Bridging the Gap in Health Literacy: Harnessing the Power of Large Language Models to Generate Plain Language Summaries from Biomedical Texts

Anonymous ACL submission

Abstract

Health literacy is essential for individuals to navigate the healthcare system and make informed decisions about their health. Low health literacy levels have been associated with negative health outcomes, particularly among older populations, those who are financially restricted or with lower educational attainment. Plain language summaries (PLS) are an effective tool to bridge the gap in health literacy by simplifying content found in biomedical and clinical documents, in turn, allowing the general audience to better understand health-related documentation. However, manually translating biomedical texts to PLS and guaranteeing they can be understood by a lay audience is a time-consuming and challenging task. This study assessed the performance of Natural Language Processing (NLP) for classifying if a biomedical text is written in plain language, and Large Language Models (LLMs), Generative Pre-trained Transformer (GPT) 3.5 and GPT 4, for automating the generation of PLS from technical biomedical texts. The classification model achieved high precision (97.2%) in identifying if a text is written in plain language. GPT 4, a state-of-the-art LLM, successfully generated PLS that were semantically equivalent to those generated by domain experts and which were rated high in accuracy, readability, completeness, and usefulness. Our findings demonstrate the value of using LLMs and NLP to translate biomedical texts into plain language summaries, and their potential to be used as a supporting tool for healthcare stakeholders to empower patients and the general audience to understand healthcare information and make informed healthcare decisions.

1 Introduction

Health literacy refers to an individual's capacity to access, understand, and use health information (Nielsen-Bohlman et al., 2004). It allows patients and their families to navigate healthcare systems, comprehend and act upon a diagnosis or medical instruction, adhere to medication regimens, and make informed decisions, otherwise considered daunting, regarding participation in clinical trials, treatment options, or medical procedures (Berkman et al., 2011a,b; Miller, 2016). Low health literacy levels have been consistently associated with higher mortality rates, increased instances of preventable hospitalizations, and poor treatment adherence (Berkman et al., 2011a). Paradoxically, while health literacy is crucial for positive health outcomes, the 2015 European Health Literacy Survey revealed that almost half of the respondents had inadequate health literacy, particularly among older populations, those who are financially restricted, or who have lower educational attainment (Sørensen et al., 2015; Bahador et al., 2020).

With the growing expectation for individuals to participate in healthcare decisions, enhancing health literacy becomes a significant attribute in improving public health and reducing health disparities (Nielsen-Bohlman et al., 2004; Stormacq et al., 2019; Schillinger, 2021). Improving health literacy in the population extends beyond actions taken to increase individual health literacy levels. In line with the General Data Protection Regulation (GDPR) principle of transparency, stakeholders such as healthcare providers, policymakers, and pharmaceutical companies should strategize to improve their organizational health literacy (OHL) by ensuring the clarity and comprehensibility of health documentation (GDPR, 2023; Trezona et al., 2018).

One strategy is simplifying clinical and scientific research language into lay-friendly summaries, known as plain language summaries (PLS). Some different techniques and guidelines can be used to translate complex scientific and biomedical concepts into PLS, for example, eliminating the use of technical jargon, replacing passive voice with active, or using short sentences and paragraphs (Bahador et al., 2020; Centers for Disease Control and Prevention, 2022). However, authoring a PLS can be time-consuming and challenging, particularly in areas like clinical settings which typically involve documents with technical and domain-specific vocabulary.

With the advancement of technology, new methods have been developed to automate the simplification of biomedical texts. In 2022, a review by Oldov et al. analyzed 32 tools or methods using either a rule-based approach or Natural Language Processing (NLP) and concluded that NLP methods offer more promising outputs but were limited by the scarcity of training data, resulting in continued reliance on rule-based methods (Ondov et al., 2022). Large Language Models (LLMs) with their immense data training potential and text generation capabilities, present a promising solution to tackle this challenge and automate the generation of PLS from technical documents.

Intending to bridge the gap in health literacy by facilitating the translation of biomedical texts to comprehensible summaries designed for patients, our study demonstrates the potential of NLP to develop a classification system to identify if a text is written in plain language, and LLMs to automate the generation of accurate, complete, and comprehensible PLS.

2 Materials and Methods

Our methodology, outlined in Figure 1, consisted of 3 main steps: 1) collecting and processing of sample texts in technical and plain language, 2) conducting a quantitative analysis of the plain and technical texts to generate a plain language classification model and a qualitative analysis of the texts to generate the prompts for the LLMs, and 3) assessing the use of the LLMs to generate PLS from technical texts.

2.1 Data Collection and Processing

We collected biomedical texts, both in technical and plain language (see the data sources in Table A1), and assembled them into a dataset of 14,441 texts. This “main dataset” was then divided into training and testing sets, consisting of 4,596 plain and 6,721 technical texts for training, and 1,149 plain and 1,975 technical texts for testing.

We enlarged each dataset by treating each paragraph of a minimum of 250 words as a distinct unit while excluding texts with fewer than 250 words. As a result, our "augmented dataset" had 61,354 texts, divided into 16,731 plain and 31,740 technical for training, and 5,090 plain and 7,793 technical for testing.

2.2 Analysis of Plain Language

We conducted qualitative and quantitative analyses of the texts to identify unique linguistic traits and variables that classify a text as plain language.

2.2.1 Qualitative Analysis

Driven by the varying and broad-scope guidance on creating high-quality PLS (Stoll et al., 2022), we analyzed a subset of our plain texts and created a ’criteria checklist’ (see Table 1) with the linguistic attributes most commonly present in plain texts. Key resources used in this process were guides and reviews, such as Your Guide to CLEAR WRITING by CDC (Centers for Disease Control and Prevention, 2022), Federal Plain Language Guidelines (The Plain Language Action and Information Network, 2011), Health Literacy Universal Precautions Toolkit by Agency for Healthcare Research and Quality (AHRQ) (Brach, 2023), Just Plain Clear Glossary by United Health Group (United Health Group, 2023), EU 536/2014 Summary of Clinical Results for Laypersons (European Union, 2023), and results presented by Stoll et al, in their systematic review of theory, guidelines, and empirical research on PLS (Stoll et al., 2022). We used the resultant checklist to complement the qualitative findings described in the next section and aid in developing the prompt detailed in the section LLM Prompt for Plain Language Summary Generation.

2.2.2 Quantitative Analysis

We computed readability metrics and language variables for each text in the augmented dataset using the Readability Library (Crummy, 2023) and SpaCy (SpaCy, 2023), respectively. This resulted in 64 variables presenting each text’s readability and linguistic traits (see Table A2).

We analyzed the language variables in our dataset to identify their potential to classify a text as technical or plain. We used a statistical hypothesis test for each of the variables of the *main dataset*. For each variable, we created a random sample of size n from the plain texts $(X_1, X_2, \cdots, X_n \sim P_X)$ and a random sample of size n from the technical texts $(Y_1, Y_2, \cdots, Y_n \sim Q_Y)$, and tested if our data supported either of the following hypotheses:

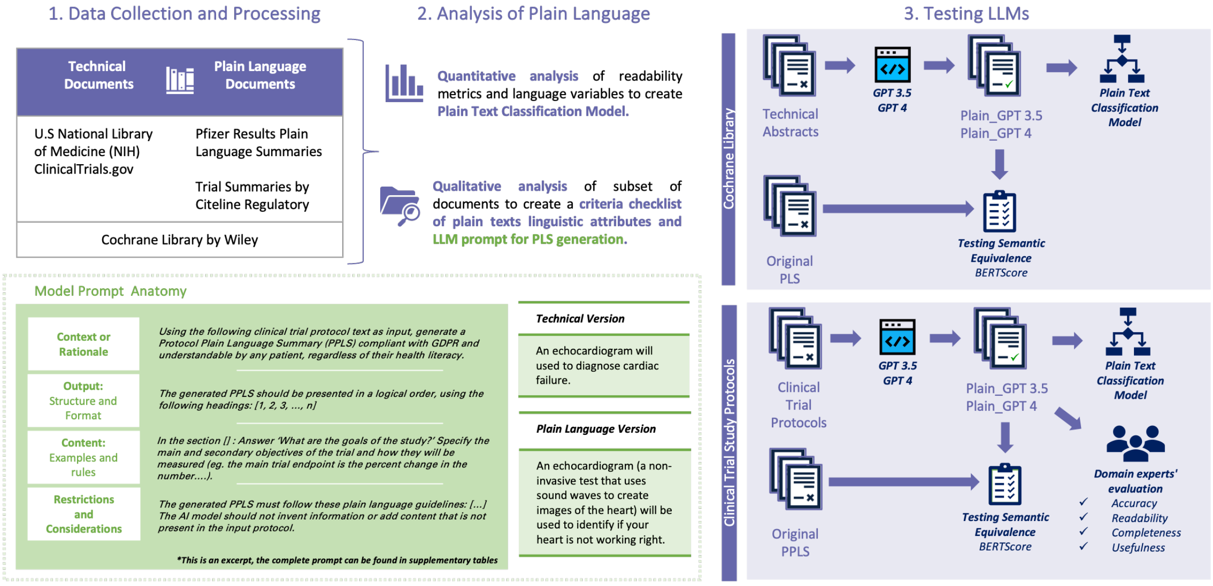

1. Data Collection and Processing
Technical Documents
Plain Language Documents
U.S National Library of Medicine (NIH) ClinicalTrials.gov
Pfizer Results Plain Language Summaries
Trial Summaries by Citeline Regulatory
Cochrane Library by Wiley
2. Analysis of Plain Language
Quantitative analysis of readability metrics and language variables to create Plain Text Classification Model.
Qualitative analysis of subset of documents to create a criteria checklist of plain texts linguistic attributes and LLM prompt for PLS generation.
Model Prompt Anatomy
Context or Rationale
Using the following clinical trial protocol text as input, generate a Protocol Plain Language Summary (PPLS) compliant with GDPR and understandable by any patient, regardless of their health literacy.
Output: Structure and Format
The generated PPLS should be presented in a logical order, using the following headings: [1, 2, 3, ..., n]
Content: Examples and rules
In the section [] : Answer 'What are the goals of the study?' Specify the main and secondary objectives of the trial and how they will be measured (eg. the main trial endpoint is the percent change in the number....).
Restrictions and Considerations
The generated PPLS must follow these plain language guidelines: [...] The AI model should not invent information or add content that is not present in the input protocol.
*This is an excerpt, the complete prompt can be found in supplementary tables
Technical Version
An echocardiogram will used to diagnose cardiac failure.
Plain Language Version
An echocardiogram (a non-invasive test that uses sound waves to create images of the heart) will be used to identify if your heart is not working right.
3. Testing LLMs
Cochrane Library
Technical Abstracts
GPT 3.5
GPT 4
Plain_GPT 3.5
Plain_GPT 4
Plain Text Classification Model
Original PLS
Testing Semantic Equivalence
BERTScore
Clinical Trial Study Protocols
Clinical Trial Protocols
GPT 3.5
GPT 4
Plain_GPT 3.5
Plain_GPT 4
Plain Text Classification Model
Domain experts' evaluation
✓ Accuracy
✓ Readability
✓ Completeness
✓ Usefulness
Original PPLS
Testing Semantic Equivalence
BERTScore

Figure 1: **Methodology.** Our methodology involved three steps: 1) collection and processing of biomedical texts (technical documents and plain language documents) into datasets for training and testing, 2) quantitative analysis of the texts to create a plain language classification model, and qualitative analysis to identify linguistic traits in plain texts to guide the engineering of a prompt that could translate biomedical text into Plain Language Summaries (PLS) using Language Learning Models (LLMs; and 3) testing the effectiveness of the LLMs in generating PLS quantitatively with our classification model and with semantic equivalence (BERTScore) and qualitatively with domain experts' evaluation.

- *Null Hypothesis*, $H_0 : P = Q$, the distributions of the proportion of the variable of interest for both samples (text and technical) are the same.

- *Alternative Hypothesis*, $H_1 : P \neq Q$, the distributions of the proportion of the variable of interest for both samples (text and technical) are different.

We evaluated the null hypothesis by comparing our 2 distributions using non-parametric tests: Wilcoxon, Kolmogorov-Smirnov (KS), and Mann–Whitney U. Given the multiple hypothesis tests, one for each variable, we adjusted the significance levels to control the probability of Type I errors by using the Bonferroni correction to lower the alpha value by dividing the desired significance level $\alpha = 0.05$ by the total number of tests $m = 64$ which gives a new significance level $\alpha' \approx 0.0008$.

Figure 2 illustrates examples of the comparison of the distributions of some of the variables in technical and plain texts. Out of the 64 variables, only 'Interjections' and 'Passive Voice' did not provide sufficient evidence to reject the null hypothesis (p-value > 0.0008). The other 62 variables were significantly distinct between the types of text and were included in our classification model.

2.3 Plain Texts Classification Model

We used the *augmented dataset* - train and the 62 distinct variables between text types (Section Quantitative Analysis), to build the classification model. We used Gradient Boosting (GB) and Random Forest (RF) machine learning models.

2.4 LLM Prompt for Plain Language Summary Generation

Our objective was to design a prompt for LLMs capable of translating biomedical technical documents into PLS.

Beginning with a clinical trial protocol from ClinicalTrials.Gov (see data sources in Table A1), we used a simple initial prompt: *'Using the following clinical trial protocol text as input, create a plain language summary'*. We tested this prompt using both GPT3.5 and GPT4, analyzed the generated output, and iteratively refined the prompt by adding details and instructions.

We aimed to produce a PLS that met the following qualitative criteria: **(1) Accuracy:** The content is clinically and scientifically accurate. **(2) Readability**: the content is comprehensible by a layperson, as defined by the plain language criteria checklist (Table 1). **(3) Completeness:** The content adheres to the expectations of a Protocol Plain

Linguistic Attributes	PLS Characteristics
• Use simple and everyday words. Avoid technical, medical, or scientific terms, jargon, or complex terminology (e.g., explain technical terms such as copayment, electrocardiogram, pyrexia, screening, double-blind). • Readability level 6 or below • Active voice over passive voice • Mostly 1-2 syllable words • Sentences of less than 20 words • Short paragraphs of 3-5 sentences • Simple numbers that do not require any math (e.g., 4 out of every 10 community members, not 40% of community members)	• Approximate length of 700-900 words • Specific structure and content by domain (e.g., EU-CTR suggested a specific structure and content for lay protocol synopsis)

Table 1: PLS Criteria Checklist of linguistic attributes and characteristics as defined by qualitative analysis of sample texts and Plain Language guidelines frequently used by domain experts.

Language Summary (PPLS) as specified by EU CTR No 536/2014 (United Health Group, 2023). **(4) Usefulness:** The generated PLS can be used as a first version to draft the study PPLS.

Our final prompt, provided in Figure B1, was designed specifically to generate a PLS of a clinical trial protocol. It includes the following elements:

- **Context:** a clear rationale on why a PLS is needed for the given clinical trial protocol.
- **Output:** the desired structure and format for the generated summary, including the specific sections of the output.
- **Content:** the expected content within each section, with examples and rules to guide the generation process.
- **Restrictions:** limitations of the output (e.g., word count limitations, the inclusion of only the information provided in the original protocol, and adherence to the criteria checklist for plain language as set out in Table 1).

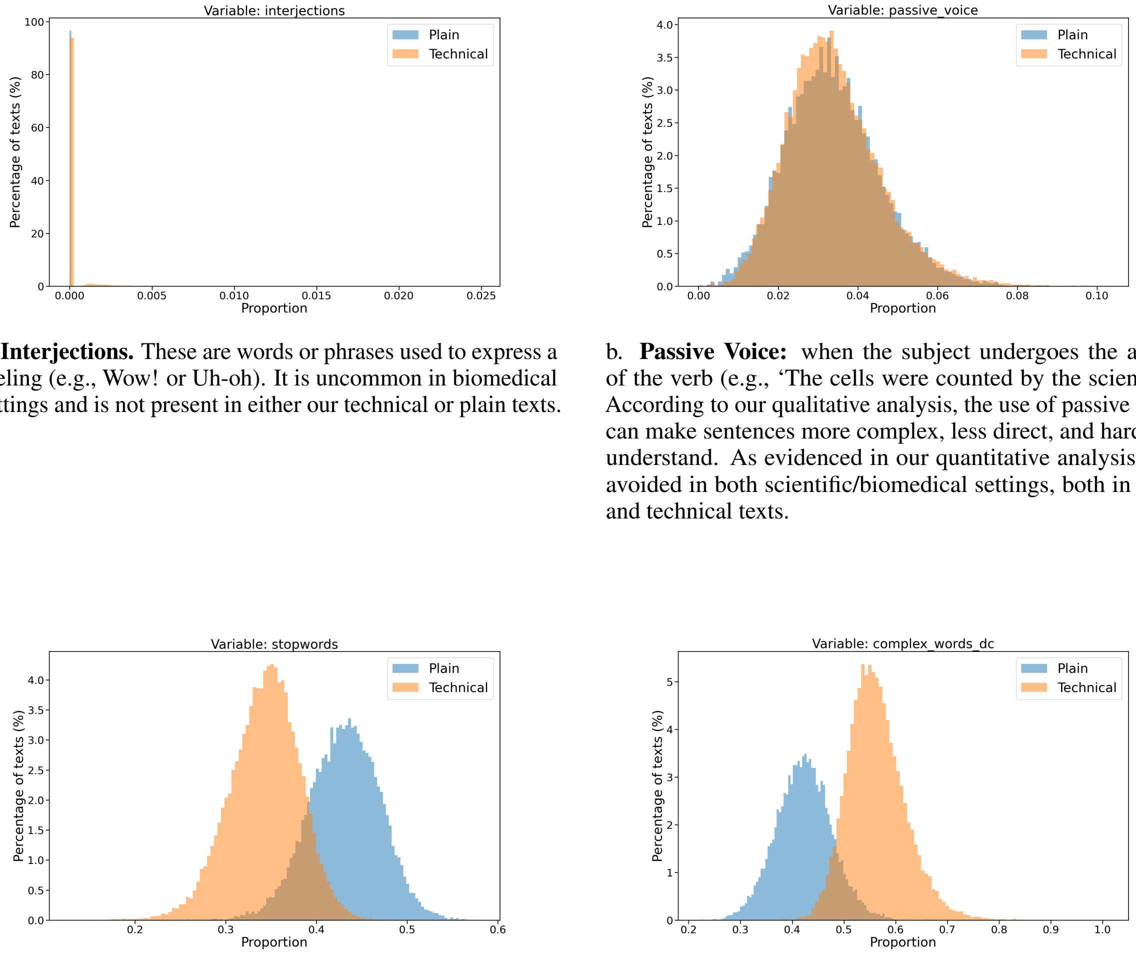
After finalizing the prompt for generating a PPLS, we used the same approach to create a prompt to generate Cochrane Reviews PLS (see the description of this data source in Table A1, and the prompt in Figure B2).

We used our prompts with GPT 3.5 and GPT 4 to translate technical biomedical texts, Cochrane Reviews and Study Protocols, into their respective PLS: Cochrane PLS and Protocol PLS. We quantitatively tested the generated PLS for plainness and semantic equivalence. For PPLS, we also performed a qualitative assessment of the outputs by three experts in Clinical Trial Operations and Regulatory Medical Writing, who rated each GPT 3.5 and GPT 4 text on a 5-point Likert Scale (1-Strongly Disagree to 5-Strongly Agree). They evaluated the texts for accuracy, readability, completeness, and usefulness as defined in the Section 3.2.

3 Results

3.1 Plain Text Classification Model

The classification models accurately distinguished whether an input text was plain or technical. The Gradient Boosting model showed slightly superior results with a precision rate of 97.2% (See Table 2).



a. **Interjections.** These are words or phrases used to express a feeling (e.g., Wow! or Uh-oh). It is uncommon in biomedical settings and is not present in either our technical or plain texts.

b. **Passive Voice:** when the subject undergoes the action of the verb (e.g., 'The cells were counted by the scientist'). According to our qualitative analysis, the use of passive voice can make sentences more complex, less direct, and harder to understand. As evidenced in our quantitative analysis, it is avoided in both scientific/biomedical settings, both in plain and technical texts.

c. **Stopwords.** The proportion of words such as 'a' 'the' are is higher in plain texts, most likely as they aid in the fluency and comprehension of a text by acting as connectors between words, enhancing the coherence and naturalness of sentences for readers.

d. **Complex Words.** The proportion of words with three or more syllables is higher in technical texts, consistent with our qualitative assessments and plain language guidelines.

Figure 2: Comparison of the distribution of a sample of readability metrics or language variables between plain and technical texts.

Metric	RF	GB
F1 Score	0.971	0.975
Accuracy	0.980	0.982
Recall	0.973	0.977
Precision	0.969	0.972

Table 2: Comparison of tested classification models in terms of F1 Score, Accuracy, Recall, and Precision.

3.2 LLM Prompt for Plain Language Summary Generation

3.2.1 Cochrane Reviews: Plain Language Summaries

We randomly selected a sample of 600 Cochrane texts from the main dataset: 300 technical abstracts and the corresponding 300 plain summaries. We then used our prompt in both GPT 3.5 and GPT 4 to generate the plain language summary from the technical abstracts resulting in 300 Plain-GPT 3.5 and 300 Plain-GPT 4 summaries.

We tested the LLM-generated texts with our best model, Gradient Boosting, for plain language classification, and BERTScore to test semantic equivalence against the original Cochrane plain summaries. Our model classified 96% of GPT 3.5 texts and 99.6% of GPT 4 texts as plain. Hence, our prompt is effective in generating PLSs that meet quantitative plain language requirements as defined in our classification model, with GPT 4 showing higher adherence.

The semantic equivalence score, BERTScore, confirmed both GPT 3.5 and GPT 4 successfully retained the original message. However, GPT 4 produced plain summaries that outperformed GPT 3.5 in all parameters (Precision, Recall, and F1-Score) with a significant difference (p-value < 0.05, see Table 3).

3.2.2 Protocol Plain Language Summaries

We randomly selected a sample of nine clinical trial protocols from ClinicalTrials.Gov. Given that their corresponding PPLS were not yet publicly published, we used Trial Summaries by Citeline Regulatory to find the corresponding Results Plain Language Summaries (RPLS) and extracted four sections that are equivalent in a PPLS: ‘Why is this study needed?’: Background and hypothesis of the trial (Rationale), ‘Who will take part in this study?’ (Population), ‘How is this study designed?’ (Trial Design), and ‘What treatments are being given during the study?’ (Interventions).

Quantitative Analysis

We used our prompt specific for PPLS with both GPT 3.5 and GPT 4 to generate the plain language summary from the technical protocols. We used our Gradient Boosting model to verify if LLM-generated texts were plain and BERTScore to check semantic equivalence to the content on the RPLS. All LLM-generated PPLS were classified as plain, and BERTScore confirmed a semantic agreement with the content in the RPLS (see Table 4). Consistent with Cochrane results, GPT 4 produced PPLS with higher semantic equivalence than GPT 3.5 (no statistical analysis due to the small sample size).

Qualitative Analysis

Ratings by 3 domain experts who evaluated each LLM-generated text, demonstrated that GPT 4 outperformed GPT 3.5 in all four criteria: Accuracy, Readability, Completeness, and Usefulness, as indicated by an average score of 4.71 for GPT 4 texts as compared to 3.93 for GPT 3.5 (see Figure 3 and Table 5).

In terms of accuracy, both GPT 3.5 and GPT 4 received high scores. Reviewers noted that both language models exhibited scientific accuracy and relied exclusively on the input text (study protocol). Notably, even when the content in the original RPLS contained inconsistencies (e.g. incorrect age limit or indication), both language models generated accurate PLS. This finding suggests that language models can be used to automatically generate a first draft of a PLS while minimizing data inaccuracies resulting from human error.

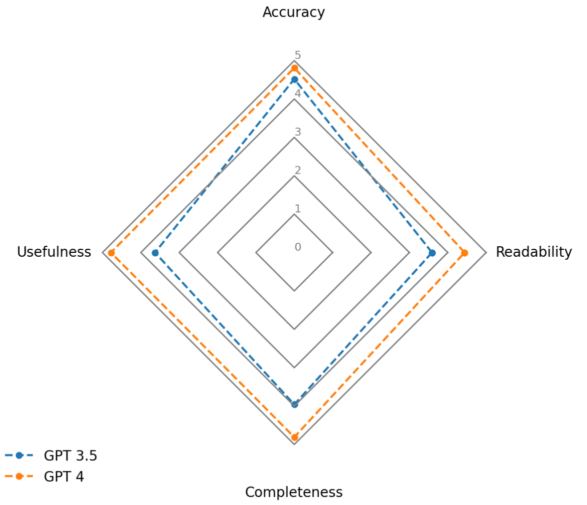


Figure 3: Radar diagram comparing the qualitative assessment of the LLM-generated texts in 4 criteria: Accuracy, Readability, Completeness, and Usefulness.

Regarding readability, both GPT 3.5 and GPT 4 generated texts that were likely to be understood by a lay audience. This observation aligned with the results obtained through the classification model. However, GPT 3.5 occasionally employed complicated medical jargon (e.g., ’chronic’, ’randomized’, ’double-blind’) and longer words and sentences (e.g., ’approximately 640 adults’ vs ’about 640 adults’). Similarly, GPT 4, despite its outstanding performance, occasionally preferred passive voice over active voice, compromising clarity and concise writing. This highlights the importance of quality control by a healthcare professional who should verify the content and style of the automatically generated PLS draft.

Completeness, which assessed the compliance of PPLS content and structure with EU CTR No 536/2014 guidelines, revealed inconsistencies in the outputs generated by GPT 3.5. These inconsistencies manifested as the creation of new, unrequested sections and summaries, with significant variation among the nine generated PLS. Conversely, GPT 4 consistently generated PLSs that adhered to the specified format and content expectations, and complied with the guidelines, showing a remarkable value in automating the time-consuming task of guaranteeing the content to be standardized and aligned with industry-specific and rigorous guidelines.

The usefulness ratings, indicating the suitability of the generated PLSs as draft versions, correlated with the findings in other criteria. GPT 3.5 received moderate scores in generating draft PLS, while

Semantic Equivalence	Plain_GPT 3.5	Plain_GPT4	p-value
Precision	0.790 ± 0.010	0.791 ± 0.015	0.027
Recall	0.772 ± 0.017	0.773 ± 0.016	0.003
F1-Score	0.780 ± 0.015	0.782 ± 0.014	0.001

Table 3: Semantic equivalence score (BERT) between the GPT-generated plain summaries from Cochrane technical abstract vs. original Cochrane PLS.

Semantic Equivalence	PPLS_GPT 3.5	PPLS_GPT4
Precision	0.8040 ± 0.0068	0.8073 ± 0.0208
Recall	0.7940 ± 0.0138	0.7975 ± 0.0129
F1-Score	0.7989 ± 0.0076	0.8023 ± 0.0109

Table 4: Semantic equivalence score (BERT) between the GPT-generated PPLS from clinical trial protocols vs. the original content written for the PLS.

Metric	GPT 3.5	GPT 4
Accuracy	4.52	4.81
Readability	3.59	4.44
Completeness	3.96	4.81
Usefulness	3.63	4.78
Overall Score	3.93	4.71

Table 5: Ratings for GPT 3.5 and GPT 4 plain summaries in 4 criteria: Accuracy, Readability, Completeness, and Usefulness.

GPT 4 scored 4·78, indicating that the generated PLS were highly suitable as draft versions of the PLS.

4 Discussion

In this study, we used NLP and LLMs to improve health literacy by generating PLS from biomedical texts. Our two-part strategy involved creating a classification model for identifying if a text was written in plain language and using LLMs (specifically GPT 3.5 and GPT 4) for the automated generation of the PLS.

The classification model achieved over 97% accuracy, indicating its effectiveness in distinguishing between the text types: technical and plain. This is a very useful stand-alone strategy that could support authoring teams in identifying if their texts targeted for patients or the general audience are compliant with plain language guidelines.

The LLMs exhibited outstanding performance in generating PLS, with GPT 4 outperforming GPT 3.5 in creating content that was both plain and semantically similar. In a qualitative review by domain experts, GPT 4 also surpassed GPT 3.5 by generating high-quality drafts of PLS. These drafts were scientifically accurate, compliant with plain language requirements, and met expectations in content and structure. These results underline the value of LLMs in supporting healthcare stakeholders to streamline the generation of plain documents, and with that, promote equitable access to biomedical information, engagement of the lay audience in health-related decision-making, and improved health outcomes.

Our study highlights the importance of using well-designed, structured, and domain-specific prompts to guarantee the creation of high-quality, easily comprehensible PLS. This is particularly vital when accuracy in biomedical facts is essential. This requires the collection of feedback from stakeholders who are experts in the domain or field of interest. Such feedback would help to fine-tune the prompts and guarantee that the output fulfills the purposes of different document types. Our study exemplified this with various document types (e.g., Cochrane reviews, PPLS), some of which adhere to strict industry standards.

While the findings of our study are promising, they also underscore opportunities for further research to fully harness the potential of NLP and LLMs in this context. Future studies could involve direct audience feedback in evaluating the understandability of PLS. This would ensure that the generated content aligns with the comprehension levels of the intended audience, such as patients in clinical settings, and would provide cues for ways in which they could improve their interaction with biomedical content, improving adherence to treatment plans or educating them about a disease or diagnosis. Additionally, depending on the intended use and field of interest, refining the models could potentially account for specific linguistic nuances, exploring advanced techniques like Retrieval Augmented Generation (RAG) could enhance factual accuracy, and expanding the dataset to include a wider range of texts and languages could enhance the generalizability of the classification model and applicability of the LLMs. Different interesting opportunities to leverage NLP and LLMs to serve society by simplifying what would otherwise be daunting.

In conclusion, by leveraging the capabilities of NLP and LLMs, we have taken a significant step towards bridging the gap between complicated biomedical texts and comprehensible summaries designed for the general audience. This framework paves the way for prospective innovations in the field of health literacy, which, in turn, holds the potential to enhance health outcomes and foster health equity.

5 Limitations

Our study has taken a significant step towards leveraging NLP and LLM to bridge the gap between complicated biomedical texts and comprehensible summaries designed for the general audience. However, the following are limitations that we've identified, and which should be considered by fellow researchers and users interested in applicability of our methodology to generate PLS from biomedical texts:

Dataset

The size and diversity of the dataset we used to train our classification model and to define our prompt is not representative of all types of biomedical text. Despite collecting an extensive and diverse data set, the type of texts we have used may pose a limitation in the generalizability of our findings. Most especially, it may impact the precision and accuracy of our classification model when applied to texts from different biomedical subfields. This outlines an opportunity for those interested in replicating our findings to specific document types (e.g., other biomedical subfields or even other languages) to enrich their dataset with such types of documents.

Qualitative Assessment

Our qualitative evaluation was conducted by a few domain experts on a limited number of texts (e.g., nine clinical trial protocols). While this sheds light on the value of LLM to generate high-quality PLS, a broader applicability of our results requires the collection of feedback from different and larger sets of stakeholders to continue fine-tuning the prompts (e.g., patients, medical writers, and clinicians). Working with clinical data requires high accuracy, thus applying our findings to real-world settings must follow rigorous testing to guarantee PLS are appropriate for the targeted audience. Additionally, we encourage using the automated PLS as a first draft which would then benefit from proofreading and quality control (i.e. human oversight), most especially in highly regulated settings.

Type of LLMs

Our findings relied on GPT, a non-open source LLM. We were focused on having a proof of concept to test the capacity to which LLMs could generate high-quality PLS. Yet, there's still much to explore with the advent of newer and more robust LLMs, including open-source alternatives. Also, we were unable to test on open models due to resource limitations in using OpenAI-comparable models.

References

B. Bahador, S. Baedorf Kassis, H. Gawrylewski, and et al. 2020. Promoting equity in understanding: A cross-organizational plain language glossary for clinical research. *Medical Writing*, 29(4):10–15.

N. D. Berkman, S. L. Sheridan, K. E. Donahue, and et al. 2011a. Health literacy interventions and outcomes: an updated systematic review. *Evidence Report/Technology Assessment*, 199:1–941.

N. D. Berkman, S. L. Sheridan, K. E. Donahue, D. J. Halpern, and K. Crotty. 2011b. Low health literacy and health outcomes: an updated systematic review. *Annals of Internal Medicine*, 155(2):97–107.

C. Brach. 2023. Ahrq health literacy universal precautions toolkit, 3rd edition. AHRQ Publication No. 23-0075, Accessed November 20, 2023.

Centers for Disease Control and Prevention. 2022. Your guide to clear writing. Accessed November 15, 2023.

Crummy. 2023. Beautiful soup 4 4.10. Accessed December 2022.

European Union. 2023. Q&a: Clinical trial regulation (eu) no 536/2014 2023. Accessed December 26, 2023.

GDPR. 2023. General data protection regulation (gdpr) - the principle of transparency. Accessed December 22, 2023.

T. A. Miller. 2016. Health literacy and adherence to medical treatment in chronic and acute illness: A meta-analysis. *Patient Education and Counseling*, 99(7):1079–1086.

L. Nielsen-Bohlman, A. M. Panzer, and D. A. Kindig. 2004. *Health Literacy: A Prescription to End Confusion*. National Academies Press.

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

Pfizer. 2023. Plain language study results summaries. Accessed September 2023.

Pharma Intelligence UK Limited. 2023. Citeline trial summaries citeline regulatory. Accessed September 2023.

D. Schillinger. 2021. Social determinants, health literacy, and disparities: Intersections and controversies. *HLRP: Health Literacy Research and Practice*, 5(3):233–243.

Selenium. 2023. Selenium 4.4. Accessed December 2022.

SpaCy. 2023. Spacy. Accessed November 2023.

M. Stoll, M. Kerwer, K. Lie, and A. Chasiotis. 2022. Plain language summaries: A systematic review of theory, guidelines, and empirical research. *PLoS ONE*, 17(6):e0268789.

C. Stormacq, S. Van den Broucke, and J. Wosinski. 2019. Does health literacy mediate the relationship between socioeconomic status and health disparities? integrative review. *Health Promotion International*, 34(5):e1–e17.

Kristine Sørensen, Jürgen M. Pelikan, Florian Röthlin, Kristin Ganahl, Zofia Slonska, Gerardine Doyle, James Fullam, Barbara Kondilis, Demosthenes Agrafiotis, Ellen Uiters, Maria Falcon, Monika Mensing, Kancho Tchamov, Stephan van den Broucke, and on behalf of the HLS-EU Consortium Brand, Helmut. 2015. Health literacy in Europe: comparative results of the European health literacy survey (HLS-EU). *European Journal of Public Health*, 25(6):1053–1058.

The Plain Language Action and Information Network. 2011. Federal plain language guidelines. Accessed November 20, 2023.

A. Trezona, G. Rowlands, and D. Nutbeam. 2018. Progress in implementing national policies and strategies for health literacy-what have we learned so far? *International Journal or Environmental Research and Public Health*, 15(7):1554.

United Health Group. 2023. Just plain clear glossary. Accessed December 5, 2023.

U.S National Library of Medicine (NIH). 2023a. Clinicaltrials.gov. Accessed November 2023.

U.S National Library of Medicine (NIH). 2023b. Clinicaltrials.gov api. Accessed November 2023.

A Supplemental Material

Data Source	**Text Type**	**Overview**	**Count of Texts**	**Extraction Method**
U.S National Library of Medicine (NIH), ClinicalTrials.gov	Technical	Largest and publicly available database of clinical research studies and information about their results (U.S National Library of Medicine (NIH), 2023a).	100	ClinicalTrials.Gov API (Application Programming Interface) that provides access to all posted information on study records (U.S National Library of Medicine (NIH), 2023b).
Cochrane Library by Wiley	Technical and Plain	International not-for-profit organization that presents trusted synthesized reviews of biomedical research projects in 2 formats: a technical abstract and plain language summary.	8465 Research Projects (13,922 texts). *Texts shorter than 250 were excluded.	Python Libraries: Selenium to automate web browser interactions with Python (2023) and Beautiful Soup for web scraping (2023).
Pfizer Results Plain Language Summaries	Plain	Plain Language Study Results Summaries (RPLS) of the design and results of Pfizer clinical studies Pfizer (2023). Specific sections of the RPLS containing tables or diagrams were excluded during processing.	125	Given the diversity of clinical trial sponsors (Pfizer, GSK, etc.), specific sections of interest of the RPLS PDF documents were mapped and extracted (e.g., what happened during the Study?).
Trial Summaries by Citeline Regulatory	Plain	Trial results summaries (RPLS) for studies that started in late 2015 and beyond as provided by the study sponsors (e.g., AstraZeneca, GSK, Amgen, Astellas, Sanofi) (Pharma Intelligence UK Limited, 2023).	294	Automatic extraction of PDF content led to errors such as missing letters, combined words, or words separated by syllables. We then used GPT 3.5 API on the extracted texts to correct those texts errors only and guarantee texts were exactly as found in the RPLS PDFs.

Table A1: Overview of the data sources used in the study. All the texts in our data sources can be found in our GitHub Data Repository (blinded).

Readability Indexes	Flesch-Kincaid, Automated Readability Index (ARI), Coleman-Liau, Flesch Reading Ease, Gunning Fog, Lasbarhets index (LIX), Simple Measure of Gobbledygook (SMOG), Dale-Chall, and Anderson's Readability Index (RIX)
Linguistic Characteristics	Complex words, variability of type of words, stop words, proper nouns, long words, punctuation, numbers, symbols, organization, sentences per paragraph, characters per word, type token ratio, total characters in the text, total syllables in the text, syllables per word, count of cardinal numbers in the text, auxiliary verbs, active tokens (number of tokens that are actively being processed in the texts), determiners, percentages, pronouns, use of active voice, use of passive voice, use of the verb 'To be,' dates, count of sentences in the text, words per sentence, nominalization (verbs, adjectives, or other linguistic elements that were turn into nouns), subordinating conjunctions, grammatical particles, adverbs, auxiliary verb, prepositions, adjectives, nouns, use of conjunctions, interjections, ordinal numbers, mention of persons, nationalities, religious or political affiliations, works (eg, art, books, movies), geopolitical entities, quantities, facilities (building, airports, roads), geographical locations, products, laws, times, or money, and miscellaneous (elements that didn't fit any of the previous categories).

Table A2: Variables that were used to describe the readability and linguistic characteristics of the technical and plain biomedical texts.

Using the following abstract of a biomedical study as input, generate a Plain Language Summary (PLS) understandable by any patient, regardless of their health literacy. Ensure that the generated text adheres to the following instructions which should be followed step-by-step:

a. Specific Structure: The generated PPLS should be presented in a logical order, using the following headings:

1. Plain Protocol Title
2. Rationale
3. Objectives
4. Trial Design
5. Trial Population
6. Interventions

b. Sections should be authored following these parameters:

1. **Plain Protocol Title:** Simplified protocol title understandable to a layperson but including specific indication for which the study is meant.
2. **Rationale:** Include the phrase 'Researchers are looking for a better way to treat [condition]; background or study rationale describing the condition: what it is, what it may cause, and why it is a burden for the patients; the reason and main hypothesis for the study; and why the study is needed, and the study medication has the potential to treat the condition.
3. **Objectives:** Answer 'What are the goals of the study?' Specify the main and secondary objectives of the trial and how they will be measured (e.g., the main trial endpoint is the percent change in the number of events from baseline to a specified time or the total number of adverse reactions at a particular time after baseline).
4. **Trial Design:** Answer 'How is this study designed?' Include the description of the design and the expected amount of time a person will be in the study.
5. **Trial Population:** Answer 'Who will participate in this study?' Include a description of the study and patient population (age, health condition, gender), and the key inclusion and exclusion criteria.
6. **Interventions:** Answer 'What treatments are being given during the study?' Include a description of the medication, vaccine, or treatment(s) being studied, the route of administration, the duration of treatment, and any study-related diagnostic and monitoring procedures used. Include justification if a placebo is used.

c. Consistency and Replicability: The generated PPLS should be consistent regardless of the order of sentences or the specific phrasing used in the input protocol text.
d. Compliance with Plain Language Guidelines: The generated PPLS must follow these plain language guidelines:

- Have readability grade level of 6 or below.
- Do not have jargon. All technical or medical words or terms should be defined or broken down into simple and logical explanations.
- Active voice, not passive.
- Mostly one or two-syllable words.
- Sentences of 15 words or less.
- Short paragraphs of 3-5 sentences.
- Simple numbers (e.g., ratios, no percentages).

e. No Extra Content: The AI model should not invent information or add content that is not present in the input protocol. The PPLS should only present information from the original protocol in a simplified and understandable manner.
f. Aim for an approximate PPLS length of 700-900 words.

Figure B1: Prompt to translate a protocol into a plain language summary compliant with EU CTR No 536/2014.

Using the following abstract of a biomedical study as input, generate a Plain Language Summary (PLS) understandable by any patient, regardless of their health literacy. Ensure that the generated text adheres to the following instructions which should be followed step-by-step:

a. Specific Structure: The generated PLS should be presented in a logical order, using the following order:

1. Plain Title
2. Rationale
3. Trial Design
4. Results

b. Sections should be authored following these parameters:

1. **Plain Title:** Simplified title understandable to a layperson that summarizes the research that was done.
2. **Rationale:** Include: background or study rationale providing a general description of the condition, what it may cause or why it is a burden for the patients; the reason and main hypothesis for the study; and why the study is needed, and why the study medication has the potential to treat the condition.
3. **Trial Design:** Answer 'How is this study designed?' Include the description of the design, description of study and patient population (age, health condition, gender), and the expected amount of time a person will be in the study.
4. **Results:** Answer 'What were the main results of the study', include the benefits for the patients, how the study was relevant for the area of study, and the conclusions from the investigator.

c. Consistency and Replicability: The generated PLS should be consistent regardless of the order of sentences or the specific phrasing used in the input protocol text.

d. Compliance with Plain Language Guidelines: The generated PLS must follow all these plain language guidelines:

- Have readability grade level of 6 or below.
- Do not have jargon. All technical or medical words or terms should be defined or broken down into simple and logical explanations.
- Active voice, not passive.
- Mostly one or two syllable words.
- Sentences of 15 words or less.
- Short paragraphs of 3-5 sentences.
- Simple numbers (e.g., ratios, no percentages).

e. Do not invent Content: The AI model should not invent information. If the AI model includes data other than the one given in the input abstract, the AI model should guarantee such data is verified and real.

f. Aim for an approximate PLS length of 500-900 words.

Figure B2: Prompt to translate Cochrane technical abstract into a plain language summary.