Two-Pronged Human Evaluation of ChatGPT Self-Correction in Radiology Report Simplification

Ziyu Yang
CIS, Temple University
zyyang@temple.edu

Santhosh Cherian
Temple University Hospital
santhosh.cherian@tuhs.temple.edu

Slobodan Vucetic
CIS, Temple University
vucetic@temple.edu

Abstract

Radiology reports are highly technical documents aimed primarily at doctor-doctor communication. There has been an increasing interest in sharing those reports with patients, necessitating providing them patient-friendly simplifications of the original reports. This study explores the suitability of large language models in automatically generating those simplifications. We examine the usefulness of chain-of-thought and self-correction prompting mechanisms in this domain. We also propose a new evaluation protocol that employs radiologists and laypeople, where radiologists verify the factual correctness of simplifications, and laypeople assess simplicity and comprehension. Our experimental results demonstrate the effectiveness of self-correction prompting in producing high-quality simplifications. Our findings illuminate the preferences of radiologists and laypeople regarding text simplification, informing future research on this topic.

1 Introduction

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 and Lin, 2003), increase patient involvement in care (Delbanco et al., 2012), and improve outcomes (Rosenkrantz and Flagg, 2015). However, the health literacy (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, whose primary purpose is doctor-doctor communication. 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).

There is an increasing interest in patient-friendly reporting. One way to accomplish this is to ask radiologists to supplement their expert-language reports with patient-friendly summaries. One downside of this approach is a negative impact on radiologists' cognitive load and productivity. Another downside is the curse of knowledge (Camerer et al., 1989), making it challenging for radiologists to simplify their reports. 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

Findings of the Association for Computational Linguistics ACL 2024, pages 4701–4714
August 11-16, 2024

constitutes a reasonable radiology simplification, we think humans should perform the evaluation.

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 paper, 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.

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.

2 Related Work

2.1 Radiology text simplification

Traditionally, text simplification referred to lexical simplification that paraphrases text (Chen et al., 2018; Biran et al., 2011; Weng et al., 2018). More recently, it shifted towards semantic simplification that seeks to simplify grammatically complex text (Shardlow, 2014; Leroy et al., 2016). 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. We think that summarization is not the best term because it implies text compression, while text simplification is more interested in text understanding, which allows creation of text longer than the original.

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).

2.2 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 com-

pare 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), Gunning Fog Index (GFI), and Automated Readability Index (ARI) are widely recognized metrics that estimate 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.

2.3 Prompting strategies for LLMs

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.

3 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.

3.1 Factuality (for experts)

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.

- **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

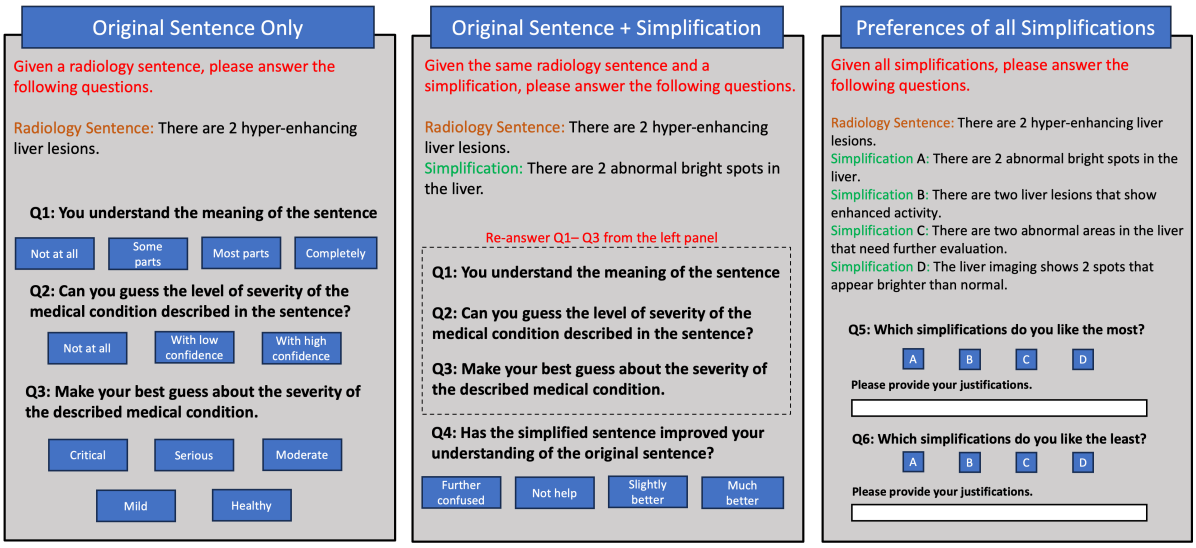

Figure 1: Layperson evaluation of radiology report simplifications. (a) (**left panel**) evaluates whether laypeople understand the original sentence. (b) (**middle panel**) evaluates whether simplification improves understanding. (c) (**right panel**) evaluates the preferences given a set of candidate simplifications and asks for justification.

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. Appendix A.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.

3.2 Simplicity (for laypeople)

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 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 1. 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 in Appendix A.2. 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 1. This free-text re-

sponse can be used in qualitative analysis of laypeople's simplification preferences.

4 Prompting and Self-Correction

4.1 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 A.3.1.

4.2 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 2. 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 Appendix A.4. 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 in Appendix A.3.4.

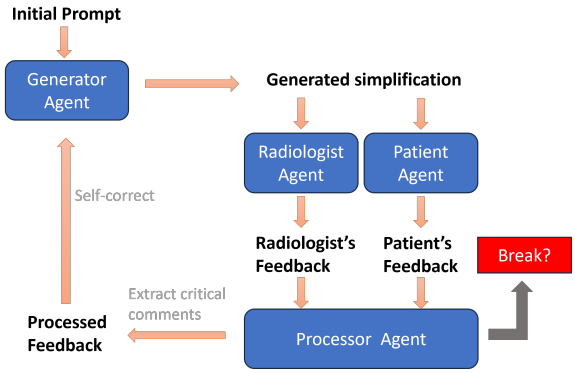

Figure 2: The workflow of self-correction mechanism. Processor agent decides when to stop the process.

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 Experimental Design

5.1 Data

For our experimental evaluation, we identified 40 diverse, representative sentences from the radiology reports in the public database MIMIC III (Johnson et al., 2016). To arrive at those 40 sentences, a radiologist read 100 randomly selected MIMIC III radiology reports and marked self-contained sentences about findings. The list of marked sentences was then narrowed down by removing redundancy and trying to maintain a diversity of findings in the reports. The selected sentences include a variety of findings (lesions, masses, obstructions, nodular surface, infection), conditions (being enlarged, abnormal, shrunken) of many anatomical parts in the abdomen (liver, kidney, pancreas, intestine, bones). 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

Table 1: FKGL, GFI, ARI 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
GFI ↓	19.011	14.632	11.942	10.990	12.371
ARI ↓	9.940	6.463	4.941	4.404	6.006
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**

5.2 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 4.1. The remaining two use self-correction explained in Section 4.2. 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.

5.3 Automated Metrics

In our study we did not generate ground truth simplifications because there might be a variety of acceptable simplifications with different lengths and levels of detail. As a result, our automated metrics only include three reference-free measures of simplicity: Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), and Automated Readability Index (ARI). The FKGL gives a grade-level score, indicating the minimum education level needed to comprehend the text. The GFI estimates the years of formal education required for understanding by focusing on sentence length and the frequency of complex words. The ARI, on the other hand, calculates readability based on the number of characters per word and words per sentence.

5.4 Human Evaluation Protocol

As described in Section 3, 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 3.1. 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 the Appendix A.2. The distribution of severity scores on our data is: Critical: 4, Serious: 2, Moderate: 7, Mild: 20, Healthy: 7. 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. For each of the 40 sentences, each layperson was given the questionnaire in Figure 1. First, they answered 3 questions about the original sentence (left panel). Then, we selected one of the four simplified sentences using the Latin square design and asked them 4 questions from the middle panel to see if it improved their understanding compared to the original sentence alone. As the layperson was already starting to understand

the original sentence, we did not repeat the middle panel questions for the 3 remaining simplifications. Instead, we asked a layperson the right panel questions to find which simplifications they liked the most and least. The simplifications in the right panel were listed at random to prevent bias.

Selected sentences, generated simplifications, and evaluation answers from the radiologist and laypeople are released.[1]

6 Results

6.1 Human Evaluation Results

Top half of Table 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 automated readability scores in the first three rows of Table 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.

6.2 Do Simplifications Help?

In the bottom half of Table 1, we evaluate laypeople responses about simplicity. Q1, Q2, and Q3 in Figure 1 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[2]. 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 2: Confidence levels vs Mean squared errors and Accuracy for Q3

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

6.3 Confidence vs Accuracy

Table 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 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 3, 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.

6.4 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 1). The findings are shown in Table 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.

[1] https://github.com/Ziyu-Yang/Human-Evaluation

[2] 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

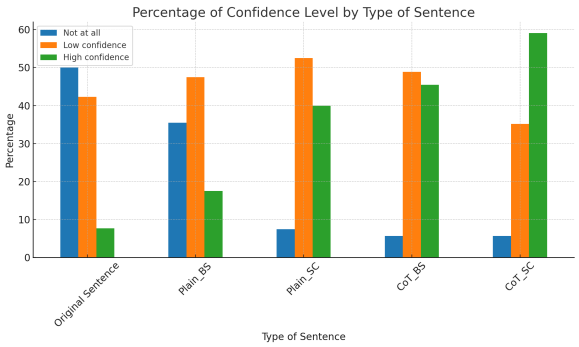

Figure 3: Distribution of confidence level (Q2) by laypeople given the original sentence and four types of simplifications

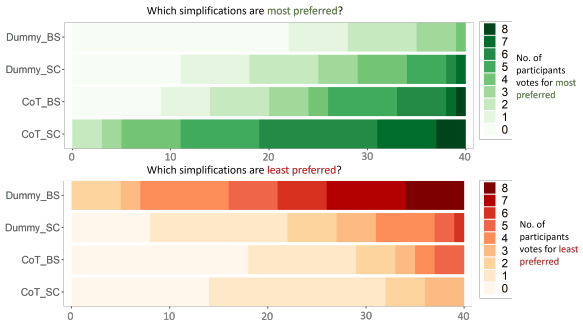

Figure 4: The horizontal stacked histogram of laypeople vote distribution for the most and least preferred simplifications.

Table 3: Majority votes for the most and least preferences for all 40 sentences.

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

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 4. 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.

6.5 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 4, we show and 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 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**)

6.6 Preferences of Laypeople

In Table 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

Table 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.

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.

7 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.

8 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.

References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Or Biran, Samuel Brody, and Noémie Elhadad. 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*, pages 496–501.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Colin Camerer, George Loewenstein, and Martin Weber. 1989. The curse of knowledge in economic settings: An experimental analysis. *Journal of political Economy*, 97(5):1232–1254.

Andrea Chaves, Cyrille Kesiku, and Begonya Garcia-Zapirain. 2022. Automatic text summarization of biomedical text data: A systematic review. *Information*, 13(8):393.

Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, and Hong Yu. 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(1):e26.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2022. Controllable sentence simplification via operation classification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2091–2103.

Tom Delbanco, Jan Walker, Sigall K Bell, Jonathan D Darer, Joann G Elmore, Nadine Farag, Henry J Feldman, Roanne Mejilla, Long Ngo, James D Ralston, et al. 2012. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. *Annals of internal medicine*, 157(7):461–470.

Ashwin Devaraj, William Sheffield, Byron C Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. *arXiv preprint arXiv:2204.07562*.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2023. Appls: A meta-evaluation testbed for plain language summarization. *arXiv preprint arXiv:2305.14341*.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Matthew K Hong, Clayton Feustel, Meeshu Agnihotri, Max Silverman, Stephen F Simoneaux, and Lauren Wilcox. 2017. Supporting families in reviewing and communicating about radiology imaging studies. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 5245–5256.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa St¨uber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. 2022. Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882*.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural crf model for sentence alignment in text simplification. *arXiv preprint arXiv:2005.02324*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Mark Kutner, Elizabeth Greenburg, Ying Jin, and Christine Paulsen. 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*.

John P Lalor, Hao Wu, Li Chen, Kathleen M Mazor, and Hong Yu. 2018. Comprehenotes, an instrument to assess patient reading comprehension of electronic health record notes: development and validation. *Journal of medical Internet research*, 20(4):e139.

Gondy Leroy, David Kauchak, and Alan Hogue. 2016. Effects on text simplification: Evaluation of splitting up noun phrases. *Journal of health communication*, 21(sup1):18–26.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large scale language model society. *arXiv preprint arXiv:2303.17760*.

Siting Liang, Klaus Kades, Matthias Fink, Peter Full, Tim Weber, Jens Kleesiek, Michael Strube, and Klaus Maier-Hein. 2022. Fine-tuning bert models for summarizing german radiology findings. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 30–40.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chang Liu, Yuanhe Tian, and Yan Song. 2023a. A systematic review of deep learning-based research on radiology report generation. *arXiv preprint arXiv:2311.14199*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Junru Lu, Jiazheng Li, Byron C Wallace, Yulan He, and Gabriele Pergola. 2023. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. *arXiv preprint arXiv:2302.05574*.

Qing Lyu, Josh Tan, Mike E Zapadka, Janardhana Ponnatapuram, Chuang Niu, Ge Wang, and Christopher T Whitlow. 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*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Brian Ondov, Kush Attal, and Dina Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

OpenAI. 2023. Gpt-4 technical report.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (masi) for semantic and pragmatic annotation.

Andrew B Rosenkrantz and Eric R Flagg. 2015. Survey-based assessment of patients' understanding of their own imaging examinations. *Journal of the American College of Radiology*, 12(6):549–555.

Stephen E Ross and Chen-Tan Lin. 2003. The effects of promoting patient access to medical records: a review. *Journal of the American Medical Informatics Association*, 10(2):129–138.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Rebecca L Sudore, Kristine Yaffe, Suzanne Satterfield, Tamara B Harris, Kala M Mehta, Eleanor M Simonsick, Anne B Newman, Caterina Rosano, Ronica Rooks, Susan M Rubin, 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.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jun-Cheng Weng, Yu-Syuan Chou, Guo-Joe Huang, Yeu-Sheng Tyan, and Ming-Chou Ho. 2018. Mapping brain functional alterations in betel-quid chewers using resting-state fmri and network analysis. *Psychopharmacology*, 235:1257–1271.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Ziyu Yang, Santhosh Cherian, and Slobodan Vucetic. 2023. Data augmentation for radiology report simplification. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1877–1887.

Paul Hyunsoo Yi, Sean Kenney Golden, John B Harringa, and Mark A Kliewer. 2019. Readability of lumbar spine mri reports: will patients understand? *American Journal of Roentgenology*, 212(3):602–606.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*.

A Appendix

A.1 Designed Survey for Radiologist

The exact design of the survey for the radiologist is shown in Figure 5

A.2 Definitions of Severity Levels

- 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.

A.3 Prompts

This section provides details about the design of four ChatGPT agents used in the self-correction mechanism outlined in Figure 2.

A.3.1 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.

The response from Generator is saved as evaluated in our experiments. In addition, the response from Genarator is used to start the self-correction mechanism illustrated in Figure 2.

A.3.2 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 6 defines Radiologist persona. Text in green is an instruction consistent with the survey we designed for human radiologists and that was used in human evaluation of simplifications.

A.3.3 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 7, 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.

A.3.4 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 Processor:

- **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

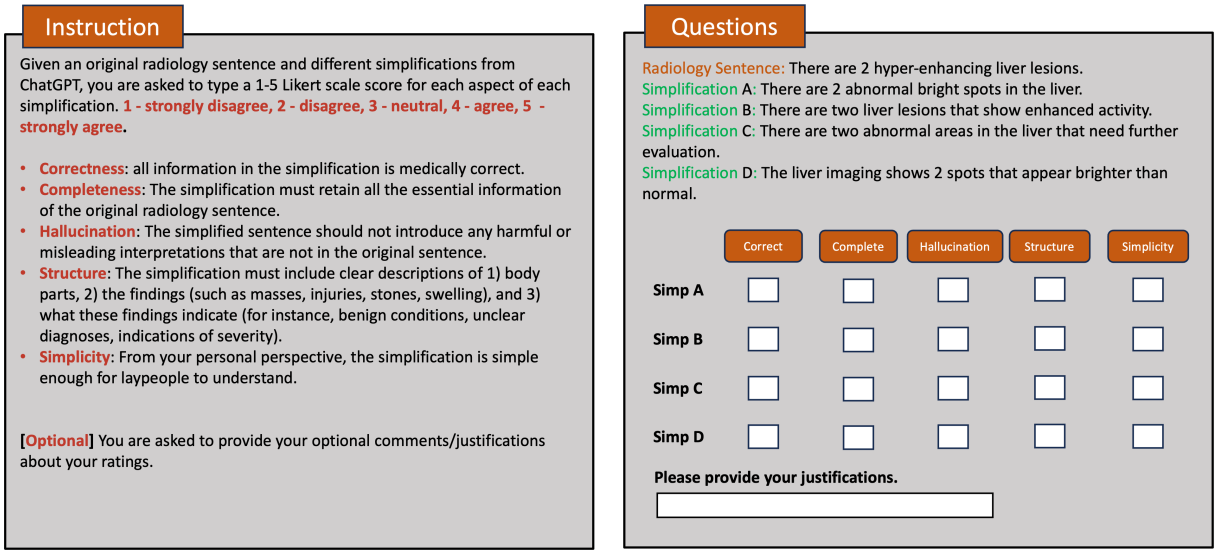

Figure 5: Expert evaluation of radiology report simplification. (**left panel**) lists instructions, (**left panel**) is a survey form with text boxes for ratings and justification.

FEEDBACK>
Patient's feedback: <PROCESSED FEEDBACK>
Can you improve your simplification while keeping it concise?

A.4 Personas

The designed personas for the Radiologist and Patient agents are shown in Figure 6 and 7

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 6: The persona of **Radiologist** agent and task instructions

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 7: The persona of **Patient** agent and task instructions