





Automated Medical Text Simplification for Enhanced Patient Access

Liliya Makhmutova¹^a, Giancarlo Salton²^b, Fernando Perez-Tellez¹^c and Robert Ross¹^d

¹Technological University Dublin, School of Computer Science, 191 North Circular Road, Dublin, Ireland

²Unochapecó, Servidão Anjo da Guarda, 295-D - Efapi, Chapeco, Brazil

Keywords: Medical Texts Simplification, LLM Evaluation.

Abstract: Doctors and patients have significantly different mental models related to the medical domain; this can lead to different preferences in terminology used to describe the same concept, and in turn, makes medical text often difficult to understand for the average person. However, getting access to a good understanding of patient notes, medical history, and other health-related documents is crucial for patients' recovery and sticking to a diet or medical procedures. Large language models (LLM) can be used to simplify and summarize text, yet there is no guarantee that the output will be correct and contain all the needed information. In this paper, we create and propose a new multi-modal medical text simplification dataset with pictorial explanations following along the aligned simplified and use it to evaluate the current state-of-the-art large language model (SOTA LLM) for the simplification task for the dataset and compare it to human-written texts. Our findings suggest that the current general-purpose LLMs are still not reliable enough for such in the medical sphere, though they may simplify texts quite well. The dataset and additional materials may be found at https://github.com/LiliyaMakhmutova/medical_texts_simplification.


1 INTRODUCTION


Medical texts can be very difficult to understand for patients, which may lead to health problems. More importantly, patients often don't have access to their medical records, and where they have, the patients often cannot understand the meaning due to the very different mental models and background knowledge that patients and clinicians have (Slaughter, 2005; Rote-gard et al., 2006). This leads to patients' partial exclusion from the recovery process and sub-optimal outcomes.


Medical texts usually contain lots of special terminology, many abbreviations, lack of coordination, subordination, and explanations in sentences making it harder to understand causal relationships. Moreover, medical texts usually consist of short ungrammatical sentences (Kandula et al., 2010). This makes their understanding difficult not only for laymen but also for healthcare professionals from other fields. Given these challenges, a machine learning model for medical text simplifications may be very beneficial


both in terms of democratising information access and improving outcomes. Although a model under no circumstances should add irrelevant information (making up some facts), it may incorporate true knowledge that is not mentioned in a report to make a medical text clearer and more understandable for a patient. So, for example, a model might add "Your mother or father are likely to have similar conditions too" explaining "genetic" reasons, but should not judge whether the blood sugar level in a patient is normal or not.

Currently, there are multiple datasets related to medical text simplification available (Basu et al., 2023; Luo et al., 2020; Sakakini and Lee, 2020; Van et al., 2020; Luo et al., 2022; Trienes et al., 2022). With the recent advance in the quality of the LLMs (OpenAI, 2023; et al., 2023; Chowdhery and et al., 2022; Li et al., 2023; Touvron and et al., 2023), more and more studies are investigating the quality of the LLMs' output by various benchmarks (Ariyaratne et al., 2023; Nascimento et al., 2023; Liao et al., 2023). Although current SOTA LLMs can produce texts of exceptional quality, there may be many problems related to it. The produced text may be biased, contain offensive language, or even include some made-up facts. The latter problem is known as hallucination (Manakul et al., 2023). The hallu-

^a <https://orcid.org/0000-0003-2191-4330>

^b <https://orcid.org/0000-0002-4301-7000>

^c <https://orcid.org/0000-0003-4978-2843>

^d <https://orcid.org/0000-0003-1449-1827>

cination may be also related to data leakage (Borkar, 2023) in a way that a model may reveal some output from its training data (which may be very bad for medical privacy), and there is no known direct way of controlling it.

In our paper, we make three contributions to the medical text simplification problem. Firstly, we create a multi-modal aligned dataset, which simplifies a subset of texts from Vydiswaran (2019) line-by-line, with some pictures illustrating the devices or procedures where appropriate. Secondly, we compared the human simplifications from the dataset with the output of an LLM tasked with simplification (namely ChatGPT (OpenAI, 2023) in this case). The comparison was based on multiple metrics such as similarity score, perplexity, and POS-tagging distribution, as well as congruence, fluency, and simplicity. We also conducted a survey, where we asked respondents' opinions on the quality of simplifications including questions on factual accuracy, complexity, structure, etc. Finally, we adapted a widely-used protocol on the quality of general simplifications (adequacy, fluency, and simplicity), by adding some rules for simplification in the medical field.

2 RELATED WORK

The importance of medical text simplification for patient has been noted by several authors. For example, Kandula et al. (2010) advocates a lexical-centric approach to the challenge by applying the Open Access and Collaborative Consumer Health Vocabulary (OAC. CHV) for terminology simplification.

Prior studies have shown that there are significant differences between a patient's and a healthcare professional's mental model of the medical domain and they prefer different terms to describe the same concept (Slaughter, 2005; Rotegard et al., 2006). This has been further reinforced by works that emphasise the clinical concerns related to simplification. The ethical (and not only) concerns related to simplification are outlined by Gooding (2022).

While most work in the area has looked at the simplification task, it is notable that some related work looks at the opposite problem. For example, Manzini et al. (2022) introduced a tool that solves the opposite task: by human description inputted, it can output the corresponding term in a structured vocabulary of phenotypic abnormalities found in human disease.

Cao et al. (2020) created the dataset for style transfer to simplify medical texts. They scraped the Merck Manuals (MSD Manuals) to find aligned sentences and hired experts to select sentences from each ver-

sion of the group to annotate pairs of sentences that have the same meaning but are written in different styles. They also developed benchmarks.

After the release of ChatGPT in November 2022 by OpenAI, it has become a widely used tool for solving everyday tasks due to its excellent zero-shot and few-shot abilities in various domains. That's why many papers nowadays focus on analysing its potential in many fields including medicine. Gao et al. (2023); Guo et al. (2023) compared the output of human beings and ChatGPT against each other to know more about the accuracy and integrity of using these models in scientific writing. Guo et al. (2023) propose the HC3 (Human ChatGPT Comparison Corpus) dataset, which consists of nearly 40K questions and their corresponding human/ChatGPT answers. They also answered multiple questions about ChatGPT possibilities, limitations, and prompt engineering. The studies revealed that although the ChatGPT texts are well-written and without plagiarism, it is still can be distinguished from human-written ones.

Liao et al. (2023); Jeblick et al. (2023) conducted a comparative study of human- vs ChatGPT-generated medical texts. Liao et al. (2023) compare texts to uncover differences in vocabulary, part-of-speech, dependency, sentiment, perplexity, etc. They concluded that medical texts written by humans are more concrete, more diverse, and typically contain more useful information, while medical texts generated by ChatGPT pay more attention to fluency and logic, and usually express general terminologies rather than effective information specific to the context of the problem. They also created a BERT-based model that can effectively detect medical texts generated by ChatGPT, and the F1 exceeds 95%. In the exploratory study of Jeblick et al. (2023), the authors concluded that most participating radiologists agreed that the simplified reports were factually correct, complete, and not potentially harmful to the patient, indicating that ChatGPT is in principle able to simplify radiology reports. Nevertheless, they mention that instances of incorrect text passages and missing relevant medical information were identified in a considerable number of cases, which could lead patients to draw harmful conclusions.

Although ChatGPT may be very beneficial for many tasks (summarization, information extraction, code generation, writing stories, etc.) (The New York Times, 2023; Meghan Holohan, Today, 2023; Will Douglas Heaven, MIT Technology Review, 2023), it also can lead to unforeseen consequences (especially in a sensitive sphere like medicine) (Dan Milmo, The Guardian, 2023; Ken Foxe, Irish examiner, 2023; The White Hatter, 2023; JMIR Publications, Medical

Xpress, 2023).

Currently, the following metrics for automatic simplification evaluation are used, including SARI (Xu et al., 2016), FKGL (Flesch, 1948), BLEU (Papineni et al., 2002), Levenshtein distance (Levenshtein, 1966), Type-Token ratio (Johnson, 1944), textual lexical diversity (McCarthy, 2005), etc. Most of them were created for another purpose (machine translation, lexical richness of texts, readability), so not fully suitable for simplification evaluation. Martin et al. (2020) identify four attributes related to the process of text simplification. Namely, amount of compression, amount of paraphrasing, lexical complexity, and syntactic complexity. In addition, for simplification and especially for medical texts simplification, it is crucial that no important information is missing.

Given the great variety of automatic metrics for evaluation, there has also been considerable interest in evaluation based on manual evaluation. Commonly used protocol (Jiang et al., 2020; Narayan and Gardent, 2014) usually evaluates adequacy (is the meaning preserved?), fluency (is the simplification fluent?), and simplicity (is the simplification actually simpler?). Schwarzer (2018) claim that adequacy and simplicity are negatively correlated suggesting a common, underlying fact: removing material from a sentence will make it simpler while reducing its adequacy. Still, these criteria should be supplemented by more medical-sphere-specific information.

3 METHODOLOGY

3.1 Principles of Medical Text Simplification

We propose a refined protocol for medical simplification, which also includes three criteria (congruence, fluency, and simplicity). The first criterion convergence contains two components: 1) preserve the original information, 2) don't add extra information. As much information must be preserved as in this case almost every detail is important. This includes the medical test outcomes, dates related to medical history and treatment process, medication names and dosage, diseases (history and current state), doctor and hospital names, race and other body features, reference values, etc. There are however cases (for example, some inner body parts, medical devices) that cannot be simplified and retain the meaning fully at the same time, so there is always a balance between detailed and easy-to-understand explanations. It also is very important that no other information is added and a

machine learning model doesn't add its judgments or conclusions.

Secondly, textual fluency (readability) and correctness are the other important aspects of simplification. This metric is related to overall text quality (weakly related to the "medical" characteristic of a text). The questions we address in textual fluency are: 1) Are sentences grammatically correct?, 2) Are sentences relatively short? 3) Are sentences easy to follow? The latter question includes making sure that two related concepts come as close as possible in sentences and a text and using the right ordering within sentences. A good topic explanation may be found in a book written by Stafford and Webb (2010).

Thirdly, let us discuss the simplicity aspect. Based on our analysis, some principles may help to create a more easy-to-understand simplified medical text. It may be crucial for understanding and mistake avoidance to disclose abbreviations while keeping the original abbreviation (for example, in brackets), so that a patient may refer to it in them in the source text. However, some abbreviations are quite common and can be left as they are (for example, we can keep "CT" instead of writing "computed tomography"). Besides affecting patients' understanding badly, complex medical text can sometimes be hard to read for healthcare professionals due to lots of short ungrammatical sentences. Some medical abbreviations can in turn even threaten a patient's life (National Coordinating Council for Medication Error Reporting and Prevention, 2023). For example, "Q.D." (Latin abbreviation for every day), here the period after the "Q" has sometimes been mistaken for an "I," and the drug has been given "QID" (four times daily) rather than daily. In (Health Service Executive, Code of Practice for Healthcare Records Management, 2010) a list of agreed abbreviations and other recommendations on abbreviating is provided. Another thing that may improve simplicity is repeating new or rare terminology in multiple places in a text. It may also be beneficial to include the main purpose at a high level (or briefly explain how it works) for each medication. As well as for each medical test, try to include information on the reason why it was taken. Also, for each procedure or surgery, try to explain the steps during the surgery.

3.2 Dataset

For this work, we created a small, proof-of-concept dataset. The dataset consisting of 30 triples (around 800 sentences) of the original text, human and ChatGPT-simplified texts was created from the dataset of Vydiswaran (2019). The original dataset consists of medical notes, which come from exactly

one of the following five clinical domains: Gastroenterology, Neurology, Orthopedic, Radiology, Urology. There are 1239 texts in total in the original dataset.

The original texts were first preprocessed which included removing HTML tags, replacing multiple spaces with single spaces, and enumeration of each sentence. Simplified text was created out of complex text (Vydiswaran, 2019) under the previously outlined congruence, fluency, and simplicity principles by a non-native English speaker with no medical or healthcare professional background. New texts were aligned with the original, with each sentence with the number N in the original text corresponding to a sentence with the number N in the simplified instance. In some cases, one sentence in an original text can correspond to multiple sentences in simplified text (each of them is also numbered N). In this way, we would make sure that aligned pairs are created as the alignment is crucial for simplification tasks (Jiang et al., 2020).

Automatically created simplifications were obtained through the use of the OpenAI chat prompt where the following prompt was used: “Please simplify the text so that non-professionals could understand it”. ChatGPT tends to produce summarization rather than simplification on longer texts, so, for long texts (typically more than 20 sentences), the text was inputted by parts (with the following prompt after the main one within the same chat context: “Could you also simplify one more follow-up text so that non-professional could understand it: <NEXT PART OF THE COMPLEX TEXT>”). It was decided not to add any examples or guidance for clarity reasons.

3.3 Questionnaire

To provide a subjective evaluation, a survey has been conducted via Prolific (Prolific, 2014). Each respondent was required to be fluent in English and use a computer or tablet to take the survey. No other restrictions were posed. Forty-seven people participated in the evaluation (out of the total pool of around 120,000 preselected Prolific users).¹

The survey consisted of three sections. In the first two sections, full-text simplifications were compared against each other with questions intended to ascertain the easiness of getting the main idea, detailedness, text quality, and easiness of understanding. Each text was presented to the participants side by side. For clarity, each sentence in all of the three variants was numbered so that the relationships be-

¹Specific questions are available at the Github page provided in the abstract.

tween the sentences were clear. In some cases, one sentence in a text could correspond to multiple sentences in another text.

In the third section, standalone sentences from medical texts were evaluated. The participants were given the original sentence, some context (description of the procedure from which the sentence was taken and an illustrative picture where applicable), and two possible simplifications to choose from. The questions were aimed to measure whether new forms retained clarity, factual accuracy, easiness of grasping the context, bias, misinterpretation, etc.

The survey also gathered demographic information such as gender, age group, English language proficiency (native or non-native, bilingual, etc.), education, and information on whether participants belonged to the medical profession in some way (at a student or professional level).

4 RESULTS AND ANALYSIS

4.1 Automatic Metrics Analysis

The manually and automatically simplified texts are first compared via several text analytical metrics to get an overall idea of the texts’ differences. To get the results, aligned sentences were evaluated and an average score was obtained. The metrics used are the similarity score of PubMedBERT (Deka et al., 2022) between the original and both human and ChatGPT sentences, the average number of character and words, words frequency according to Zipf frequency without stop-words (pypi, 2023b), POS-tagging distribution using spaCy library (spaCy, 2023), words dependency in sentences distribution using spaCy library (spaCy, 2023), sentiment score distribution using NLTK Vader library (NLTK, 2023), lexical readability (the Flesch Reading Ease) (pypi, 2022), and lexical richness (type-token ratio) (pypi, 2023a).

Some overall textual characteristics were also obtained. Namely, total number of sentences, vocabulary variety (total number of unique lowercased words in all texts), stemmed vocabulary variety (total number of unique lowercased and stemmed words in all texts), perplexity score (Huggingface, 2023b) of the sentences using Microsoft’s BioGPT model (Huggingface, 2023a).

Tables 1 and 2 along with Figures 1-4 summarize the obtained results. In Table 1, we can see that human produces more similar simplifications to the original ones and uses more words and characters. Surprisingly, Zipf’s word frequency score results show that the Original text uses more frequently used words

on average. In terms of perplexity, from Table 1 and Figure 1, we can deduce that ChatGPT produces more “predicted” outputs, which is in line with the findings in the paper of Liao et al. (2023)². Also, based on lexical richness (terms to words ratio), the original text is more lexically rich, while ChatGPT and Human output are identical. As for lexical readability, ChatGPT’s text corresponds to the “Fairly Easy” category, while Human and Original texts fall into the “Standard” and “Fairly Difficult” categories respectively. From Table 2, we can deduce that the ChatGPT vocabulary variety is the smallest one, while the Original text and Human’s simplifications are more varied in vocabulary (which is again in line with the results of Liao et al. (2023)).

In Figure 3, it can be seen that ChatGPT tends to use more DT (determiner or article such as “which”, “the”, “this”, etc.), while less JJ (adjectives) and CD (cardinal digits). As for human texts, it uses more prepositions (IN) and fewer personal pronouns (“me”, “I”, “he”, etc.). The original texts contain more adjectives (JJ), NNP (proper noun, singular such as personal and organizational names), and cardinal digits (CD). However, there are relatively small percentages of DT (determiner) and IN (prepositions) in the original texts. Some results depicted in Figure 3 are similar to the results of Liao et al. (2023).

From Figure 2 we can deduce that the Original text has more punctuation (PUNKT), numeric modifier governing the case of the noun (NUMMOD, for example, “dollar”), while having fewer determiners (det) and auxiliary verbs (aux). Human texts tend to include more prep (prepositions that are used to change the meaning of an adjective, verb, or noun, such as “up” in “get up”) and more pobj (object of a preposition). As for ChatGPT texts, they tend to have more determiners, nsubj (syntactic subject of a clause), dobj (accusative object of the verb), advcl (clause modifying the verb, for example, conditional or temporal clause), and poss (possession modifier, for example, “my” or “mother’s”). However, ChatGPT’s texts contain fewer compounds (multiple words that represent one morphosyntactical unit, for example, “adventure time”) and amod (an adjective that changes the meaning, for example, “blue” in “blue car”). Overall, it can be deduced that ChatGPT is producing more argumentative sentences, with more explicit connections within sentences.

²Human simplifications were created by non-native English speaker, which may also affect perplexity score.

Table 1: Comparison of Human and ChatGPT on a sentence level, averaged.

Metrics	Original	Human	ChatGPT
Similarity score	1.0	0.82	0.73
Number of characters	69	113	79
Number of words	12	21	15
Words Zipf frequency	4.3	4.7	4.9
Average perplexity	218	140	102
Lexical richness	0.25	0.17	0.17
Lexical readability	52.26	63.8	75.4

Table 2: Comparison of Human and ChatGPT overall characteristics for all texts.

Metrics	Original	Human	ChatGPT
Number of sentences	800	914	798
Vocab variety	2151	2543	1820
Stemmed vocab variety	1898	2065	1527

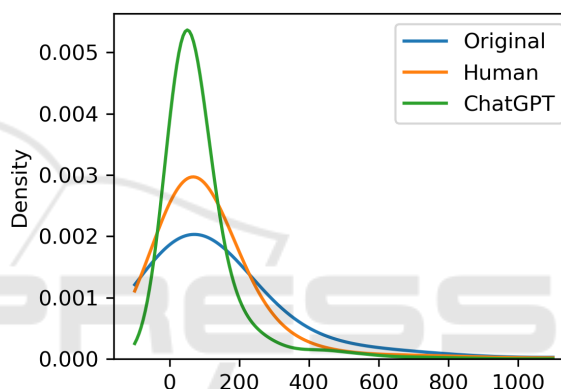


Figure 1: Kde distribution of perplexity comparison between original (complex text) and human- and ChatGPT-simplified text.

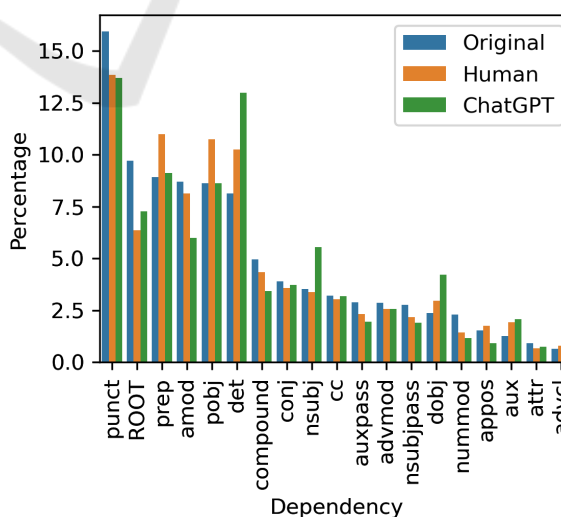


Figure 2: Words dependency distribution comparison between original (complex text) and human- and ChatGPT-simplified text.

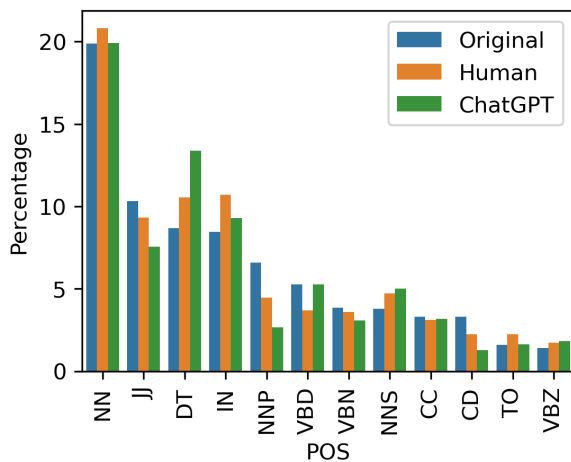


Figure 3: POS tagging distribution comparison between original (complex text) and human- and ChatGPT-simplified text.

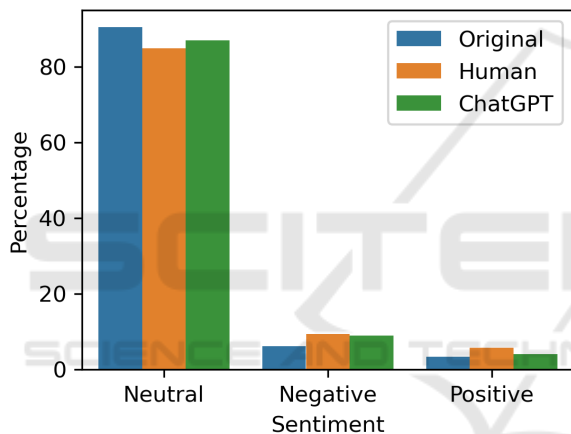


Figure 4: Sentiment distribution comparison between original (complex text) and human- and ChatGPT-simplified text.

4.2 Manual Evaluation

During the manual evaluation, some ChatGPT features based on their outputs compared to human ones were found.³ Firstly, let's consider some positive features.

1. ChatGPT can disclose abbreviations depending on the context.
2. ChatGPT has a very good rewriting ability (this is related to both general language skills and the ability to understand and simplify medical terms).

Some problems were also found in ChatGPT's medical text simplification.

³Specific examples are available at the Github page provided in the abstract.

1. ChatGPT tends to produce abstracts or summaries rather than simplification on long texts. This may be explained by the limited length of the context.
2. ChatGPT sometimes makes up some facts, which may be very dangerous in such a sensitive field as medicine. ChatGPT may even contradict its output.
3. ChatGPT somehow lacks commonsense reasoning or medical "knowledge".
4. ChatGPT may omit important facts or oversimplify. As was mentioned in the Congruence principle, it is very important to retain details of a human body or medical history for making a diagnosis.
5. ChatGPT is biased towards rewriting a text by any means, even if it has been already quite simple. Sometimes the rewriting may change the meaning. ChatGPT also tends to produce more personal sentences.
6. ChatGPT sometimes uses words such as "a", "about", "some", "called", rather than properly simplify a concept or explain. It also frequently outputs undersimplifications.

4.3 Questionnaire Results Analysis

For the subjective survey, there were almost equal numbers of female and male participants, around 80% of them are under 35, around two-thirds of them have native-equivalent English language proficiency, and more than 70% have at least an Undergraduate degree (bachelor, associate). Only five respondents are students in the medical sphere. Only two people consider themselves medical professionals (both are medical students).

The respondents were paid £9,21 per hour (average recommended value by the platform). On average, it took a survey participant around twenty minutes to complete the survey.

Figures 5-8 represent the averaged results of the first and second sections of the survey where participants were given three texts per section (Original, Human- and ChatGPT-simplified) to compare against sentence-by-sentence. The results in Sections 1 and 2 are mostly similar. However, there were remarkable variances in Human- and ChatGPT-generated text's detailedness evaluation. It was found that respondents likely consider the longest text to be the most detailed. The reason behind it may be the deduction that the longer the text, the more details it should have. There were three non-mutually exclusive options to assess

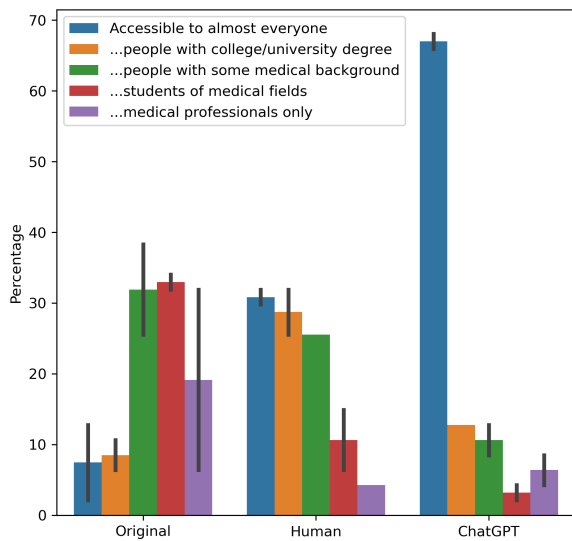


Figure 5: Participants' answers summary on the question of "Please evaluate the three texts according to their easiness of getting the general idea?".

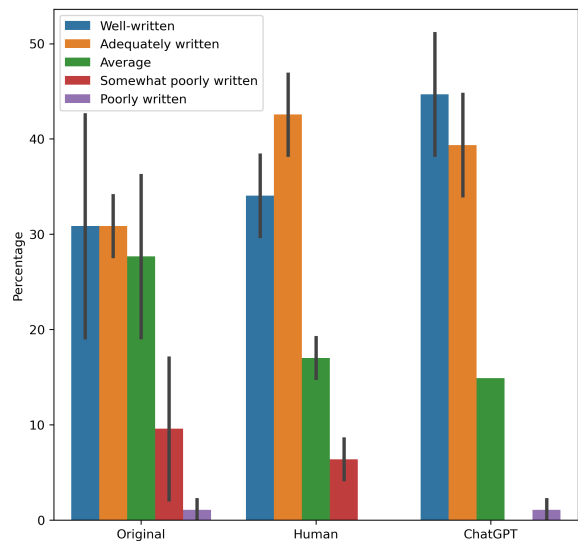


Figure 7: Participants' answers summary on the task of "Please evaluate the three texts according to their language fluency (how well are they written?)".

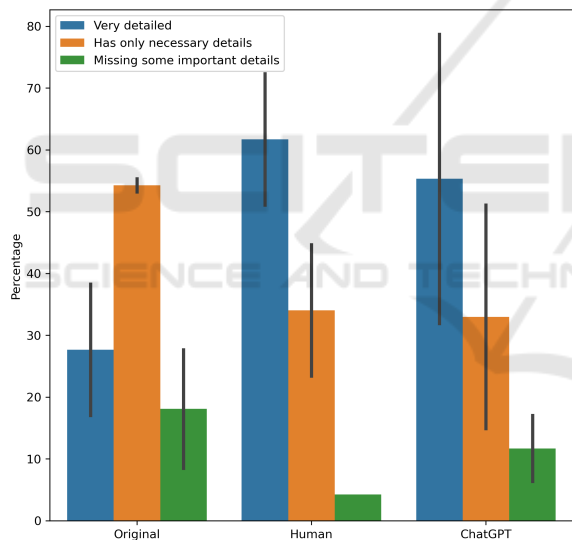


Figure 6: Participants' answers summary on the task of "Please evaluate the three texts according to the number of details provided".

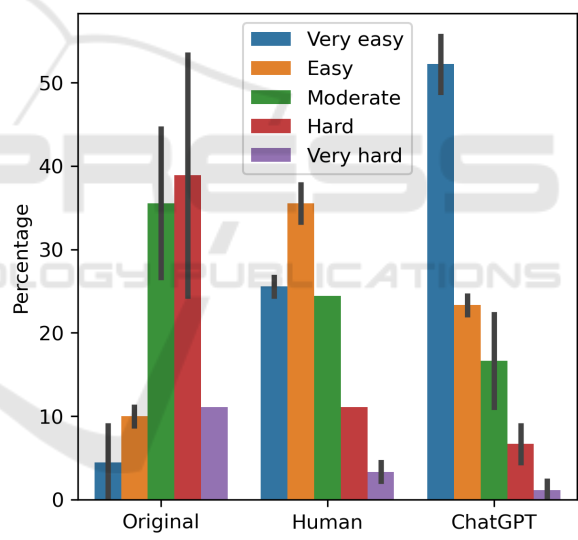


Figure 8: Participants' answers summary on the task of "Please evaluate the simplicity of these three texts (how easy is it to understand them?)".

each of the three texts (original, human-written simplification, and ChatGPT's output). Texts were presented side-by-side. In the first section, the length distribution of the texts was 421 for the Original text, 591 for human simplifications, and 512 for ChatGPT. It resulted in more than 70% of people thinking that human-written text was "Very detailed", around 30% considered ChatGPT's text to be very detailed, and only 20% classified the original text in the "Very detailed" category. As for the second section with the same setting but the other simplification triplets, the

length distribution of the texts was 738 for the Original text, 955 for human simplifications, and 940 for ChatGPT's. In this case, around 50% of people considered human text to be "Very detailed", more than 80% decided that ChatGPT's output is very detailed, and less than 40% classified original text to be very detailed. However, no new details have been added in either the human or ChatGPT texts (it was the other way around as simplifications tend to omit some details for the sake of more easily understood text). So, in terms of retaining the information, the original texts can be considered to be the most detailed ones, though

they are shorter in terms of length as they contain more descriptive terms. Unfortunately, the number of people with medical backgrounds is not enough to test for any difference between the answers of professionals' and laymen's answers.

Table 3: Section 3 survey results. Here the percentages in favor of human- or ChatGPT-produced texts are presented.

Question	Human	ChatGPT
Which option retains the main idea or meaning of the text?	51%	49%
Which option is more clear and easy to understand?	70%	30%
Which option maintains the factual accuracy of the original information?	70%	30%
Which option is better at using relatable comparisons or examples to help the audience grasp the concept more easily?	47%	53%
Which option better maintains an appropriate level of complexity, avoiding the loss of essential nuances?	75%	25%
Which simplification is better at maintaining the spirit and purpose of the original content, while making it more accessible?	77%	23%
Which option is more well-organized, well-structured, and easy to navigate?	85%	15%
Which option is more free from unnecessary details or information that doesn't contribute to the understanding of the main message?	26%	74%
Which simplification piques the audience's interest and encourages them to explore the topic further?	34%	66%
Which option is more unambiguous and straightforward?	10%	90%
Which simplification is more free from bias or misrepresentation?	70%	30%

Let's now discuss the third section where human and ChatGPT texts are compared against each other by various characteristics. Here the options were randomly shuffled for the respondents and there was no information related to the source of the text (whether it is specialist-, human- or machine-produced). For each question where appropriate, the respondents were given a pictorial or textual context, so that it

would be easier for them to understand medical texts from a question. The summarizing results are depicted in Table 3.

Overall, the results of the conducted survey suggest that people consider to be clear and easy to understand (and consider them to have an appropriate level of complexity) those simplifications that explain the process well, though may be quite long. The results also suggest that people are not always able to detect untruthful information in the simplifications. Another finding was that people are less interested in medical conditions' explanations and exact definitions. We also found that ChatGPT produces texts that people consider to be easily understood by many people.

5 DISCUSSION

During the evaluation of ChatGPT outputs against human simplifications, it was found that ChatGPT tends to produce more "average" (in terms of perplexity) and be more argumentative (it has more determiners according to POS-tagging and words dependency distributions) texts. Although in terms of language fluency, ChatGPT produces very good texts and can successfully disclose abbreviations depending on the context, it may make up some facts, lack common-sense reasoning (or medical "knowledge"), omit important facts, or oversimplify, etc. According to the survey results, we found that people sometimes cannot distinguish untruthful information in the simplifications, which may be dangerous. Another finding was that people are less interested in medical condition explanations and exact definitions in simplified texts even though they more accurately correspond to the original text. We also found that ChatGPT's simplifications are considered to be accessible to a large percentage of people.

6 CONCLUSION

We hope that our paper and dataset will help to bridge the gap between medical professionals and patients' vision. We believe that AI tools would be used more concisely in the medical sphere, because of the problems associated with omitting important information, made-up facts, oversimplification, etc. Bearing in mind the features of current SOTA LLMs, we can make a safer model for the medical field.

7 FUTURE WORK

Multiple things have been found that are worth further investigation. Firstly, some of ChatGPT's simplifications weren't found on the web in English (by keywords), so it would be interesting how the model utilizes the multilingual data it has been trained on. Is it implicitly translating the simplifications from the other languages?

Another thing we faced during the writing of this paper is that it is hard to decide which term should be simplified and which one shouldn't. For example, should we keep "placenta" word? Or maybe should we simplify it to "afterbirth"? Or is it better to explain that term?

Speaking about which terms should be simplified, it is obvious that it heavily depends on the target audience. It would be beneficial to try other prompts or techniques for ChatGPT that would be better designed for a particular group ("Simplify this text for a fifteen years old non-native English speaker. Here you will see some examples of a good simplification..."). So, chain-of-thought, explicit role statement (Salewski et al., 2023), psychological manipulations, in-context learning, self-consistency verification (Wang et al., 2023), etc. techniques may be used.

We should also take into account that our respondents from Prolific are educated enough to use this platform, so, our results weren't evaluated on illiterate people or people with poor (health) literacy. In future studies, it would be beneficial to take this group of people in account.

Lastly, as new text generative models are being released on an almost everyday basis, it would also be worth looking into the other models other than ChatGPT.

ACKNOWLEDGEMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183 and the ADAPT SFI Research Centre for AI-Driven Digital Content Technology under Grant No. 13/RC/2106_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

- Ariyaratne, S., Iyengar, K. P., Nischal, N., Chitti Babu, N., and Botchu, R. (2023). A comparison of chatgpt-generated articles with human-written articles. *Skeletal Radiology*, 52(9):1755–1758.
- Basu, C., Vasu, R., Yasunaga, M., and Yang, Q. (2023). Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.
- Borkar, J. (2023). What can we learn from data leakage and unlearning for law?
- Cao, Y., Shui, R., Pan, L., Kan, M.-Y., Liu, Z., and Chua, T.-S. (2020). Expertise style transfer: A new task towards better communication between experts and laymen. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.
- Chowdhery, A. and et al., S. N. (2022). Palm: Scaling language modeling with pathways.
- Dan Milmo, The Guardian (2023). Mushroom pickers urged to avoid foraging books on amazon that appear to be written by ai. <https://www.theguardian.com/technology/2023/sep/01/mushroom-pickers-urged-to-avoid-foraging-books-on-amazon-that-appear-to-be-written-by-ai>. Retrieved on November 7, 2023.
- Deka, P., Jurek-Loughrey, A., and P. D. (2022). Evidence extraction to validate medical claims in fake news detection. *Health Information Science*, page 3–15.
- et al., B. W. (2023). Bloom: A 176b-parameter open-access multilingual language model.
- Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., and Pearson, A. T. (2023). Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers. *npj Digital Medicine*, 6(1).
- Gooding, S. (2022). On the ethical considerations of text simplification. In Ebling, S., Prud'hommeaux, E., and Vaidyanathan, P., editors, *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.
- Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., and Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection.
- Health Service Executive, Code of Practice for Healthcare Records Management (2010). Abbreviations. <https://www.hse.ie/eng/about/who/qid/quality->

- and-patient-safety-documents/abbreviations.pdf. Retrieved on November 19, 2023.
- Huggingface (2023a). Biogpt. <https://huggingface.co/microsoft/biogpt>. Retrieved on November 7, 2023.
- Huggingface (2023b). Metric: perplexity. <https://huggingface.co/spaces/evaluate-metric/perplexity>. Retrieved on November 7, 2023.
- Jeblick, K., Schachtner, B., Dextl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B. O., Ricke, J., and et al. (2023). Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *European Radiology*.
- Jiang, C., Maddela, M., Lan, W., Zhong, Y., and Xu, W. (2020). Neural CRF model for sentence alignment in text simplification. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.
- JMIR Publications, Medical Xpress (2023). Chatgpt generates 'convincing' fake scientific article. <https://medicalxpress.com/news/2023-07-chatgpt-generates-convincing-fake-scientific.html>. Retrieved on November 7, 2023.
- Johnson, W. (1944). Studies in language behavior: A program of research. *Psychological Monographs*, 56(2):1–15.
- Kandula, S., Curtis, D., and Zeng-Treitler, Q. (2010). A semantic and syntactic text simplification tool for health content. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2010:366–70.
- Ken Foxe, Irish examiner (2023). Ucc staff told it would be almost impossible to detect students cheating with chatgpt. <https://www.irishexaminer.com/news/munster/arid-41135368.html>. Retrieved on November 7, 2023.
- Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.
- Li, Y., Bubeck, S., Eldan, R., Giorno, A. D., Gunasekar, S., and Lee, Y. T. (2023). Textbooks are all you need ii: phi-1.5 technical report.
- Liao, W., Liu, Z., Dai, H., Xu, S., Wu, Z., Zhang, Y., Huang, X., Zhu, D., Cai, H., Liu, T., and Li, X. (2023). Differentiate chatgpt-generated and human-written medical texts.
- Luo, J., Lin, J., Lin, C., Xiao, C., Gui, X., and Ma, F. (2022). Benchmarking automated clinical language simplification: Dataset, algorithm, and evaluation. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3550–3562, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Luo, Y.-F., Henry, S., Wang, Y., Shen, F., Uzuner, O., and Rumshisky, A. (2020). The 2019 national natural language processing (nlp) clinical challenges (n2c2)/open health nlp (ohnlp) shared task on clinical concept normalization for clinical records. *Journal of the American Medical Informatics Association*, 27(10).
- Manakul, P., Liusie, A., and Gales, M. J. F. (2023). Self-checkgpt: Zero-resource black-box hallucination detection for generative large language models.
- Manzini, E., Garrido-Aguirre, J., Fonollosa, J., and Perera-Lluna, A. (2022). Mapping layperson medical terminology into the human phenotype ontology using neural machine translation models. *Expert Systems with Applications*, 204:117446.
- Martin, L., de la Clergerie, É., Sagot, B., and Bordes, A. (2020). Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.
- McCarthy, P. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. PhD thesis, University of Memphis.
- Meghan Holohan, Today (2023). A boy saw 17 doctors over 3 years for chronic pain. <https://www.today.com/health/mom-chatgpt-diagnosis-pain-rcna101843>. Retrieved on November 7, 2023.
- Narayan, S. and Gardent, C. (2014). Hybrid simplification using deep semantics and machine translation. In Toutanova, K. and Wu, H., editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.
- Nascimento, N., Alencar, P., and Cowan, D. (2023). Comparing software developers with chatgpt: An empirical investigation.
- National Coordinating Council for Medication Error Reporting and Prevention (2023). Dangerous abbreviations. <https://www.nccmerp.org/dangerous-abbreviations>. Retrieved on November 7, 2023.
- NLTK (2023). Vader. https://www.nltk.org/_modules/nltk/sentiment/vader.html. Retrieved on November 7, 2023.
- OpenAI (2023). Chatgpt. <https://openai.com/chatgpt>. Retrieved on November 7, 2023.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Prolific (2014). Prolific. <https://www.prolific.com/>. Retrieved on November 20, 2023.
- pypi (2022). textstat. <https://pypi.org/project/textstat/>. Retrieved on November 7, 2023.
- pypi (2023a). Lexicalrichness. <https://pypi.org/project/lexicalrichness/>. Retrieved on November 7, 2023.
- pypi (2023b). wordfreq. <https://pypi.org/project/wordfreq/>. Retrieved on November 7, 2023.

- Rotegard, A., Slaughter, L., and Ruland, C. (2006). Mapping nurses' natural language to oncology patients' symptom expressions. *Studies in health technology and informatics*, 122, 987-8.
- Sakakini, T. and Lee, J. Y. e. a. (2020). Context-aware automatic text simplification of health materials in low-resource domains. In Holderness, E., Jimeno Yepes, A., Lavelli, A., Minard, A.-L., Pustejovsky, J., and Rinaldi, F., editors, *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 115–126, Online. Association for Computational Linguistics.
- Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., and Akata, Z. (2023). In-context impersonation reveals large language models' strengths and biases.
- Schwarzer, M. (2018). Human evaluation for text simplification : The simplicity-adequacy tradeoff.
- Slaughter, L., K. A. K. A. . P. V. L. (2005). A framework for capturing the interactions between laypersons' understanding of disease, information gathering behaviors, and actions taken during an epidemic. *Journal of biomedical informatics*, 38(4), 298–313. <https://doi.org/10.1016/j.jbi.2004.12.006>.
- spaCy (2023). Industrial-strength natural language processing. <https://spacy.io/>. Retrieved on November 7, 2023.
- Stafford, T. and Webb, M. (2010). *Mind hacks*. O'Reilly Media.
- The New York Times (2023). When doctors use a chatbot to improve their bedside manner. <https://www.nytimes.com/2023/06/12/health/doctors-chatgpt-artificial-intelligence.html>. Retrieved on November 7, 2023.
- The White Hatter (2023). Scammed by chatgpt! darkside of ai. <https://thewhitehatter.ca/news-show/scammed-by-chatgpt-darkside-of-ai/>. Retrieved on November 7, 2023.
- Touvron, H. and et al., L. M. (2023). Llama 2: Open foundation and fine-tuned chat models.
- Trienes, J., Schlötterer, J., Schildhaus, H.-U., and Seifert, C. (2022). Patient-friendly clinical notes: Towards a new text simplification dataset. In Štajner, S., Saggion, H., Ferrés, D., Shardlow, M., Sheang, K. C., North, K., Zampieri, M., and Xu, W., editors, *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.
- Van, H., Kauchak, D., and Leroy, G. (2020). AutoMeTS: The autocomplete for medical text simplification. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1424–1434, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Vydiswaran, V. (2019). Medical notes classification.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models.
- Will Douglas Heaven, MIT Technology Review (2023). Chatgpt is going to change education, not destroy it. <https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/>. Retrieved on November 7, 2023.
- Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.