to find the best answer from different outputs.

CHAPTER 4

DATA AUGMENTATION FOR RADIOLOGY REPORT SIMPLIFICATION

4.1 Introduction

In this work, we consider an alternative approach that treats patient-friendly reporting as text simplification. In this case, original text is provided as an input to simplification software, which produces text simplification as output, where it is acceptable to simplify terminology and grammar and remove technical details.

To alleviate the data scarcity issue in simplifying health reports, particularly radiology reports, this work proposes a novel approach for data augmentation. It augments manually-created labeled data with simplifications generated by a large pre-trained language model such as GPT-3 Brown et al. (2020). To improve the quality of data augmentation, the approach develops a separate deep learning model that evaluates the quality of generated simplifications. Furthermore, the approach also provides data augmentation through paraphrasing the originally labeled radiology sentences.

The proposed data augmentation approach is experimentally evaluated on a unique corpus of manually generated labeled data for radiology report simplification. The evaluation includes both automatic measures and human evaluation.

Our research claims are: 1) Our augmentation methods enable training of a more accurate model than baselines in solving low-resource radiology sentence simplification problems. 2) We address the challenge of selecting qualified augmentations for radiology sentence simplification. 3) We create unique real data containing expert-annotated simplifications for radiology reports' sentences regarding liver conditions.

4.2 Related Works

In addition to the research works in text simplification, data augmentation is a general method that automatically generates labeled data to enhance manually labeled data Liu et al. (2020). One approach is to use paraphrasing to create different variants of the original or simplified sentences Wei & Zou (2019). Another approach is to use pre-trained language models to generate labeled data Bayer et al. (2021). LAMBADA Anaby-Tavor et al. (2020) augments data for text classification tasks by encoding labels in the input. Similarly, PromptDA Wang et al. (2022) use language models to augment data for NLU tasks. Back-translation Edunov et al. (2018) is used to generate different variants of the input text.

There are several public benchmark data sets that are related to our paper. There are paragraph level medical text simplifications Devaraj et al. (2021) focusing on medical paper abstracts. There is a corpus parsed aligned sentences from Wikipedia and Simple English Wikipedia [1] Pattisapu et al. (2020); Van den Bercken et al. (2019) that has been a popular text simplification benchmark. However, none of these data sets have properties similar to the radiology text simplification task.

4.3 Problem Definition

$\mathbf{D}_{\text{Lab}} = \{(\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2), ..., (\mathbf{X}_n, \mathbf{Y}_n)\}$, where $\mathbf{X}_i$ is the ith original document, $\mathbf{Y}_i$ is its simplification provided by a human expert, and n is the number of labeled documents.

[1] `simple.wikipedia.org`

Let us also assume we are given an unlabeled corpus of documents $\mathbf{D}_{\text{Unl}} = \{\mathbf{X}_1, \mathbf{X}_2..., \mathbf{X}_m\}$, where m is the number of unlabeled documents. The objective of data augmentation is to automatically create a synthetic set $\mathbf{D}_{\text{Syn}} = \{(\mathbf{X}_1^*, \mathbf{Y}_1^*), (\mathbf{X}_2^*, \mathbf{Y}_2^*), ..., (\mathbf{X}_K^*, \mathbf{Y}_K^*)\}$, where $\mathbf{X}_i^*$ is one of the original documents from $\mathbf{D}_{\text{Lab}}$ or $\mathbf{D}_{\text{Unl}}$ or their derivative, and $\mathbf{Y}_i^*$ is its corresponding simplification from $\mathbf{D}_{\text{Lab}}$, its derivative, or an automatically generated simplification. $\mathbf{D}_{\text{Syn}}$ is appended to $\mathbf{D}_{\text{Lab}}$ and the resulting set $\mathbf{D}_{\text{Aug}} = \mathbf{D}_{\text{Lab}} \cup \mathbf{D}_{\text{Syn}}$ is called the augmented training corpus. The assumption is that a seq2seq model for text simplification trained on the augmented corpus will have higher simplification quality than the one trained on the original training corpus. This paper focuses on the radiology report simplification problem where $\mathbf{X}_i$ is a sentence and $\mathbf{Y}_i$ is its simplification. As a result, our augmentation approach explained in the next two sections is specifically tailored for this application.

4.4 Methodology

4.4.1 GPT-based Semantic Augmentation

We propose two types of augmentation. The first, referred to as semantic augmentation, relies on using the large-scale language model GPT-3 Brown et al. (2020) to generate simplifications automatically. The second, referred to as lexical augmentation, relies on modifying original documents from the labeled corpus. The first augmentation type is described in this section, while the second type is described in the next.

Generating simplifications with GPT-3

It has been demonstrated that large-scale language models such as GPT-3 are capable of doing a wide variety of NLP tasks through prompting Brown et al. (2020). Our preliminary experiments showed that the two largest GPT-3 models, Curie and Davinci, are surprisingly good at generating simplifications of radiology report sentences, despite never being trained explicitly for that task. Table 4.1 shows several representative examples of simplifications

Table 4.1: Good, Not Simple Enough, and Incorrect examples of radiology liver sentences (ORI), expert-written simplifications (SIM), and generated simplifications from GPT-Curie (GPT).

ORI:	The liver demonstrates diffusely low attenuation, consistent with fatty infiltration
SIM:	There is increased amount of fat in the liver.
GPT:	There is a lot of fat in the liver. (Good)
ORI:	The liver is stable in appearance with again note made of a hypodense lesion in the medial left lobe adjacent to the gallbladder fossa
SIM:	The liver may have one or more tumors or masses.
GPT:	The liver has a small lesion on one side, close to the gallbladder. (Good)
ORI:	The liver shows diffuse low attenuation consistent with fatty infiltration
SIM:	There is increased amount of fat in the liver.
GPT:	The liver has diffuse low attenuation, indicative of fatty infiltration. (Not Simple Enough)
ORI:	The liver is unremarkable without intra- or extra- hepatic biliary dilatation
SIM:	The liver appears normal.
GPT:	The liver is not enlarged or tender without intra- or extra-hepatic biliary dilatation. (Not Simple Enough)
ORI:	Complex hepatic abscess predominantly occupying segments VI and VII of the liver is overall little changed in size compared to date, measuring approximately 7.9 (AP) x 8.7 (TRV) x 11.4 (CC) cm
SIM:	There is infection and pus seen within the liver.
GPT:	This means that the abscess is smaller than it was before, by about 3 inches on all sides. (Incorrect)
ORI:	The liver demonstrates prominent intrahepatic biliary ducts
SIM:	The liver ducts that carry digestive enzymes may be blocked.
GPT:	The liver has many bile ducts. (Incorrect)

by the Curie model prompted by *My patient asks me to simplify this radiology sentence "original sentence":*. Each example shown contains an original radiology sentence, its simplification by GPT-3, and its simplification provided by collaboration between a radiologist and a layperson. The first two examples show that GPT-3 can provide factually correct and easy to understand simplifications comparable to the manually created simplifications. Next two examples show that GPT-3 might provide factually correct text that is not sufficiently simple. The final two examples show that GPT-3 might provide factually incorrect simplifications.

Our semantic augmentation approach uses GPT-3 to generate simplifications of unlabeled documents from $\mathbf{D}_{\text{Unl}}$ and add them to the augmented corpus $\mathbf{D}_{\text{Aug}}$. As noted in previous research Liu et al. (2021) the choice of prompting can have a significant impact on the quality of the generated text and accuracy on a particular task.

Our prompting approach relies on the in-context learning that has been used with success with GPT-3 models. Instead of relying on costly fine-tuning of a language model, it pastes a few labeled examples into the prompt and asks the language model to generate label of an unlabeled example. In our specific application, we select K labeled examples $(\mathbf{X}, \mathbf{Y})$ from $\mathbf{D}_{\text{Lab}}$ and insert each of them into template '***Sentence:*** $< \mathbf{X} >$*;* ***Simplification:*** $< \mathbf{Y} >$'. A triple pound sign, *###*, is used to separate templates for the K labeled examples. The prompt ends with '***Sentence:*** $< \mathbf{X} >$*;* ***Simplification:***', where $\mathbf{X}$ is an unlabeled document from $\mathbf{D}_{\text{Unl}}$. GPT-3 model is expected to write a simplification by mimicking the style of the labeled examples from the prompt.

As noted in previous work Brown et al. (2020) the success of prompting that uses in-context learning depends on the particular choice of K examples. Therefore, we select most related sentence simplification pairs from the training set $\mathbf{D}_{\text{Lab}}$ given any unlabeled document from $\mathbf{D}_{\text{Unl}}$. In detail, we use BERTScore Zhang et al. (2019), which leverages the pre-trained contextual embeddings from BERT Devlin et al. (2018) and matches words in unlabeled and labeled radiology sentences by cosine similarity. Thus, each prompt consists of K most related examples rated by BERTScore for an unlabeled sentence that is appended to the end. Moreover, we evaluate more example selection scenarios in our ablation study.

BERT-Checker

Language models such as GPT-3 provide token probabilities as their output. When generating text, one option is to use brute force and generate the most likely token. However, in the context of text simplification, the most likely tokens are not guaranteed to produce the best simplification. An alternative is to generate tokens by selecting among the most likely choices, which the temperature hyperparameter in GPT-3 can control. In our approach, we invoke a GPT-3 model N times for each prompt using a temperature higher than zero, which results in N different simplifications. Then, we automatically select the best one of the N generated simplifications and add it to the augmented corpus.

As seen in Table 4.1, some of the generated simplifications are good while others are not. Separating good from inadequate simplifications is a non-trivial challenge. Related work on automatic evaluation of the generated text includes GPT-3-ENS Chintagunta et al. (2021), which measures the complexity of terms in simplifications, and GPT3Mix Yoo et al. (2021), which treats the likelihood scores of generated labels as confidence scores. However, we found that the existing approaches are inappropriate for our application. Thus, we developed a novel approach called BERT-Checker.

BERT-Checker is a fine-tuned BERT model Devlin et al. (2018) to a task similar to entailment. In particular, we convert our labeled corpus into training data matching the format of the entailment task. We add label 1 to each example from $\mathbf{D}_{\text{Lab}}$ to create positive examples in new training data set, $\mathbf{D'}_{\text{Lab}} = \{[(\mathbf{X}_i, \mathbf{Y}_i), 1]\}$. To create negative examples in $\mathbf{D'}_{\text{Lab}}$, we use four different strategies as outlined next:

- **Precision:** To ensure that simplification is closely related to the original text, we corrupt the original text $\mathbf{X}$ by replacing the medical terms with randomly selected medical terms, and generate negative example from labeled example $(\mathbf{X}, \mathbf{Y})$ as $[(corrupt(\mathbf{X}), \mathbf{Y}), 0]$.

- **Simplicity:** To penalize simplifications that are too similar to the original sentence, we create negative examples by using the original text as simplification, $[(\mathbf{X}, \mathbf{X}), 0]$.

- **Correctness:** To penalize incorrect simplifications, we randomly select two labeled examples $(\mathbf{X}_1, \mathbf{Y}_1)$ and $(\mathbf{X}_2, \mathbf{Y}_2)$ and create a negative example by mixing the original and simplified text, $[(\mathbf{X}_1, \mathbf{Y}_2), 0]$.

- **Robustness:** For labeled example $(\mathbf{X}, \mathbf{Y})$ we replace the simplification with an empty string or a sentence generated by a GPT-3 given the prompt 'Generate a radiology report sentence about liver' and high temperature of 0.8 to create negative example $[(\mathbf{X}, GPT()), 0]$.

Thus, for each positive example, we generate four negative examples. As a result, we can obtain a negative dataset $\mathbf{D'}_{\text{Neg}}$. We fine-tune Clinical BERT Alsentzer et al. (2019) on the text entailment task using the generated data set.

4.4.2 Dictionary-based Lexical Augmentation

We propose lexical augmentation to supplement semantic augmentation described in the previous section. Lexical simplification refers to replacing complex terms in original documents $\mathbf{X}$ with their synonyms, which might also be complex. In the related work on text simplification of general-purpose text, EDA approach Wei & Zou (2019) paraphrases original documents by replacing randomly selected words or phrases with their synonyms in WordNet Miller (1995). We modify EDA by replacing only specialized medical terms.

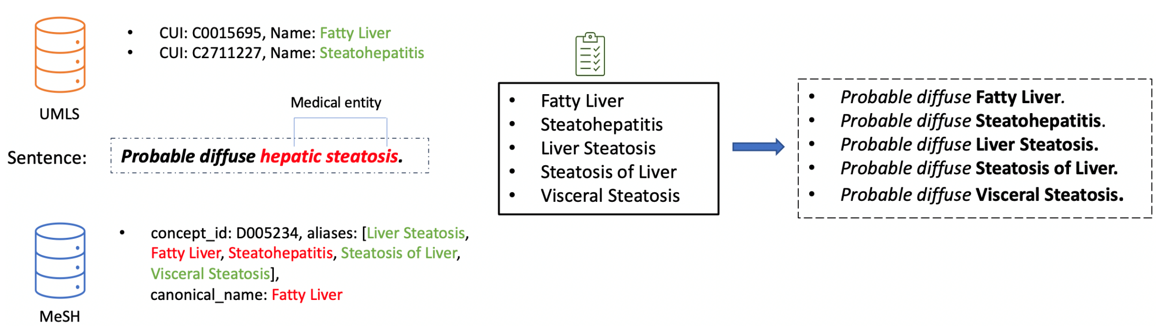

FIGURE 4.1: Workflow for lexical augmentation. It shows the linked synonyms of the entity "hepatic steatosis" from UMLS/MeSH and five synthetic sentences.

Inspired by Pattisapu et al. (2020); Hasan et al. (2016), we use medical dictionaries Medical Subject Headings (MeSH) Lipscomb (2000) and Unified Medical Language System (UMLS) Bodenreider (2004) to find the synonyms. We use pre-trained named entity recognition model Honnibal & Montani (2017) to extract medical terms from the original documents in labeled corpus $\mathbf{D}_{\text{Lab}}$. The medical terms are linked to Concept Unique Identifier (CUI) in UMLS and the concept_id in MeSH. Each medical code in UMLS and MeSH is mapped to a list of synonyms. We iteratively select a synonym to replace the medical term from the original document.

We illustrate the lexical simplification process in Fig 4.1, where *hepatic steatosis* in the sentence '*Probable diffuse hepatic steatosis*' is recognized as a medical term and replaced with its synonyms. In particular, CUI codes 'C0015695' and 'C2711227' are found to match *hepatic steatosis*, where the canonical names are *Fatty Liver* and *Steatohepatitis*. Similarly, 'D005234' from MeSH also provides several synonyms. This process identifies five synonyms used to create five different versions of the original document.

Once the synonyms for a medical term in original document $\mathbf{X}$ of labeled example $(\mathbf{X}, \mathbf{Y})$ are identified, we paraphrase the original document as $lexical(\mathbf{X})$ and generate an augmented example $(lexical(\mathbf{X}), \mathbf{Y})$. The new example is added to the augmented corpus $\mathbf{D}_{\text{Aug}}$.

4.5 Experiments

4.5.1 Data

To the best of our knowledge, there is no readily available corpus for simplifying radiology sentences. To experimentally evaluate our data augmentation approach, we created a new corpus for this purpose. In particular, we collected 540 sentences from radiology reports describing the liver condition and manually created their simplifications: 170 sentences were obtained from CT-Abdomen radiology reports from a university hospital (UH), and the remaining 370 were extracted from CT-Abdomen radiology reports from publicly available MIMIC-III Johnson et al. (2016) data. All sentences were de-identified with Health Insurance Portability and Accountability Act (HIPAA) standards in order to facilitate public accesses and human annotations.

We asked a radiologist to provide a simplification for each selected sentence. A layperson joined the radiologist to provide feedback about the generated simplifications. If the layperson thought the simplification was too complicated, this was communicated to the radiologist, who proceeded to improve the simplification. The process was repeated until the layperson could understand all the simplification and could correctly guess the severity of the described conditions.

During this sequence simplification process, the radiologist and the layperson agreed that it is sufficient to use simplification '*The liver looks normal*' for sentences explaining that nothing concerning was observed about the liver. 39% of the the university hospital sentences and 21% of the MIMIC-III sentences were simplified as '*The liver looks normal*'. For simplification of sentences that described concerning findings was to ignore technical details that might be confusing to patients. Any relevant medical terms were stated in simple terms familiar to laypeople. If possible, grammar was kept simple, and the sentences were kept short. Table 4.1 shows several examples of the original sentences (ORI) and their manual simplifications (SIM).

For our experiments, we randomly selected 100 sentences and their simplifications for training and the remaining 70 for testing for both the university hospital and MIMIC-III labeled data. Thus, we had 200 labeled examples for training denoted as $\mathbf{D}_{\text{Lab}}$, and 140 for testing. We used the remaining 200 MIMIC-III sentences as the unlabeled corpus $\mathbf{D}_{\text{Unl}}$ and used their simplifications to better evaluate the data augmentation approaches.

The corpus is available to the research community to support further research on medical text simplification.[2]

4.5.2 Data Augmentation

To implement the proposed semantic augmentation approach, we used GPT-3 Curie model (6.7B parameters) with the few-shot learning prompt described in Section 4.1 with $K = 5$ to automatically generate simplifications for each unlabeled sentence in $\mathbf{D}_{\text{Unl}}$. We

[2] https://github.com/Ziyu-Yang/Radiology-Text-Simplification-Liver

used the API provided by OpenAI [3]. We generated $N = 5$ simplifications for each liver sentence with temperature $= 0.5$, which was selected to provide a good balance between factual correctness and diversity.

We trained BERT-Checker to select the best among the $N = 5$ generated simplifications for each liver sentence. BERT-checker was fine-tuned using 80% of the training data as positives and four copies of negatives for each positive, as explained in Section 4.2. BERT-Checker was a fine-tuned BERT base model (110M parameters) consisting of 12 transformer encoder layers. A fully connected linear layer was added to BERT on its [CLS] output to score the simplification quality. The binary cross-entropy loss was used. 20% of the training data was used for validation and early stopping. We fine-tuned for up to 20 epochs with the patience for early stopping of 3, batch size 16, and learning rate 1e-4. All experiments were implemented with a single GTX 1080Ti.

The accuracy of trained BERT-Checker on validation data was 0.924. Its precision (the fraction of true positives among positive predictions) was 0.899 and its recall (the fraction of positives that were predicted correctly) was 0.958. We consider it to be high enough accuracy for BERT-Checker to be used to determine the quality of simplifications produced by GPT-3.

In the lexical augmentation, we annotated the recognized entities in the liver sentences from the labeled corpus with Type Unique Identifier (TUI) [4]. TUI is the code to represent hierarchical semantic types of all medical concepts in UMLS and MeSH. Specifically, we only paraphrased terms that belong to "*T023 — Body Part, Organ, or Organ Component*" or "*T033 — Finding*" groups. Because many medical concepts have only one synonym, many sentences mentioned only a single body part other than the liver, and a single finding, we finally obtained 242 unique lexical augmentations from $\mathbf{D}_{\text{Unl}}$. In order to control the effect of augmentation size, we randomly selected 200 of them for further experiments.

[3] `https://openai.com/api/`

[4] https://lhncbc.nlm.nih.gov/semanticnetwork/index.html

4.5.3 BART Model

BART Lewis et al. (2019) is a pre-trained model that uses a seq2seq architecture with a bidirectional encoder and a left-to-right decoder. It achieves state-of-the-art performance on many seq2seq benchmarks. We fine-tuned a BART base model (406M parameters) on different mixes of 450 labeled and augmented data to create different radiology simplification models. The fine-tuning was implemented using PyTorch-lightning[5]. 20% of the training data was used for validation and early stopping. We used the cross entropy loss. We used the same training setting as for BERT-Checker.

4.5.4 Baselines

We first introduce two model baselines that do not use augmentations. Then we introduce two baseline augmentation methods that are appropriate to our task.

Model Baselines

The first baseline is BART base model fine-tuned with the labeled data (**BART-base**). As the second baseline, we used simplifications by the same implementation of GPT-Curie model that is used to augment the labeled data. Specifically, we selected the most related $K = 5$ sentences from the labeled set to a test sentence as the few-shot prompt, generated $N = 5$ simplifications and used BERT-Checker to select the best one. We name this baseline **GPT-FS**.

Augmentation Baselines

We implemented and evaluated two widely used baseline data augmentation methods: 1) Easy Data Augmentation (**EDA**) Wei & Zou (2019),a rule-based augmentation that includes synonym replacement, random insertion, random swap, and random deletion. We reproduced this baseline with its source code [6]. 2) Back translation (**BT**), that uses a pre-

[5] `https://www.pytorchlightning.ai/`

[6] `https://github.com/jasonwei20/eda_nlp`

trained machine translation model to translate sentences into another language and then translate them back to English. The back-translated English sentences are fused with the corresponding simplifications to provide augmented data. Following previous work Brown et al. (2020), we used GPT-3 Curie to back translate the original sentences to French and back to English. French was selected because it provided a good balance between factual correctness and diversity of generated back-translations.

We generated 200 augmented examples for each baseline approach.

4.6 Evaluation Methods

4.6.1 Automated Evaluation

We used multiple automated metrics to evaluate text simplification accuracy. **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) Lin (2004) is a set of metrics used for seq2seq tasks. It calculates the overlapping of unigrams, bigrams, and the longest common subsequences between the expert-provided and machine-generated simplifications. Similarly, **BLEU** (bilingual evaluation understudy) Papineni et al. (2002) also evaluates overlap of n-grams between the simplifications. Unlike ROUGE and BLEU, **BERTScore** Zhang et al. (2019) computes a contextual similarity score between tokens in the simplifications. **SARI** Xu et al. (2016) is a gold standard edit-based metric for text simplification evaluation. Unlike other metrics, it compares the machine-generated simplification with respect to both the original sentence and the human-provided simplification. To evaluate simplicity, we used **FKGL** (Flesch Kincaid Grade Level) Kincaid et al. (1975), which is a widely used readability formula that assesses the approximate reading grade level of a text. The lower score indicates simpler texts.

4.6.2 Human Evaluation

Applying automatic evaluation metrics is insufficient to compare quality of simplifications by different methods. Therefore, we also used human evaluation. We asked a medical doctor (family physician) that was distinct from the radiologist who provided the simplifications to evaluate the machine-generated simplifications. We asked the evaluator to use 1-5 Likert scale to evaluate the following four aspects of each simplification, the first three being consistent with.

Factuality refers to medical correctness of the simplification. Score one means that the simplification is factually incorrect and five that it is correct. Scores between one and five mean that some information is imprecise, missing, or hallucinated. Lower scores mean there are more serious factual errors. **Fluency** measures the quality of grammar and readability, regardless of factual correctness. If a simplification is both easy to read and grammatically correct it gets a score of five. This measure is consistent with the fluency measure explained in Nisioi et al. (2017). **Simplicity** evaluates whether the evaluator thought the laypeople would be able to understand the simplification, regardless of factual correctness. Score of five means that the evaluator thought that any patient would be able to completely understand the simplification.

During the initial stages of human evaluation of factuality and simplicity, we observed that the evaluator occasionally preferred machine-generated simplifications to the radiologist-provided ones. That is why we introduced **Consistency**, which measures how closely the simplification matches the radiologist-provided simplification. Score of five means that the simplification is almost identical to the radiologist-provided simplification. We note that Consistency is related to SARI automatic measure Xu et al. (2016).

4.7 Results

4.7.1 Quantitative Results

We fine-tuned BART model including augmented data from baseline methods (EDA, BT), and our lexical and semantic augmented data (LEX, SEM). BART-base and GPT-FS were created according the description in previous sections. First two rows of Table 4.2 refer to fine-tuned BART and few-shot prompted GPT-3 Curie using only the radiologist-provided

Table 4.2: Comparison of augmentation methods. # Aug is the number of augmented examples.

	# Aug	ROUGE 1/2/L	BLEU	SARI	BERTScore	FKGL↓
		Baseline Models				
GPT-FS	0	56.00/42.20/54.64	0.2363	0.5455	0.9457	5.392
BART-base	0	59.81/50.34/58.87	0.4240	0.5324	0.9411	5.560
		Augmentation Methods				
EDA	200	60.90/51.90/60.06	0.4461	0.5460	0.9429	5.315
BT	200	63.47/53.50/62.33	0.4504	0.5740	0.9470	5.133
HUMAN	200	**71.06/62.89/70.20**	**0.5322**	**0.6047**	**0.9566**	**4.870**
LEX	200	68.58/60.84/68.24	**0.5391**	0.5769	**0.9559**	5.353
SEM	200	66.11/56.55/64.76	0.4709	0.5875	0.9510	5.629
AUG-SUB	200	67.92/58.81/67.04	0.5020	0.5960	0.9524	5.314
AUG	400	**69.03/60.37/68.51**	0.5036	**0.6029**	0.9550	**5.021**

labeled data. The remaining rows refer to inclusion of augmented data to BART tuning. Rows EDA and BT refer to the baseline augmentation methods. Row HUMAN refers to the augmentation provided by the radiologist, and serves to establish the upper bound on accuracy improvement due to augmentation. LEX and SEM rows represent our lexical and semantic augmentation methods. AUG-SUB and AUG use 200 and 400 combined semantic and lexical augmentations, respectively. Aug column shows the number of augmented examples.

We observe that our proposed augmentations are superior to baselines LEX and SEM on almost all metrics. SEM is better than LEX on SARI measure. AUG is better than LEX and SEM on ROGUE, SARI and FKGL. AUG is the best overall augmentation method coming very close to the HUMAN upper bound, after noting that SARI and FKGL are the most useful measures for evaluation of simplicity. We note that GPT-FS has lower overall scores than any of the BART models.

4.7.2 Human Evaluation Results

For Table 4.3, we asked a medical doctor to evaluate 60 randomly selected simplifications from the test data (30 from each source). We evaluated the most relevant four models

Table 4.3: Human evaluation results (Factuality, Fluency, Simplicity, and Consistency) on 60 selected testing data.

Method	Factual	Fluency	Simp	Cons
BART-base	3.38	4.85	4.67	3.18
BT	3.22	4.88	4.58	3.13
GPT-FS	4.18	5.00	4.55	3.91
AUG	4.22	4.95	4.62	4.10

from Table 4.2: BART-base, BT, GPT-FS,and AUG. The results show that all methods have comparable Simplicity and Fluency. AUG and GPT-FS have better Factuality and Consistency than BART-base and BT. AUG is slightly better than GPT-FS on those two important measures, indicating that fine-tuning BART with augmentation produced by few-shot prompted GPT-3 Curie is better than directly using few-shot prompted GPT-3 Curie for simplification.

Table 4.4: Comparison between different versions of semantic augmentations. # Aug is the number of augmented examples. ROUGE refers to ROUGE-L.

Method	# Aug	ROUGE	BLEU	SARI
First-run	200	60.21	0.4053	0.5590
Similarity	200	55.31	0.3547	0.5387
Five-runs	781	55.77	0.3755	0.5391
SEM	200	**64.76**	**0.4709**	**0.5875**

4.7.3 Ablation Study

We first evaluated the ability of BERT-Checker to recognize high-quality simplifications. We compared the version we implemented in our experiments (SEM row in Table 4) with three different variants: 'First-run' always selects the first generated simplification, 'Similarity' selects the best simplification based on BERTScore, 'Five-runs' uses all simplifications generated by GPT-3 Curie as augmentations. After removing duplicates, there are 781 augmentations produced by 'Five-runs'. Table 4.4 shows all three variants are inferior to SEM, showing that any of the ablations would significantly deteriorate the results. The results confirm that the quality of augmentations is critical for success of data augmentation approaches.

Next, we evaluated the importance of GPT-3 prompting. As noted in previous research Liu et al. (2020), the choice of prompting can significantly impact the quality of the generated text. Thus, we designed an ablation study to compare different prompting approaches for data augmentation.

Table 4.5: Comparison of different prompting on data augmentation. Top 4 rows indicate baseline methods of building prompts. SEM indicates our proposed semantic augmentation method.

Prompts	ROUGE	BLEU	SARI
BART-grader	46.62	0.2862	0.4917
BART-patient	55.81	0.3516	0.5255
BART-top1	58.65	0.3955	0.5511
BART-rd5	53.94	0.3170	0.5360
SEM	**64.76**	**0.4709**	**0.5875**

In our prompt design that has the following form: ***Sentence:*** $< \mathbf{X} >$*;* ***Simplification:*** $< \mathbf{Y} >$', we included $K = 5$ most related labeled examples to the original test sentence in the prompt. We first explored whether the number of few-shot examples matters. We repeated the data augmentation process with $K = 1$ (BART-top1 in the table). Table 4.5 shows that $K = 5$ resulted in better performance than $K = 1$. Next, we evaluated whether the way we select examples matters. Instead of $K = 5$ closest labeled examples, we selected $K = 5$ random labeled examples (BART-random in the table). From Table 4.5, we can see that random labeled examples resulted in lower accuracy.

We also explored prompting that does not rely on few-shot learning. One design was explained in section 4.4.1, '*My patient asks me to simplify this radiology sentence* $< \mathbf{X} >$', we refer to as BART-patient in the table. Similarly, inspired by a GPT-3 prompt for the summarization task, we used prompt: *My second grader student asks me to simplify the following sentence:* $< \mathbf{X} >$, we refer to as BART-grader in the table. These two prompts are the so-called 'zero-shot' prompts. As shown in Table 4.5, the 'grader' and 'patient' prompts result in inferior accuracy compared to the few-shot prompting. Few-shot prompting usually generates more comprehensive outputs than zero-shot prompting.

4.8 Conclusion

This paper proposes two novel augmentation methods to enhance the limited labeled data for the radiology sentence simplification problem. Our evaluation using automatic measures and human evaluation shows that data augmentation can substantially improve the quality of simplification models. The ablation results show that the proposed innovations in automatic creation of simplifications for data augmentation are very effective.

4.9 Limitations

The main limitation of our study is that we only considered simplification of radiology sentences. In future work, it will be important to expand the approach to simplify whole paragraphs, because very often radiologists use multiple sentences to discuss a single observation. Simplifying single sentences can thus be suboptimal because important context from the previous and subsequent sentences might be lost. The second limitation of the study is that our corpus only included sentences related to liver. It will be important in the future work to evaluate the proposed approach on a wider variety of radiology sentences. The third limitation is that we obtained simplifications from a single radiologist. It will be important for future study to include simplifications from multiple radiologists to ensure generalizability of the proposed approach. The fourth limitation is that we used a single medical doctor to evaluate the quality of the simplifications. It would be important in future studies to ask multiple medical doctors to evaluate the quality, which would allow estimating the inter-rater variability. The fifth limitation is that we did not use laypeople to evaluate the quality of simplification. This would require some innovation in the human evaluation process because laypeople are not able to evaluate factual correctness and because it would be important to understand how simplifications improve the overall understanding of the radiology reports. The final limitation is a relatively small size of the labeled data set created for this study. Obtaining high-quality simplifications is very costly because it requires collaboration between radiologists and laypeople.

CHAPTER 5

TWO-PRONGED HUMAN EVALUATION OF CHATGPT SELF-CORRECTION IN RADIOLOGY REPORT SIMPLIFICATION

5.1 Introduction

There is no broadly accepted protocol for human evaluation of simplified expert text Van den Bercken et al. (2019); Devaraj et al. (2022); Lu et al. (2023), including what questions to ask and who should answer them. In this work, we propose a novel evaluation protocol following two ideas. First, we observe that laypeople should not be asked factuality and completeness questions due to the lack of expert knowledge and that radiologists should not be asked about simplicity due to the curse of knowledge bias. Thus, our protocol employs laypeople and radiologists with slightly different questions. Second, we observe that a good simplification is the one that increases understanding compared to the original text, but also that there can be a dangerous mismatch between perceived and actual understanding. Thus, laypeople are asked both about their perception and their actual increase in understanding when an expert text is supplemented by its simplification.

Modern LLMs can solve various NLP tasks with high success through prompting and without necessitating fine-tuning Brown et al. (2020). The quality of output is very sensitive to prompting. While prompting is sometimes considered an art form, there are a few strategies that work more often than not. One is Chain-of-thought (CoT) Wei et al. (2022). Another is self-correction Chen et al. (2023); Madaan et al. (2023). Huang et al. (2022) self-improves an LLM through iterative fine-tuning. Bai et al. (2022) leverages AI-generated feedback through reinforcement learning. Li et al. (2023) allows LLMs to self-improve their generations without training. They instantiate multiple LLMs models as different agents and let them collaborate towards better generation.

Another contribution of this paper is in evaluating the capabilities of the state-of-the-art LLMs without constructing a large parallel text corpus. Arguably, the best publicly available LLM at the moment is ChatGPT OpenAI (2023), and recent papers Jeblick et al. (2022); Lyu et al. (2023) indicate that both its 3.5 and 4 versions can provide high-quality radiology report simplifications only through prompting. In this paper, we provide an in-depth evaluation of chain-of-thought (CoT) prompting Wei et al. (2022) and self-correction Madaan et al. (2023). In the CoT approach, LLMs are prompted to justify an answer before providing the answer. In the self-correction approach, LLMs are prompted to critique their original response and asked to consider the critique to give an improved response. Both methods have been shown to work well in several applications Fu et al. (2023); Chen et al. (2023). To our knowledge, they have yet to be evaluated on radiology report simplification.

We designed experiments to answer the following research questions: ($\mathcal{Q}_1$) Is the proposed human evaluation protocol insightful? ($\mathcal{Q}_2$) Are CoT and self-correction helpful in the simplification of radiology reports? ($\mathcal{Q}_3$) What is the relationship between perceived and actual understanding of radiology reports? ($\mathcal{Q}_4$) What kinds of simplifications are preferred by experts and laypeople? The answers should be informative for future research towards high-quality simplifications of expert texts.

5.2 Evaluation Protocol

As described in the previous section, prior text simplification research used human evaluation, but did not clarify the roles of experts and laypeople in evaluation. In the following, we will propose an evaluation protocol that defines those roles.

5.2.1 Factuality

Factuality refers to correctly maintaining the original information. Motivated by Devaraj et al. (2022) and Jeblick et al. (2022), we measure three aspects of factuality.

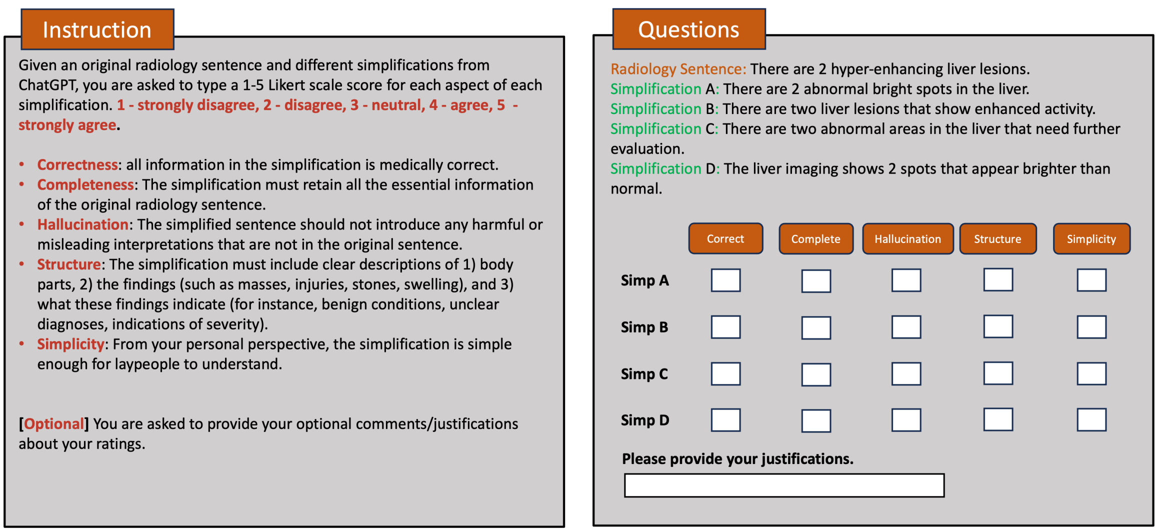

FIGURE 5.1: Expert evaluation of radiology report simplification. (**left panel**) lists instructions, (**left panel**) is a survey form with text boxes for ratings and justification.

- **Correctness**: (Factualness/Substitution) Evaluates whether the simplification correctly interprets the information in the original sentence.

- **Completeness**: (Adequacy/Meaning preservation/Deletion) Evaluates if there is any significant information loss in the simplification compared to the original text. Simplifications should retain all critical information from the original text, but it might be permissible to ignore less important information.

- **Hallucination**: (Addition) Evaluates if simplifications contain wrong statements or hallucinate new information that may misguide laypeople.

We introduce a new measure that is related to Completeness.

- **Structure**: Refers to a desire that simplifications follow a certain structure. Specifically, a good radiology simplification should mention: *body parts, findings, and consequences*. Body parts specify the anatomies and organs referred to in the radiology sentence (such as kidneys). Findings refer to the key observations in the radiology sentence (such as injuries or masses). Consequences refer to what findings indicate, which might not be explicitly stated in the original sentence, such as severity, certainty, and follow up.

Only experts can adequately evaluate factuality and structure. Figure 5.1 shows the exact survey design for expert evaluation of simplifications we used in our experiments. Note that we also ask the radiologists to evaluate the simplicity of generated simplifications for our analysis.

5.2.2 Simplicity

In prior work, simplicity mostly refers to readability, which measures text fluency and complexity of terms and grammar correctness. However, LLMs typically generate very fluent text, so evaluating that aspect is not very informative. Instead, it is more relevant to measure how well laypeople comprehend the text.

Clarity: Instead of asking evaluators to provide a single score for the simplicity Jiang et al. (2020), we evaluate their understanding by devising a set of questions to measure the usefulness of simplifications. The critical objective of radiology report simplification is to improve the clarity about the severity of the described conditions. There are two important dimensions of clarity: how well people understand the text and how well they

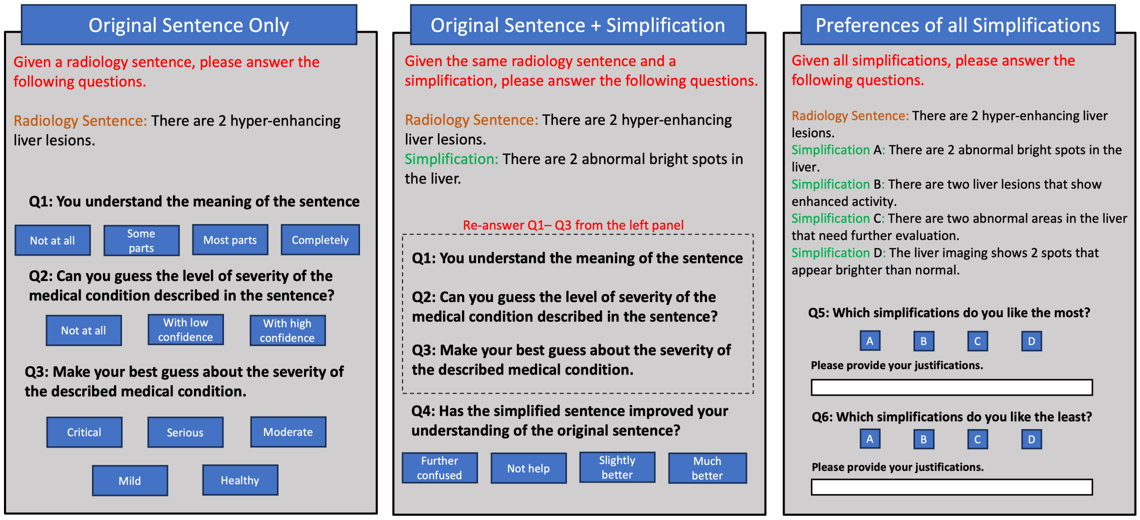

FIGURE 5.2: Layperson evaluation of radiology report simplifications. (a) (**left panel**) evaluates whether laypeople understand the original sentence. (b) (**middle panel**) evaluates whether simplification improves understanding. (c) (**right panel**) evaluates the preferences given a set of candidate simplifications and asks for justification.

believe they understand. Different combinations of those two dimensions can have different consequences for patients. For example, being confident while misinterpreting the text might lead to being too concerned or relaxed. Uncertainty is a clear indication that simplification was not adequate. Our survey is sequenced as in Figure 5.2. We ask laypeople if they think they understand the original text (4 levels). Then, we ask them specifically if they think they understand the severity of the described condition (3 levels). This is followed by asking them to guess the severity, according to 5 severity levels defined as follow.

- CRITICAL (5): Describes a medical condition that poses a threat to a person's life. A critical condition requires urgent care and close monitoring.
- SERIOUS (4): Describes a condition that requires medical attention but is not immediately life-threatening. Treatment may involve hospitalization, medication, or other interventions.
- MODERATE (3): Describes a condition that is not severe but may require medical attention and treatment. The condition may cause discomfort or affect a person's

ability to carry out normal activities.

- MILD (2): Describes a condition that is not serious. The condition may cause minor discomfort or inconvenience but is unlikely to have a significant impact on a person's overall health.

- HEALTHY (1): Findings that are considered normal or benign with no significant abnormalities.

This allows us to compare with the actual severity provided by a radiologist. We repeat those questions by suplementing the original sentence with simplification. Finally, we also ask them about their subjective opinion about the helpfulness of the simplification.

We considered other ways to measure how well laypeople understand the text, such as quizzing them about the body parts and the meaning of the findings. We decided against it because it would be cumbersome to consistently convert the responses into numbers given a wide variety of radiology sentences. Also, asking this question would compound health literacy and simplicity. For example, even if a patient cannot fully understand the medical meaning, it could still be essential to hear that a condition impacting some part of their abdomen is not critical but requires a follow-up.

Preferences: Inspired by the design for evaluating text summarization Goyal et al. (2022), we also ask evaluators to choose the most and the least preferred simplifications among multiple choices. In addition, layperson evaluators are encouraged to provide justifications for their selections, as shown in the right panel of Figure 5.2. This free-text response can be used in qualitative analysis of laypeople's simplification preferences.

5.3 Prompting and Self-Correction

5.3.1 Prompting ChatGPT

Our preliminary results showed that ChatGPT can provide good simplifications of radiology sentences. Since we did not have a sufficiently large corpus of parallel text for radiology report simplification, we opted to use prompting without any fine tuning. Due to costs, we used ChatGPT-3.5 for all experiments in our study.

Prompt selection is partly an art form, so it was beyond the scope of this paper to comprehensively search for the best prompt for this application. Instead, we constructed two representative prompts after some trial and error – one very simple (Plain) and another that relies on the Chain-of-thought (CoT) strategy, which makes ChatGPT think aloud while generating a response. All designed prompts can be found in Appendix **??**.

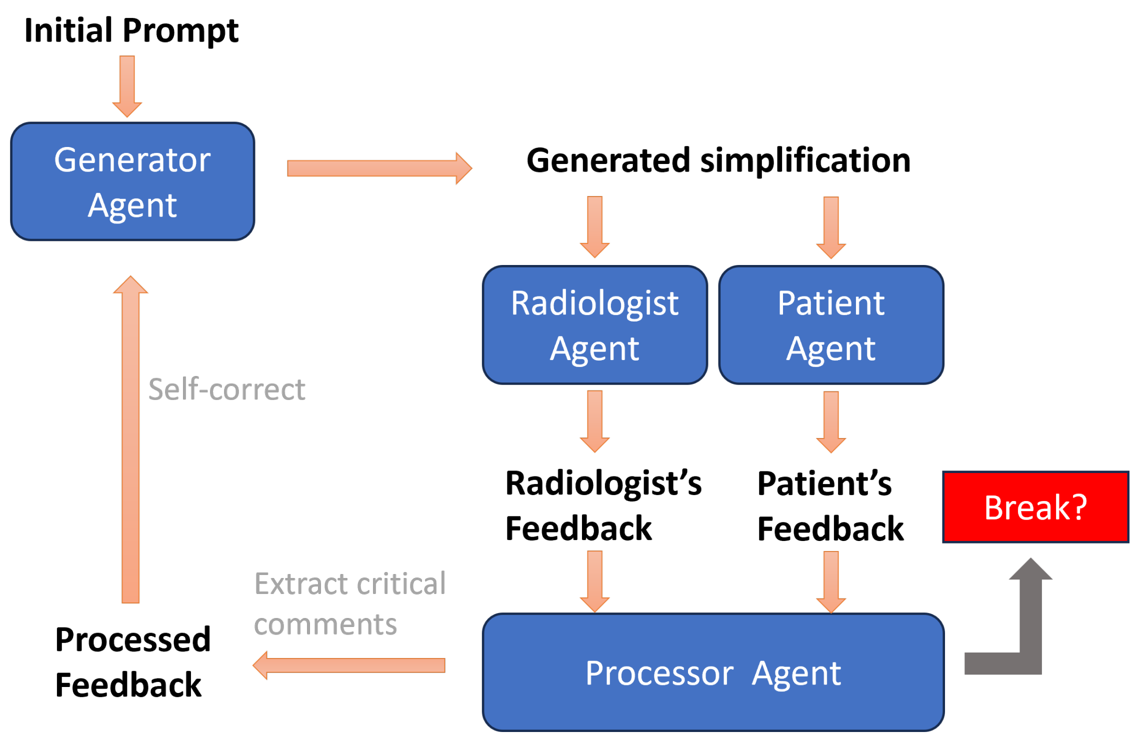

FIGURE 5.3: The workflow of self-correction mechanism. Processor agent decides when to stop the process.

5.3.2 Self-Correction Mechanism

Inspired by Madaan et al. (2023), we devised a Self-Correction mechanism for radiology report simplification. It relies on four differently instantiated ChatGPT agents: **Generator**, **Radiologist**, **Patient**, and **Processor**. The proposed workflow is shown in Figure 5.3. Given an original radiology sentence, Generator is asked to generate a simplification. Then, Radiologist and Patient provide feedback about the simplification. Finally, Processor summarizes the feedback and provides the summary to Generator who is asked to improve the simplifications. This process iterates among these four agents until Processor determines that no further improvement is needed. This self-correction mechanism can be applied to LLMs without any model training.

Inspired by Park et al. (2023), we instantiated Radiologist and Patient agents as distinct personas through distinct initial prompts shown in Figure 5.4 and 5.5. On the other hand, Generator and Processor agents are not prompted to become personas and are asked to provide an objective output. They are initialized using prompts that specify the task. Generator keeps the memory of conversation since it needs to refine the simplification based on the feedback from other agents. Generator is first provided a simple prompt for simplification.

Feedback generated by Radiologist and Patient agents is summarized by Processor to reduce the redundancy. We asked Processor agent to first decide if there is any critical comment or improvement suggestion in the generated feedback. If so, Processor summarizes the feedback and passes it back to Generator using a 'refine prompt'. Otherwise, Processor generates a string starting with "No". In this case, the last simplification is saved as the self-correct simplification. The prompts are shown as follow. **Generator Agent** Generator is initialized with a simple objective prompt. We considered two specific prompts as follows:

- Plain prompt:

  Simplify the sentence: RADIOLOGY SENTENCE .

- CoT prompt:

  Sentence: RADIOLOGY SENTENCE .

  Can you list all the complicated medical terms and provide explanations that are understandable by laypeople? Finally, write a simplification of the original sentence that laypeople can understand.

You are an experienced radiologist, specialized in evaluating the quality of simplified radiology sentences. Your role is to analyze a given sentence, comparing it to its simplified version and assessing the quality and simplicity of the simplification.

Your assessment should be based on five key criteria:
Correctness: It's crucial that the simplification correctly reflects the content of the original radiology sentence.
Completeness: The simplification must retain all the essential information of the original radiology sentence.
Hallucination: The simplified sentence should not introduce any harmful or misleading interpretations that are not present in the original radiology sentence.
Structure: The simplification must include clear descriptions of the 1) body parts , 2) the findings (such as masses, injuries, stones, swelling), and 3) what these findings indicate (for instance, benign conditions, unclear diagnoses, indications of severity).
Simplicity: Do you think the Simplification is simple enough for laypeople without medial knowledge?

Radiology Sentence: <*RADIOLOGY SENTENCE*>
Simplification: <*SIMPLIFICATION*>

Please provide your feedback.

FIGURE 5.4: The persona of Radiologist agent and task instructions. It is used to initialize the ChatGPT model.

The response from Generator is saved as evaluated in our experiments. In addition, the response from Generator is used to start the self-correction mechanism. **Radiologist Agent** Human radiologists can adequately evaluate the factualness of radiology report simplifications. We mimic this by creating a Radiologist agent with an initial prompt that ask ChatGPT to pretend to have a persona of radiologist, following the related idea presented in Park et al. (2023); Li et al. (2023).

Text in blue in Figure 5.4 defines Radiologist persona. Text in green is an instruction consistent with the survey we designed for human radiologists and that was used in human

You are a person who does not have any medical knowledge. You've never taken a medical-related course or class and struggle to grasp medical concepts. Your task is to evaluate the clarity and comprehensibility of simplified medical information, providing feedback on whether you understand all the words and concepts presented. However, you must not comment on the factuality of the information nor attempt to further simplify the sentence yourself. Your perspective is vital to ensure the simplification does not contain any complicated medical terms that you cannot understand.

Simplification: <*SIMPLIFICATION*>

Please provide your feedback.

FIGURE 5.5: The persona of Patient agent and task instructions. It is used to initialize the ChatGPT model.

evaluation of simplifications.

Patient Agent Similar to Radiologist agent, we created a Patient agent to provide feedback about the understandability of the simplification from Generator. As shown in Figure 5.5, we asked Patient to act as a layperson who lacks medical knowledge and cannot understand complex medical concepts. Further instructions and warnings are specified to avoid generating comments that are beyond the ability of a layperson.

Processor Agent The feedback generated by Radiologist and Patient agents is summarized by Processor to reduce the redundancy. We asked Processor agent to first decide if there is any critical comment or improvement suggestion in the generated feedback. If so, Processor summarizes the feedback and passes it back to Generator using a 'refine prompt'. Otherwise, Processor generates a string starting with "No". In this case, the last simplification is saved as the self-correct simplification. The following is a prompt we used for asking the Processor agent to determine and parse critical comments. Moreover, Processor also decide whether break the iteration loop:

- Initial prompt for Processor:

  Feedback: FEEDBACK

  Are there any critical comments or improvement suggestions in Feedback? If so, extract them starting with "Yes". Otherwise, say "No".

The following prompt is used to ask Generator to improve its previous simplification:

- Refine prompt for Generator:

  Radiologist's feedback: PROCESSED FEEDBACK

  Patient's feedback: PROCESSED FEEDBACK

  Can you improve your simplification while keeping it concise?

The proposed variant of self-correction mechanism is designed to imitate a conversation that could occur between a real radiologist and a patient to generate a good simplification of a radiology report.

5.4 Experimental Design

5.4.1 Data

For our experimental evaluation, we worked with a radiologist to manually identify 40 diverse, representative sentences from the radiology reports in the public database MIMIC III Johnson et al. (2016) as our data set. These sentences were selected from randomly sampled radiology reports about the abdomen. Sentences were selected to range from relatively simple to relatively complex. Attention was paid to ensuring that the chosen sentences were self-contained and did not require reading the surrounding sentences to understand their meaning. This relatively small number allowed us to obtain statistically valid results while keeping the costs of our study manageable.

5.4.2 Types of Simplifications

For each of the 40 radiology sentences, we produced four simplifications using ChatGPT-3.5. The first is Plain_BS, which uses the plain prompt, while the second is CoT_BS, which uses the CoT prompt, both introduced in Section 5.3.1. The remaining two use self-correction explained in Section 5.3.2. The initial Generator prompt in self-corrected Plain_SC is the plain prompt, while it is CoT prompt in CoT_SC. We used the same default temperature value for ChatGPT of 0.8 in all generations.

5.4.3 Human Evaluation Protocol

As described in Section 5.2, we used two types of human evaluators to assess the quality of simplifications.

Radiologists. We recruited one radiologist to evaluate the factuality of all simplifications as described in Section 5.2.1. For further analysis, we also asked them to evaluate the simplicity via the question, "*Do you think laypeople can understand the sentence?*". Likert scores in the range 1-5 were used for all questions. In addition, the radiologist was encouraged to provide justifications for the ratings. Moreover, we asked the radiologist to estimate the severity of described medical condition in each sentence using the five levels of severity as described in Section 5.2.2. The severity question is the same as Q3 in the survey for laypeople. This allowed us to evaluate the accuracy of laypeople's guesses of severity.

Laypeople. We recruited eight laypeople to assess if the simplifications improve understanding. The participants were a mix of undergraduate and graduate students from a computer science department, none of whom had any training in medicine. Thus, they are representatives of highly-educated laypeople. Each layperson was asked to assess each original sentences together with one randomly selected simplification using the approach explained in Section 5.2.2. Then, they were asked to select the best and worst simplifications among the four choices as illustrated in Figure 5.2. Full instructions and data usages were informed before the evaluation process.

Table 5.1: FKGL scores and human evaluation results. In laypeople's evaluation, Q1: You understand the sentence? Q2: Can you guess the severity? Q3: What is the severity? Q4: Does simplification help you? Categorical answers are mapped to numeric types. Mean squared error (MSE) and accuracy (ACC) are presented for Q3.

Metrics	Original Sentence	Plain_BS	Plain_SC	CoT_BS	CoT_SC
FKGL ↓	12.344	8.813	7.010	7.178	8.548
Radiologist's Evaluation					
Correctness	5.000	4.725	4.650	4.500	4.625
Completeness	5.000	4.900	4.675	4.775	4.875
Hallucination	5.000	4.925	4.900	4.850	4.825
Structure	5.000	4.850	4.900	4.825	4.875
Simplicity	1.500	3.100	4.200	4.375	4.575
Laypeople's Evaluation					
Q1 (1 to 4)	1.801	2.475	3.225	3.341	**3.602**
Q2 (1 to 3)	1.579	1.825	2.325	2.398	**2.534**
Q3 (MSE ↓)	1.699	1.650	1.188	1.341	**1.068**
Q3 (ACC)	38.4%	38.8%	38.8%	42.0%	**52.3%**
Q4 (-1 to 2)	N/A	0.613	1.288	1.477	**1.705**

5.5 Results

5.5.1 Human Evaluation Results

Top half of Table 5.1 shows results from the evaluation conducted by a radiologist. The 'Original Sentence' column denotes the scores assigned to the original radiology sentences. The factuality of the original radiology sentences was rated as five, by default. Simplicity score for the original sentences was very low (1.50), indicating that most of the original sentences are not expected to be understood by laypeople. Simplicity score was much larger for simplified sentences and was the largest for the self-correction with CoT (CoT_SC) approach. This result is consistent with the FKGL scores in the first row of Table 5.1. It can be seen that readability of all 4 simplifications is significantly smaller (freshmen high school level) than for the original sentences (college level).

Factuality scores for all four types of simplifications remained close to perfect. Hallucination and Structure scores were particularly high. Correctness scores were comparably lower, indicating occasional lack of precision in simplifications. Interestingly, factuality scores of Plain_BS are higher than for the other three simplification methods. This reflects the trade-off between simplicity and factuality. We consider CoT_SC the best approach because it achieved the highest simplicity with a very marginal decrease in factuality.

5.5.2 Do Simplifications Help?

In the bottom half of Table 5.1, we evaluate laypeople responses about simplicity. Q1, Q2, and Q3 in Figure 5.2 were designed to assess laypeople understanding of both the original sentences and their simplified versions. Q4 directly evaluated the effectiveness of these simplifications. We converted the categorical responses into numerical values[1]. For the responses to Q3, we compared the participants' severity level choices with those of the radiologist and computed the Mean Squared Error (MSE) and Accuracy (ACC).

All simplifications had significantly higher simplicity scores than the original sentences on all questions. Notably, CoT_SC achieved the highest scores accross all simplicity questions which is consistent with the radiologist's rating.

Table 5.2: Confidence levels vs Mean squared errors and Accuracy for Q3. Higher confidence leads more accurate answers.

	Not at all	Low confidence	High confidence
MSE	1.920	1.380	**0.930**
Accuracy	30.7%	39.8%	**55.5%**

5.5.3 Confidence vs Accuracy

Table 5.2 compares the correlation between the laypeople's confidence and the actual understanding of the severity of described medical conditions. When laypeople report the lowest confidence (Not at all), they also achieve the lowest accuracy in predicting severity (30.7%), and when they report the highest confidence, the accuracy is the largest (55.5%). However, there is still a significant gap between confidence and actual understanding. Even

[1] Q1: 'Not at all' - 1; 'Completely' - 4. Q2: 'Not at all' - 1; 'High confidence' - 3. Q3: 'Critical' - 1; 'Healthy' - 5. Q4: 'Furthered confused' - -1; 'Much better' - 2

when highly confident, laypeople could correctly predict severity in just over half (55.5%) of the sentences. We conclude that the simplifications might need to state the severity level explicitly.

To gain a deeper insight, in Figure 5.6, we show the distribution of confidence levels by laypeople for the original sentences and for each type of simplifications. We can see that all four types of simplifications are helpful, with CoT_SC being the most successful.

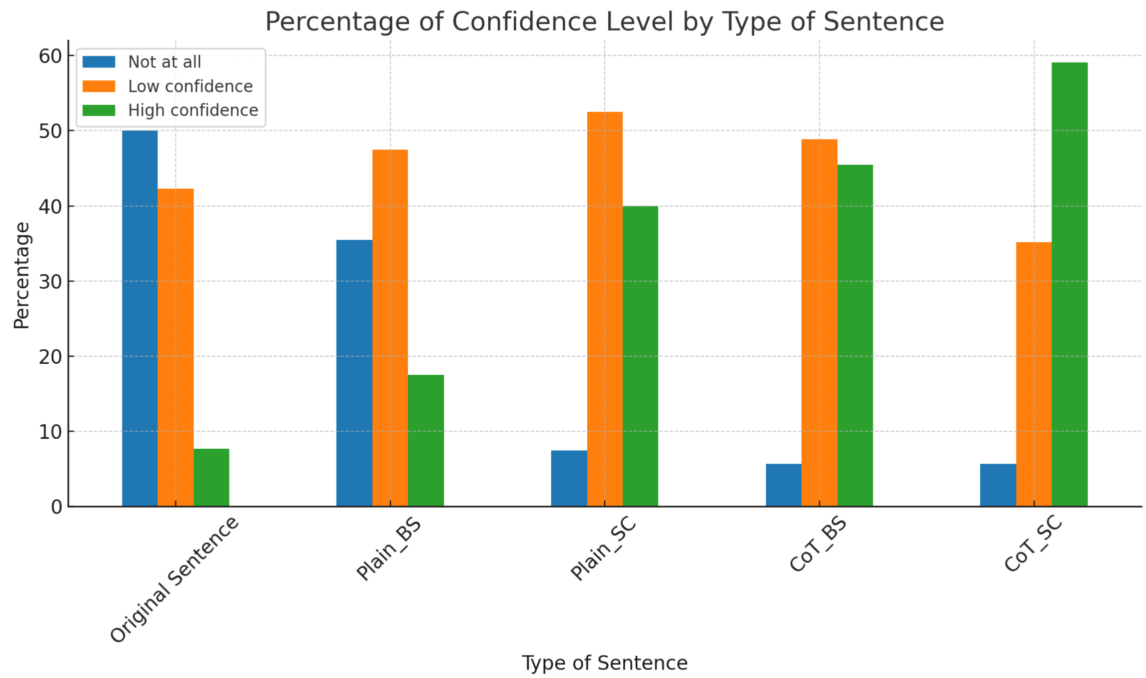

FIGURE 5.6: Distribution of confidence level (Q2) by laypeople given the original sentence and four types of simplifications

5.5.4 *Which Simplifications are Preferred by Laypeople?*

In this subsection, we report on the preferences of laypeople towards different types of simplifications (right panel in Figure 5.2). The findings are shown in Table 5.3, illustrating how often a specific simplification was deemed the most or least preferred based on the majority vote by the eight participants. The CoT_SC simplification was the clear favorite compared to the other three variants. On the other hand, Plain_BS simplification was the least favorite.

Table 5.3: Majority votes for the most and least preferences for all 40 sentences. The summation of the counts in each row may exceed 40 because we allow multiple selections.

	Plain_BS	Plain_SC	CoT_BS	CoT_SC
Most↑	2	7	15	**27**
Least↓	32	7	5	**2**

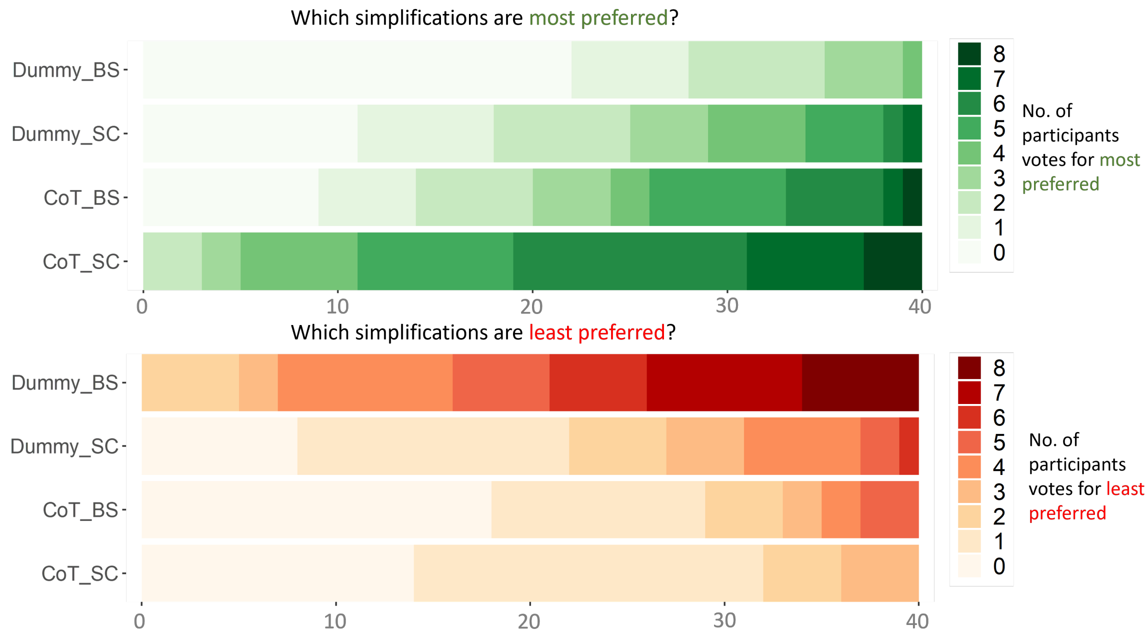

FIGURE 5.7: The horizontal stacked histogram of laypeople vote distribution for the most and least preferred simplifications.

To further investigate laypeople's preferences, we adopted the analysis technique outlined in Goyal et al. (2022). We calculated the inter-annotator agreement, applying Krippendorff's alpha with MASI distance Passonneau (2006), to account for the possibility of multiple selections for the best or worst simplifications in our research design. The alpha scores for the most and least preferred options were 0.177 and 0.132 respectively.

To visualize these results more clearly, we plot the distribution of laypeople's votes for each simplification version in Figure 5.7. The distribution of votes highlights that CoT_SC was chosen as the most preferred simplification by at least half of the participants in over 90% of the sentences. Additionally, both CoT_BS and Plain_SC received votes as the most preferred simplification from at least one participant for approximately 80% of the sentences. Unsurprisingly, based on the vote distribution, Plain_BS was the least preferred simplification version.

5.5.5 *Factual Errors in Generations*

ChatGPT is providing impressive simplifications, but it is not perfect. We observed that the slight decrease in factuality scores was caused by a few outliers. We noticed that during the self-correction process, Generator agent can occasionally generate incorrect information and hallucinate as it is trying to addresses feedback from Radiologist and Patient agents. There were in total eight factual errors observed by the radiologist evaluator among simplifications by the two self-correction approaches. In Table 5.4, we show and

Table 5.4: Four representative examples where the radiologist pointed out factual errors in generated simplifications. The radiologist's comments are included.

ORI:	Few stable less than 5 mm hypodensities, too small to characterize, probably simple cysts.
SIM:	There are a few tiny fluid-filled sacs, called cysts, that show up as lighter areas on the scan. These cysts are very small, measuring less than 5 mm.
Comments:	It says there are cysts - the radiologist was not definitive. Original text said it is PROBABLY cysts. (**Completeness error**)
ORI:	Atrophic kidneys are noted bilaterally with multiple simple cysts.
SIM:	The kidneys have become smaller and are not functioning well. There are fluid-filled sacs in both kidneys.
Comments:	You cannot assess whether they functional well based on CT. (**Hallucination error**)
ORI:	Status post hysterectomy.
SIM:	After having a hysterectomy, which is a surgery to remove the uterus, the person is in the recovery phase.
Comments:	The sentence simply says surgery was done sometime in the past - not that the patient is in recovery. (**Hallucination error**)
ORI:	Small calcified granulomata in the left hepatic lobe.
SIM:	There are small hardened clusters of immune cells in the left part of the liver.
Comments:	They are not immune cells. (**Correctness error**)

analyze four examples to point to typical errors. ChatGPT may generate hallucinations through over-interpreting the sentences. We can also see that ChatGPT may not use perfectly precise simple words to explain the complex medical jargon. However, ChatGPT is overall very impressive in producing high-quality text simplifications and is quite capable of self-correction.

Table 5.5: Representative examples where laypeople voted CoT_SC as their most preferred simplification. Laypeople justifications are listed. The ground-truth severity level is appended to the end of the original sentences. The laypeople's comments are included.

ORI:	There are metastatic bony lesions throughout the thorax. (Critical)
CoT_SC:	The doctors found that cancer from somewhere else has spread to the bones in the chest.
Comments:	It explains well.
ORI:	There is grade 1 anterolisthesis of L4 on L5. (Mild)
CoT_SC:	The diagnosis indicates a **mild** condition where the fourth bone in the lower part of the spine is slipping slightly forward over the fifth bone. This finding suggests a relatively **low level of severity**.
Comments-1:	The language is simple and easy to understand. I like that it tells the severity of the condition.
Comments-2:	Accurate and detailed.
ORI:	There is no intrahepatic or extra-hepatic biliary ductal dilatation. (Healthy)
CoT_SC:	The tubes both inside and outside the liver that carry bile (a digestive fluid) are not enlarged or widened, which is **usually a good sign** and means there is likely no blockage.
Comments-1:	I like that it mentions "usually a good sign" because it indicates the severity of the condition.
ORI:	Partially visualized central pulmonary arteries are not dilated. (Healthy)
CoT_SC:	The part of the blood vessels in the middle of the lungs that we can see is not enlarged, which is **normal**.
Comments-1:	Actually tells me that it's normal.
Comments-2:	Normal was important to me.
ORI:	The osseous structures are diffusely demineralized. (Moderate)
CoT_SC:	The bones throughout the body have lost minerals and become weaker. This widespread loss of minerals may have implications for the overall strength and health of the bones.
Comments:	Straight to the point. I needed to know that the bones are weakened.

5.5.6 *Preferences of Laypeople*

In Table 5.5, we show five examples from laypeople responses and discuss why CoT_SC could enhance comprehension. These examples are selected because they are representatives of sentences with different severity levels. We observe that participants prefer a simplification that 1) explains the medical condition in detail, 2) uses simple language, 3) indicates the severity of the condition. The CoT_SC simplification in the second example implies a mild severity level, which is not explicitly stated in the original sentence.

5.6 Conclusion

This paper introduces a two-pronged approach for human evaluation of radiology report simplifications. It proposes a specialized variant of the self-correction mechanism that allows ChatGPT to generate high-quality simplifications. The analysis of results derived from human evaluation show that our proposed evaluation protocol successfully reveals diverse facets of simplification quality.

5.7 Limitations

The first limitation of our study is that it focuses on simplification of individual sentences. Descriptions of some radiology findings are complex and require multiple sentences. While we do not expect LLMs to struggle with simplifying multiple sentences, an additional challenge would be extracting multi-sentence findings.

The second limitation is associated with simplifying the whole reports that often have multiple findings. While a trivial approach might consist of chunking the text into logical units and simplifying each unit separately, this approach might result in overly long simplification. Thus, it might be necessary to identify and simplify only the most significant findings from the report.

The third limitation is that we used only 40 original radiology sentences in the experimental evaluation. Ideally, we would like to consider a much larger set of sentences. However, the cost associated with this would be prohibitive. There are large computational costs associated with the self-correcting algorithms because they require multiple calls to ChatGPT to create a single simplification. There are also significant costs associated with human evaluation. It took laypeople over two hours on average to finish all the needed evaluations. It took the radiologist even longer. We estimated that 40 sentences were the minimum that allowed us to evaluate our ideas. We note that we made an effort to make those sentences representative of the radiology report diversity.

The fourth limitation of the study is that we obtained expert evaluation from a single radiologist. In fact, we recruited two more volunteer radiologists for our research, but neither was able to to finish the evaluation due to its length. Thus, we decided not to use their partial responses in the paper. It will be important for future studies to recruit multiple radiologists to estimate the factualness better and obtain a more complete understanding of their simplification preferences. It would also allow us to measure the inter-rater reliability. To be more successful, we need to make our survey easier to complete.

The fifth limitation is that our laypeople were college-educated individuals. It would be important in future research to recruit a more diverse group of laypeople and paint a more complete picture of the quality of simplifications and preferred types of simplifications.

CHAPTER 6

CONCLUSION

This compilation of studies has pushed forward the understanding and methodologies in the field of medical-related sentence simplification, with a particular focus on radiology reports. Across these investigations, we observe a concerted effort to leverage large language models (LLMs) for the simplification task, demonstrating the utility and challenges of applying such models to medical texts. The first study showcased the development and evaluation of GPT models tailored to radiology sentence simplification, emphasizing the effectiveness of prompting and fine-tuning strategies. It highlighted the superiority of a methodical approach to selecting the best answer from multiple GPT outputs over a baseline strategy.

The second work introduced novel data augmentation methods aimed at addressing the scarcity of labeled data in the domain. It underscored the significant impact of these methods on the quality of simplification models through both automatic measures and human evaluation. The study also illuminated the necessity of extending the simplification task beyond individual sentences to entire paragraphs, noting the limitations of focusing solely on liver-related sentences and the importance of diversifying both the data and evaluators involved in future research.

The third study presented a dual approach to evaluating radiology report simplifications, incorporating a specialized variant of the self-correction mechanism for ChatGPT. This study's evaluation protocol effectively captured various aspects of simplification quality. Despite its contributions, it acknowledged several limitations, including the focus on individual sentences, the challenge of simplifying entire reports with multiple findings, the small sample size used in experiments, and the need for broader participant diversity in evaluations.

Future Research

Objective 1: Comprehensive Paragraph-Level Simplification

Future studies should prioritize extending the simplification task to paragraph-level texts, especially considering the complexity and contextual nature of radiology reports. This involves developing advanced natural language processing algorithms capable of understanding and preserving the coherence and context of multi-sentence paragraphs. We should focus on creating models that can discern and retain critical information while eliminating redundant or non-essential content. This expansion will address the challenge of simplifying complex medical findings that cannot be adequately communicated in single sentences, ensuring that simplifications remain accurate and contextually informed.

Objective 2: LLM-based Automated Evaluator

Across our investigations, we observed the impressive creative capabilities and knowledge encapsulated within large language models (LLMs). For future endeavors, our focus should shift towards enhancing the generalizability of the human evaluation process. This can be achieved by adapting a pre-trained LLM to serve as a reliable evaluator for specific tasks, thereby ensuring a more uniform and accurate assessment framework.

Objective 3: Better Understanding Radiology Reports with Radiographs

With the rapid advancements in LLMs, the integration of multimodal tasks has gained increasing traction within the natural language processing (NLP) domain. Graphical content significantly aids in the comprehension of textual information by providing additional context and detail. In the field of radiology, radiographs, including X-rays and gamma rays, are often paired with corresponding radiology reports. By incorporating both graphical and textual inputs, the generation of simplifications could lead to more precise and accurate results than those achieved through purely text-based NLP approaches. This multimodal approach holds the promise of unlocking a deeper and more nuanced understanding of radiology reports, potentially revolutionizing the way medical information is simplified and communicated.

BIBLIOGRAPHY

Adduru, V., Hasan, S. A., Liu, J., Ling, Y., Datla, V. V., Qadir, A., & Farri, O. (2018), "Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification," in *KHD@ IJCAI*.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019), "Publicly available clinical BERT embeddings," *arXiv preprint arXiv:1904.03323*.

Alva-Manchego, F., Scarton, C., & Specia, L. (2019), "Cross-Sentence Transformations in Text Simplification." in *WNLP@ ACL*, pp. 181–184.

Alva-Manchego, F., Scarton, C., & Specia, L. (2020), "Data-driven sentence simplification: Survey and benchmark," *Computational Linguistics*, 46, 135–187.

Alva-Manchego, F., Scarton, C., & Specia, L. (2021), "The (un) suitability of automatic evaluation metrics for text simplification," *Computational Linguistics*, 47, 861–889.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., & Zwerdling, N. (2020), "Do not have enough data? Deep learning to the rescue!" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 7383–7390.

Apfel, F. & Tsouros, A. D. (2013), "Health literacy: the solid facts," *Copenhagen: World Health Organization*, pp. 3–26.

Attal, K., Ondov, B., & Demner-Fushman, D. (2023), "A dataset for plain language adaptation of biomedical abstracts," *Scientific Data*, 10, 8.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. (2022), "Constitutional AI: Harmlessness from AI Feedback," *arXiv preprint arXiv:2212.08073*.

Bayer, M., Kaufhold, M.-A., & Reuter, C. (2021), "A survey on data augmentation for text classification," *ACM Computing Surveys*.

Biran, O., Brody, S., & Elhadad, N. (2011), "Putting it simply: a context-aware approach to lexical simplification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 496–501.

Bodenreider, O. (2004), "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, 32, D267–D270.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020), "Language models are few-shot learners," *Advances in neural information processing systems*, 33, 1877–1901.

Chandrasekar, R., Doran, C., & Bangalore, S. (1996), "Motivations and methods for text simplification," in *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Chaves, A., Kesiku, C., & Garcia-Zapirain, B. (2022), "Automatic Text Summarization of Biomedical Text Data: A Systematic Review," *Information*, 13, 393.

Chen, J., Druhl, E., Polepalli Ramesh, B., Houston, T. K., Brandt, C. A., Zulman, D. M., Vimalananda, V. G., Malkani, S., & Yu, H. (2018), "A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews," *Journal of medical Internet research*, 20, e26.

Chen, X., Lin, M., Schärli, N., & Zhou, D. (2023), "Teaching large language models to self-debug," *arXiv preprint arXiv:2304.05128*.

Chintagunta, B., Katariya, N., Amatriain, X., & Kannan, A. (2021), "Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization," in *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pp. 66–76.

Cook, T. S., Oh, S. C., & Kahn Jr, C. E. (2017), "Patients' use and evaluation of an online system to annotate radiology reports with lay language definitions," *Academic radiology*, 24, 1169–1174.

Cripwell, L., Legrand, J., & Gardent, C. (2022), "Controllable Sentence Simplification via Operation Classification," in *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2091–2103.

Damay, J. J. S., Lojico, G. J. D., Lu, K. A. L., & Tarantan, D. B. (2006), "SimText: Text simplication of medical literature," .

Delbanco, T., Walker, J., Bell, S. K., Darer, J. D., Elmore, J. G., Farag, N., Feldman, H. J., Mejilla, R., Ngo, L., Ralston, J. D., et al. (2012), "Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead," *Annals of internal medicine*, 157, 461–470.

Devaraj, A., Wallace, B. C., Marshall, I. J., & Li, J. J. (2021), "Paragraph-level simplification of medical texts," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2021, p. 4972, NIH Public Access.

Devaraj, A., Sheffield, W., Wallace, B. C., & Li, J. J. (2022), "Evaluating factuality in text simplification," *arXiv preprint arXiv:2204.07562*.

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

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018), "Understanding back-translation at scale," *arXiv preprint arXiv:1808.09381*.

Enayati, S., Yang, Z., Lu, B., & Vucetic, S. (2021), "A Visualization Approach for Rapid Labeling of Clinical Notes for Smoking Status Extraction," in *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pp. 24–30, Online, Association for Computational Linguistics.

Fu, Y., Peng, H., Khot, T., & Lapata, M. (2023), "Improving language model negotiation with self-play and in-context learning from ai feedback," *arXiv preprint arXiv:2305.10142*.

Garimella, A., Sancheti, A., Aggarwal, V., Ganesh, A., Chhaya, N., & Kambhatla, N. (2022), "Text Simplification for Legal Domain:{I} nsights and Challenges," in *Proceedings of the Natural Legal Language Processing Workshop 2022*, pp. 296–304.

Goldberg-Stein, S. & Chernyak, V. (2019), "Adding value in radiology reporting," *Journal of the American College of Radiology*, 16, 1292–1298.

Goyal, T., Li, J. J., & Durrett, G. (2022), "News summarization and evaluation in the era of gpt-3," *arXiv preprint arXiv:2209.12356*.

Guo, Y., Qiu, W., Wang, Y., & Cohen, T. (2021), "Automated lay language summarization of biomedical scientific reviews," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 160–168.

Guo, Y., Qiu, W., Leroy, G., Wang, S., & Cohen, T. (2022), "Cells: A parallel corpus for biomedical lay language generation," *arXiv preprint arXiv:2211.03818*.

Guo, Y., August, T., Leroy, G., Cohen, T., & Wang, L. L. (2023), "APPLS: A Meta-evaluation Testbed for Plain Language Summarization," *arXiv preprint arXiv:2305.14341*.

Hasan, S. A., Liu, B., Liu, J., Qadir, A., Lee, K., Datla, V., Prakash, A., & Farri, O. (2016), "Neural clinical paraphrase generation with attention," in *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pp. 42–53.

Hong, M. K., Feustel, C., Agnihotri, M., Silverman, M., Simoneaux, S. F., & Wilcox, L. (2017), "Supporting families in reviewing and communicating about radiology imaging studies," in *Proceedings of the 2017 CHI conference on human factors in computing systems*, pp. 5245–5256.

Honnibal, M. & Montani, I. (2017), "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," To appear.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2022), "Large language models can self-improve," *arXiv preprint arXiv:2210.11610*.

Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., St"uber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B., Ricke, J., et al. (2022), "ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports," *arXiv preprint arXiv:2212.14882*.

Jiang, C., Maddela, M., Lan, W., Zhong, Y., & Xu, W. (2020), "Neural CRF model for sentence alignment in text simplification," *arXiv preprint arXiv:2005.02324*.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., & Mark, R. G. (2016), "MIMIC-III, a freely accessible critical care database," *Scientific data*, 3, 1–9.

Kalia, V. (2020), "Radiology Reports Must Adapt as Patient Needs Evolve and Access Improves," *Academic Radiology*, 27, 440–441.

Kandula, S., Curtis, D., & Zeng-Treitler, Q. (2010), "A semantic and syntactic text simplification tool for health content," in *AMIA annual symposium proceedings*, vol. 2010, p. 366, American Medical Informatics Association.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975), "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Tech. rep., Naval Technical Training Command Millington TN Research Branch.

Kutner, M., Greenburg, E., Jin, Y., & Paulsen, C. (2006), "The Health Literacy of America's Adults: Results from the 2003 National Assessment of Adult Literacy. NCES 2006-483." *National Center for education statistics*.

Laban, P., Schnabel, T., Bennett, P., & Hearst, M. A. (2021), "Keep it Simple: Unsupervised Simplification of Multi-Paragraph Text," *arXiv preprint arXiv:2107.03444*.

Lalor, J. P., Wu, H., Chen, L., Mazor, K. M., & Yu, H. (2018), "ComprehENotes, an instrument to assess patient reading comprehension of electronic health record notes: development and validation," *Journal of medical Internet research*, 20, e139.

Leroy, G., Kauchak, D., & Hogue, A. (2016), "Effects on text simplification: Evaluation of splitting up noun phrases," *Journal of health communication*, 21, 18–26.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019), "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*.

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023), "Camel: Communicative agents for" mind" exploration of large scale language model society," *arXiv preprint arXiv:2303.17760*.

Liang, S., Kades, K., Fink, M., Full, P., Weber, T., Kleesiek, J., Strube, M., & Maier-Hein, K. (2022), "Fine-tuning BERT models for summarizing German radiology findings," in *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pp. 30–40.

Lin, C.-Y. (2004), "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81.

Lipscomb, C. E. (2000), "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association*, 88, 265.

Liu, C., Tian, Y., & Song, Y. (2023a), "A Systematic Review of Deep Learning-based Research on Radiology Report Generation," *arXiv preprint arXiv:2311.14199*.

Liu, P., Wang, X., Xiang, C., & Meng, W. (2020), "A survey of text data augmentation," in *2020 International Conference on Computer Communication and Network Security (CCNS)*, pp. 191–195, IEEE.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021), "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *arXiv preprint arXiv:2107.13586*.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023b), "Gpteval: Nlg evaluation using gpt-4 with better human alignment," *arXiv preprint arXiv:2303.16634*.

Llanos, L. C., Bouamor, D., Zweigenbaum, P., & Rosset, S. (2016), "Managing linguistic and terminological variation in a medical dialogue system," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3167–3173.

Lourenco, A. P. & Baird, G. L. (2020), "Optimizing radiology reports for patients and referring physicians: mitigating the curse of knowledge," *Academic radiology*, 27, 436–439.

Lu, J., Li, J., Wallace, B. C., He, Y., & Pergola, G. (2023), "Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization," *arXiv preprint arXiv:2302.05574*.

Lyu, Q., Tan, J., Zapadka, M. E., Ponnatapuram, J., Niu, C., Wang, G., & Whitlow, C. T. (2023), "Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: Promising results, limitations, and potential," *arXiv preprint arXiv:2303.09038*.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. (2023), "Self-refine: Iterative refinement with self-feedback," *arXiv preprint arXiv:2303.17651*.

Miller, G. A. (1995), “WordNet: a lexical database for English,” *Communications of the ACM*, 38, 39–41.

Mukherjee, P., Leroy, G., Kauchak, D., Rajanarayanan, S., Diaz, D. Y. R., Yuan, N. P., Pritchard, T. G., & Colina, S. (2017), “NegAIT: A new parser for medical text simplification using morphological, sentential and double negation,” *Journal of biomedical informatics*, 69, 55–62.

Nisioi, S., Štajner, S., Ponzetto, S. P., & Dinu, L. P. (2017), “Exploring neural text simplification models,” in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pp. 85–91.

Oh, S. C., Cook, T. S., & Kahn, C. E. (2016), “PORTER: a prototype system for patient-oriented radiology reporting,” *Journal of digital imaging*, 29, 450–454.

Ondov, B., Attal, K., & Demner-Fushman, D. (2022), “A survey of automated methods for biomedical text simplification,” *Journal of the American Medical Informatics Association*, 29, 1976–1988.

OpenAI (2023), “GPT-4 Technical Report,” .

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002), “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318.

Park, J. S., O’Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023), “Generative agents: Interactive simulacra of human behavior,” *arXiv preprint arXiv:2304.03442*.

Passonneau, R. (2006), “Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation,” .

Pattisapu, N., Prabhu, N., Bhati, S., & Varma, V. (2020), “Leveraging Social Media for Medical Text Simplification,” in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 851–860.

Qenam, B., Kim, T. Y., Carroll, M. J., & Hogarth, M. (2017), “Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation,” *Journal of medical Internet research*, 19, e417.

Rets, I. & Rogaten, J. (2021), “To simplify or not? Facilitating English L2 users’ comprehension and processing of open educational resources in English using text simplification,” *Journal of Computer Assisted Learning*, 37, 705–717.

Rosenkrantz, A. B. & Flagg, E. R. (2015), “Survey-based assessment of patients’ understanding of their own imaging examinations,” *Journal of the American College of Radiology*, 12, 549–555.

Ross, S. E. & Lin, C.-T. (2003), “The effects of promoting patient access to medical records: a review,” *Journal of the American Medical Informatics Association*, 10, 129–138.

Sakakini, T., Lee, J. Y., Duri, A., Azevedo, R. F., Sadauskas, V., Gu, K., Bhat, S., Morrow, D., Graumlich, J., Walayat, S., et al. (2020), “Context-Aware Automatic Text Simplification of Health Materials in Low-Resource Domains,” in *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pp. 115–126.

Schick, T. & Schütze, H. (2020), “Exploiting cloze questions for few shot text classification and natural language inference,” *arXiv preprint arXiv:2001.07676*.

Shardlow, M. (2014), “A survey of automated text simplification,” *International Journal of Advanced Computer Science and Applications*, 4, 58–70.

Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., & Singh, S. (2020), “Autoprompt: Eliciting knowledge from language models with automatically generated prompts,” *arXiv preprint arXiv:2010.15980*.

Sudore, R. L., Yaffe, K., Satterfield, S., Harris, T. B., Mehta, K. M., Simonsick, E. M., Newman, A. B., Rosano, C., Rooks, R., Rubin, S. M., et al. (2006), “Limited literacy and mortality in the elderly: the health, aging, and body composition study,” *Journal of general internal medicine*, 21, 806–812.

Sun, R., Jin, H., & Wan, X. (2021), “Document-level text simplification: Dataset, criteria and baseline,” *arXiv preprint arXiv:2110.05071*.

Surya, S., Mishra, A., Laha, A., Jain, P., & Sankaranarayanan, K. (2018), “Unsupervised neural text simplification,” *arXiv preprint arXiv:1810.07931*.

Van, H., Kauchak, D., & Leroy, G. (2020), “AutoMeTS: the autocomplete for medical text simplification,” *arXiv preprint arXiv:2010.10573*.

Van den Bercken, L., Sips, R.-J., & Lofi, C. (2019), “Evaluating neural text simplification in the medical domain,” in *The World Wide Web Conference*, pp. 3286–3292.

Wang, Y., Xu, C., Sun, Q., Hu, H., Tao, C., Geng, X., & Jiang, D. (2022), “PromDA: Prompt-based Data Augmentation for Low-Resource NLU Tasks,” *arXiv preprint arXiv:2202.12499*.

Wei, J. & Zou, K. (2019), “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6383–6389, Hong Kong, China, Association for Computational Linguistics.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. (2022), “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in Neural Information Processing Systems*, 35, 24824–24837.

Weng, J.-C., Chou, Y.-S., Huang, G.-J., Tyan, Y.-S., & Ho, M.-C. (2018), "Mapping brain functional alterations in betel-quid chewers using resting-state fMRI and network analysis," *Psychopharmacology*, 235, 1257–1271.

Weng, W.-H., Chung, Y.-A., & Szolovits, P. (2019), "Unsupervised clinical language translation," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3121–3131.

Woodsend, K. & Lapata, M. (2011), "Learning to simplify sentences with quasi-synchronous grammar and integer programming," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 409–420, Association for Computational Linguistics.

Xu, W., Callison-Burch, C., & Napoles, C. (2015), "Problems in current text simplification research: New data can help," *Transactions of the Association for Computational Linguistics*, 3, 283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016), "Optimizing statistical machine translation for text simplification," *Transactions of the Association for Computational Linguistics*, 4, 401–415.

Yang, Z., Cherian, S., & Vucetic, S. (2023), "Data Augmentation for Radiology Report Simplification," in *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1877–1887.

Yi, P. H., Golden, S. K., Harringa, J. B., & Kliewer, M. A. (2019), "Readability of lumbar spine MRI reports: will patients understand?" *American Journal of Roentgenology*, 212, 602–606.

Yoo, K. M., Park, D., Kang, J., Lee, S.-W., & Park, W. (2021), "GPT3Mix: Leveraging Large-scale Language Models for Text Augmentation," *arXiv preprint arXiv:2104.08826*.

Zeng, Q. T. & Tse, T. (2006), "Exploring and developing consumer health vocabularies," *Journal of the American Medical Informatics Association*, 13, 24–29.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020), "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," in *International Conference on Machine Learning*, pp. 11328–11339, PMLR.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019), "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*.

Zhang, Y., Ding, D. Y., Qian, T., Manning, C. D., & Langlotz, C. P. (2018), "Learning to summarize radiology findings," *arXiv preprint arXiv:1809.04698*.

Zhu, Z., Bernhard, D., & Gurevych, I. (2010), "A monolingual tree-based translation model for sentence simplification," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1353–1361.