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

Felipe Arias-Russi[1,2], Carolina Salazar-Lara[3], Rubén Manrique[1]
[1]Systems and Computing Engineering Department, Universidad de los Andes, Bogotá D.C.
[2]Department of Mathematics, Universidad de los Andes, Bogotá D.C.
[3]Department of Biomedical Engineering, Universidad de los Andes, Bogotá D.C.

{af.ariasr, c.salazar499, rf.manrique}@uniandes.edu.co

Abstract

Health literacy enables individuals to navigate healthcare systems and make informed decisions. Plain language summaries (PLS) can bridge comprehension gaps by simplifying complex biomedical texts, yet their manual creation is both time-consuming and challenging. This study advances the field by (1) constructing a novel corpus of paired technical and plain language texts from medical trial libraries, (2) developing machine learning classifiers to rapidly identify plain language features, and (3) establishing a multi-dimensional evaluation framework that integrates computational metrics with human expertise. We iteratively optimized prompts for diverse large language models (LLMs)—including GPT models, Gemini 1.5, DeepSeek-R1, and Llama-3.2—to generate PLS variants aligned with domain-specific guidelines. Our classifier achieved 97.5% accuracy in distinguishing plain from technical language, and the generated summaries demonstrated high semantic equivalence to expert-written versions.

1 Introduction

Health literacy refers to an individual's capacity to access, understand, and use health information (Nielsen-Bohlman et al., 2004). This ability is essential for patients and their families to effectively navigate healthcare systems, comprehend medical instructions, adhere to treatment regimens, and make informed decisions about clinical trials, treatments, or procedures (Berkman et al., 2011a,b; Miller, 2016). However, inadequate health literacy remains a widespread problem, one that has been linked to increased mortality, higher rates of preventable hospitalizations, and poorer treatment adherence (Berkman et al., 2011a). In particular, the 2015 European Health Literacy Survey found that nearly half of the respondents, particularly older adults, people with financial constraints, or those with lower educational attainment, exhibit insufficient health literacy (Sørensen et al., 2015; Bahador et al., 2020).

In today's healthcare landscape, where patient participation in decision-making is increasingly critical, improving health literacy is essential to reduce disparities and improve public health outcomes (Nielsen-Bohlman et al., 2004; Stormacq et al., 2019; Schillinger, 2021). Moreover, aligning with the transparency principles of the General Data Protection Regulation (GDPR) (GDPR, 2023; Trezona et al., 2018), stakeholders are compelled to ensure that health documentation is both clear and accessible.

Plain language summaries (PLS) offer a viable solution by translating complex clinical and scientific texts into accessible language (Bahador et al., 2020; Centers for Disease Control and Prevention, 2022). However, the manual production of such summaries is labor-intensive and particularly challenging in fields dominated by technical terminology. While large language models (LLMs) have demonstrated promise in automating the generation of lay summaries, previous efforts have largely centered on text generation, often overlooking the need for systematically curated training data and comprehensive evaluation frameworks.

To bridge these gaps, our work introduces a novel resource and an integrated methodological framework that addresses key challenges in health communication. By compiling a corpus of paired technical and plain language texts from medical trial libraries, we provide a valuable dataset that underpins the development of machine learning classifiers capable of rapidly distinguishing between plain and technical language. Using state-of-the-art LLMs and iteratively refining our prompts, we generate plain-language variants that adhere to domain-specific guidelines. Furthermore, our evaluation framework, which combines automated metrics with an expert in health literacy assessments, of-

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 269–284

fers critical insights into the factors that define an effective plain-language summary.

Through this integrated approach, our study not only provides practical tools for producing patient-centered medical summaries but also enhances our understanding of the linguistic variables that support clear and accessible healthcare communication.

2 Related Work

Recent efforts in biomedical text simplification have increasingly focused on automatically generating PLS using NLP and LLMs. Ondov et al. (2022) reviewed a range of approaches and observed that, although neural methods show promise, their progress is limited by the scarcity of high-quality, parallel corpora. This data challenge was similarly highlighted by Devaraj et al. (2021), who introduced a new corpus of parallel texts specifically designed to aid the training of models that could effectively reduce jargon in biomedical information.

LLMs offer a compelling solution to overcome these limitations due to their extensive training data and advanced text generation capabilities. For instance, the BioLaySumm contest (Goldsack and Lin, 2025) targets the task of generating PLS from abstracts. In the 2023 BioLaySumm Task, Turbitt et al. (2023) demonstrated that GPT-3.5—when used in a few-shot setting—produced summaries with superior relevance and factuality compared to those of the specialized BioGPT model, despite the latter's advantage in readability. Additional studies (Veen et al., 2024; Mirza et al., 2024) further indicate that LLMs can outperform human experts in summarizing clinical texts and enhancing the clarity of informed consent documents.

However, there remains a critical need for systematically curated datasets and evaluation frameworks that combine computational metrics with human expertise. We aim to enhance existing work by building a comprehensive database of plain and technical biomedical texts. We will then implement advanced LLMs alongside a classification system to automatically ensure that the generated summaries are composed in plain language. Additionally, we will conduct a thorough evaluation of the generated PLS by domain experts, employing metrics such as readability, factuality, and accuracy, as outlined in the BioLaySumm shared task.

3 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.

3.1 Data Collection and Processing

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

We further enlarged the dataset by treating each paragraph of at least 250 words as a distinct unit, while excluding texts shorter than 250 words. As a result, our "augmented dataset" contained 61,354 texts, split into 16,731 plain and 31,740 technical texts for training, and 5,090 plain and 7,793 technical texts for testing. To mitigate source imbalance, we limited the dataset to 23,695 texts, divided into 9,093 plain and 8,654 technical for training, and 2,741 plain and 3,205 technical for testing. Additionally, we obtained a validation set of PLOS and eLife texts from (Goldsack et al., 2022; Luo et al., 2022) to evaluate the ML models on a dataset external to our own.

3.2 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.

3.2.1 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

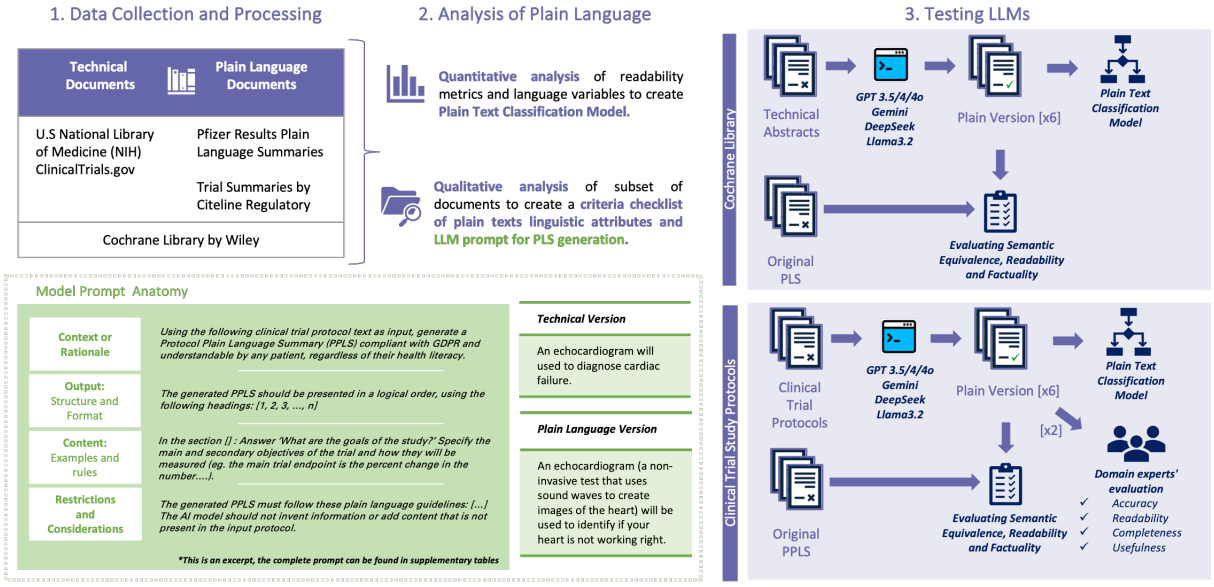


Figure 1: **Methodology.** Our methodology consists of three main steps: (1) collecting and processing biomedical texts (technical and plain language documents) to construct training and testing datasets, (2) conducting quantitative analysis to develop a plain language classification model and qualitative analysis to identify linguistic traits guiding prompt engineering for LLM-based PLS generation, and (3) evaluating LLM-generated PLS both quantitatively—using our classification model, semantic equivalence/relevance (BERTScore, Zhang et al. (2020)), factuality (AlignScore, Zha et al. (2023)), and readability metrics—and qualitatively through expert assessments.

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.

3.2.2 Quantitative Analysis

We computed readability metrics and language variables for each text in the augmented dataset using the Readability (2019) and SpaCy (2023) libraries, respectively. This resulted in 64 variables presenting each text's readability and linguistic traits (see Table B1 and Section B).

For each language variable characteristic k, we evaluated its discriminative potential for classifying texts as either technical or plain. To this end, we randomly selected a sample of size n from the plain texts, denoted by

$$X_1^{(k)},\ X_2^{(k)},\ \ldots,\ X_n^{(k)} \sim P_X^{(k)},$$

and a corresponding sample of size n from the technical texts, denoted by

$$Y_1^{(k)},\ Y_2^{(k)},\ \ldots,\ Y_n^{(k)} \sim Q_Y^{(k)}.$$

An independent hypothesis test was then conducted for each k to determine whether the distributions differ statistically between the two text types.

Specifically, for each linguistic feature k, we considered the following hypotheses:

- **Null Hypothesis** ($H_0^{(k)}$): $P_X^{(k)} = Q_Y^{(k)}$. The distributions of the characteristic k for plain and technical texts are identical.
- **Alternative Hypothesis** ($H_1^{(k)}$): $P_X^{(k)} \neq Q_Y^{(k)}$. The distributions of the characteristic k for plain and technical texts differ.

To evaluate these hypotheses, we employed several non-parametric tests, namely the Wilcoxon signed-rank test (Wilcoxon, 1945), the Kolmogorov-Smirnov test (Kolmogorov, 1933), and the Mann–Whitney U test (Mann and Whitney, 1947), ensuring robustness across different statistical assumptions. Since a total of 64 independent hypothesis tests were performed (one for each characteristic k), a Bonferroni correction (Benjamini and Hochberg, 1995) was applied to control the family-wise error rate. Thus,

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.

the nominal significance level of $\alpha = 0.05$ was adjusted to $\alpha' = \frac{0.05}{64} \approx 0.0008$.

Figure 2 illustrates examples of the distribution comparisons for selected characteristics. Notably, of the 64 characteristics examined, only 'Interjections' and 'Passive Voice' did not provide sufficient evidence to reject the null hypothesis (i.e., their p-values exceeded 0.0008), whereas the remaining 62 characteristics exhibited statistically significant differences and were subsequently incorporated into our classification model.

3.3 Plain Texts Classification Model

We used the reduction of the augmented dataset and first preprocessed the 62 linguistic variables by applying standard min-max normalization. For variables representing counts of specific word types, normalization was performed relative to the total number of words in the text. We then built our models using the processed features.

For the Gradient Boosting (GB) model, we manually set the parameters as follows: the number of estimators was fixed at 120 (i.e., the number of boosting stages), the learning rate was set to 0.25 to scale the contribution of each tree, a subsample rate of 0.8 was used to fit each base learner on 80% of the training instances, the maximum depth of each tree was limited to 5 to minimize overfitting, a minimum of 5 samples was required to split an internal node, and at least 3 samples were needed in a leaf node. A fixed random state (0) ensured reproducibility.

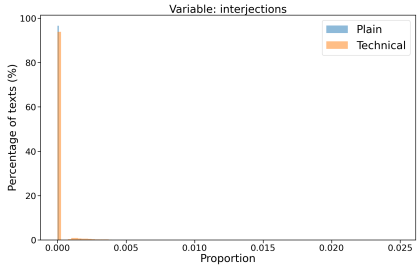
For the Random Forest (RF) model, we configured 100 estimators (trees) with a maximum tree depth of 10 and also set the random state to 0.

Note that we did not perform automated hyperparameter tuning (e.g., using grid search) or use K-fold cross-validation to select optimal training and testing splits; instead, the parameters were adjusted manually through trial and error, given the rapid training times observed.

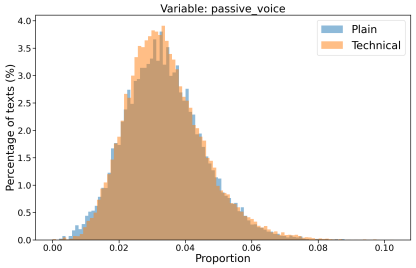
3.4 LLM Prompt for Plain Language Summary Generation

Our objective was to design a prompt for LLMs capable of translating biomedical technical documents into plain language summaries (PLS). Beginning with a clinical trial protocol from ClinicalTrials.gov (see data sources in Table A1), we used an initial simple prompt: *"Using the following clinical trial protocol text as input, create a plain language summary."* We tested this prompt using both GPT-3.5 and GPT-4, analyzed the generated outputs, and iteratively refined the prompt by adding further details and instructions.

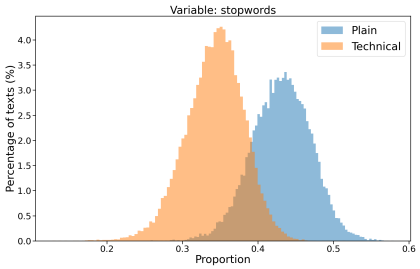
We aimed to produce a PLS that met the following qualitative criteria: **(1) Accuracy:** the content is clinically and scientifically correct; **(2) Readability:** the text is grammatically correct and easily understood by a lay audience (as defined in Table 1); **(3) Completeness:** the summary adheres to the expectations of a Protocol Plain Language Sum-



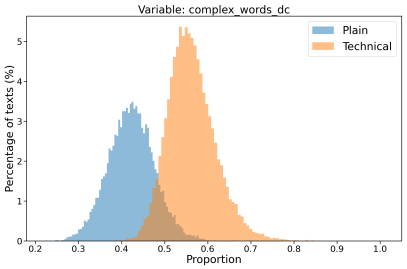
a. **Interjections.** These are words or phrases used to express a feeling (e.g., Wow! or Uh-oh). It is uncommon in biomedical settings and is not present in either our technical or plain texts.



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' and 'the' is higher in plain texts 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.

mary (PPLS) as specified by EU CTR No 536/2014 (United Health Group, 2023); and **(4) Usefulness:** the generated PLS can serve as a reliable first draft for final study documentation.

Because PPLS are intended for review by professional evaluators, they required a higher level of care and were generated in limited numbers. This qualitative evaluation method, although rigorous, did not scale efficiently to large sample sizes. To address this limitation, for the more numerous Cochrane Review PLS we adopted a scalable, quantitative evaluation approach based on the three criteria used in the BioLaySumm competition (Goldsack and Lin, 2025). Specifically, we assessed:

1. **RELEVANCE:** measuring the semantic similarity between the LLM-generated summaries and a ground-truth summary (created by a human) using BERTScore (Zhang et al., 2020);
2. **FACTUALITY:** evaluating the consistency of the generated content with the source text (i.e., ensuring that no contradictory information is introduced) using AlignScore (Zha et al., 2023); and
3. **READABILITY:** assessing grammaticality and ease of comprehension through computational metrics such as Flesch–Kincaid Grade Level (Flesch, 1948), Coleman-Liau Index (Coleman and Liau, 1975), Flesch Reading Ease, Gunning Fog Index (Gunning, 1952), SMOG readability formula, and Dale–Chall Readability Score (Chall and Dale, 1995).

In addition, we considered the CLASSIFICATION results from our best ML model, which predicts if the LLM-generated text is plain or technical.

Our final prompt (see Figure C2) for generating

a PPLS included the following elements:

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

After finalizing the prompt for generating a PPLS, we used a similar approach to create a prompt for generating Cochrane Review PLS (see Table A1 and Figure C1). This two-pronged strategy allowed us to balance the need for careful, qualitative review (for PPLS) with a scalable, quantitative evaluation method (for Cochrane PLS) that can handle larger sample sizes.

4 Results

4.1 Plain Texts Classification Model

The classification models accurately distinguished between plain and technical texts. The GB model, in particular, achieved a slightly higher F1 Score (see Table 2). Since most of the training data were derived from Cochrane texts, we further evaluated the models on a completely separate validation set composed of PLOS and eLife documents (see Table A1) to assess potential bias. The performance metrics, reported as Main/PLOS+eLife in Table 2, indicate that the models generalize well to unseen data and exhibit minimal bias.

Metric	Main (Test)		PLOS + eLife (Test)	
	RF	GB	RF	GB
Accuracy	0.968	0.9752	0.9421	0.9557
Recall	0.973	0.9813	0.9616	0.9672
Precision	0.959	0.9655	0.9255	0.9455
F1 Score	0.966	0.9734	0.9432	0.9562

Table 2: Performance comparison of classification models on the Main Dataset and the PLOS + eLife test dataset.

4.2 LLM Prompt for Plain Language Summary Generation

4.2.1 Cochrane Reviews: Plain Language Summaries

We randomly selected 600 Cochrane texts from the main dataset—300 technical abstracts and their corresponding plain language summaries (ground truth). Using our final prompt, we generated summaries for the technical abstracts and computed average metrics—READABILITY, FACTUALITY, and RELEVANCE—for each model (Table 3). The factuality metric was calculated using the original abstracts to ensure the summaries remained faithful. We also evaluated classification accuracy (i.e., whether our ML classifier recognized the summaries as plain language) as shown in Table 4.

Overall, API-based models produced summaries consistently classified as plain language, while locally executed models tended to yield more technical outputs, as indicated by lower readability scores. Among the GPT models, those with higher readability were more often recognized as plain language, although their factuality and relevance were slightly lower than those of GPT-3.5. These results suggest that some models generate easier-to-read texts, whereas others retain a more technical tone.

4.2.2 Protocol Plain Language Summaries

We randomly selected a sample of nine clinical trial protocols from ClinicalTrials.gov. Since the corresponding PPLS were not publicly available, we used Trial Summaries by Citeline Regulatory to obtain the Results Plain Language Summaries (RPLS) and extracted four sections equivalent to a PPLS: 'Why is this study needed?' (Background and hypothesis of the trial, i.e., 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 generated PPLS from technical protocols using our prompt with both API-based models (e.g., GPT-3.5, GPT-4, GPT-4o, Gemini-1.5) and locally executed models (DeepSeek R1, Llama-3.2). For each model, we computed average metrics for READABILITY, FACTUALITY (AlignScore), and RELEVANCE (BERTScore), as shown in Table 3. Our ML classifier also confirmed that nearly all outputs were recognized as plain language (see Table 4).

Quantitative Evaluation for Cochrane

	READABILITY						FACTUALITY	RELEVANCE
Model	**CLI ↓**	**FRE ↑**	**GFI ↓**	**SMOG ↓**	**FKGL ↓**	**DCRS ↓**	**AlignScore ↑**	**BERTScore ↑**
deepseek-r1-7b	16.99	22.75	17.69	12.31	14.80	9.45	0.7955	0.8496
gemini-1.3-flash	**9.60**	**66.87**	**8.75**	**9.08**	**6.90**	**5.94**	0.6333	0.8474
gpt_4-32k	12.48	48.52	13.39	11.20	10.80	7.41	0.7801	0.8519
gpt_4o	11.49	57.13	11.16	9.91	9.09	6.88	0.7383	0.8527
gpt_35-turbo-16k	15.52	28.08	17.33	12.59	13.91	8.60	0.8781	**0.8585**
llama-3.2-3b	16.42	21.96	18.58	10.79	15.73	9.39	**0.8785**	0.8490

Quantitative Evaluation for PPLS

	READABILITY						FACTUALITY	RELEVANCE
Model	**CLI ↓**	**FRE ↑**	**GFI ↓**	**SMOG ↓**	**FKGL ↓**	**DCRS ↓**	**AlignScore ↑**	**BERTScore ↑**
deepseek-r1-7b	15.70	24.73	15.03	11.88	13.89	9.88	0.9657	0.8305
gemini-1.3-flash	**9.11**	**65.09**	**8.61**	11.40	**6.74**	**5.75**	0.9331	**0.8479**
gpt_4-32k	10.86	52.26	12.15	**10.45**	10.79	6.86	0.9646	0.8472
gpt_4o	11.20	55.67	10.37	10.97	8.91	7.05	0.9515	0.8465
gpt_35-turbo-16k	14.30	29.10	16.07	13.49	13.68	8.15	0.9697	0.8434
llama-3.2-3b	13.54	35.17	14.75	11.72	13.26	8.47	**0.9826**	0.8386

Table 3: Comparison of model metrics. **Upper table:** Metrics computed as averages from generated summaries derived from 300 Cochrane abstracts. **Lower table:** Metrics computed as averages over the 9 generated PPLS produced by the LLMs. Best values are in **bold** and worst values are underlined. READABILITY metrics are lower-is-better (except FRE, where higher is preferred), while FACTUALITY and RELEVANCE are higher-is-better.

	CLASSIFICATION	
Model	**Cochrane**	**PPLS**
deepseek-r1-7b	0.5567	0.5556
gemini-1.3-flash	**1.0000**	**1.0000**
gpt_4	0.9433	**1.0000**
gpt_4o	0.9767	**1.0000**
gpt_35	0.8733	**1.0000**
llama-3.2-3b	0.7033	0.7778

Table 4: Accuracy of generated summaries as determined by our plain language classifier. Since all outputs should be plain language by instruction, these results indicate the extent to which each model adheres to this requirement.

Overall, API-based models achieved higher precision and better factuality, while locally executed models performed worse due to computational limitations. Among the GPT models, GPT-4 and GPT-4o produced the most readable summaries (and were most frequently classified as plain language), though their factuality and relevance were slightly lower than those of GPT-3.5. These results indicate that models like GPT-4o, Gemini-1.5, and GPT-4 tend to generate easier-to-read texts, whereas DeepSeek R1 and Llama-3.2 yield more technical summaries.

Qualitative Analysis

For the qualitative evaluation, only the plain language summaries generated by GPT-3.5 and GPT-4 were selected. Due to time constraints for experts, we selected only the best models based on previous results, considering that GPT-4o has minimal differences from GPT-4 in content generation. Ratings by three 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 overall score of 4.71 for GPT-4 texts compared to 3.93 for GPT-3.5 (see Figure 3 and Table 5).

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

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 proto-

col). Notably, even when the content in the original RPLS contained inconsistencies (e.g., an 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.

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 language summaries in four criteria: Accuracy, Readability, Completeness, and Usefulness.

5 Discussion

In this study, we used NLP and LLMs to improve health literacy by generating PLS from biomedical texts. Our approach involved building a robust database that generalizes well across diverse sources and developing a highly accurate classification model to distinguish technical from plain texts. This model serves as a valuable tool for ensuring that patient-targeted documents adhere to plain language guidelines, while our LLM-based generation framework leverages well-designed, domain-specific prompts to produce PLS.

Our evaluation shows that API-based models generally generate easier-to-read and more semantically faithful summaries, although they sometimes exhibit slightly lower factuality—possibly due to hallucination issues. In contrast, locally executed models, while maintaining acceptable factual accuracy, tend to yield more technical outputs, most probably because they have difficulty understanding instructions better, due to computational limitations. Qualitative feedback from domain experts confirmed that GPT 4 outperformed GPT 3.5 in terms of accuracy, readability, completeness, and usefulness. These findings highlight the value of using well-designed, domain-specific prompts and robust LLMs to streamline the generation of plain language summaries. Future research should explore the use of fully-featured, open-source models comparable to the API-based alternatives and incorporate broader stakeholder feedback to refine these methods for diverse biomedical domains.

In conclusion, by leveraging the capabilities of NLP and LLMs, our framework represents a significant step towards bridging the gap between complex biomedical texts and comprehensible summaries for the general audience, paving the way for innovations in health literacy.

6 Future Work

We plan to expand and diversify our dataset by incorporating the full collections of PLOS and eLife, obtaining more plain language samples, and employing advanced techniques to better separate and curate the data.

Future evaluations should include a larger and more diverse set of documents as well as input from multiple stakeholder groups (e.g., patients, medical writers, and clinicians). Additionally, further research should explore advanced prompt engineering techniques, such as chain-of-thought strategies, particularly for open-source models.

7 Limitations

Our study has some limitations. First, our dataset is predominantly composed of Cochrane texts with very few samples from other sources (e.g., Pfizer), which may lead to overfitting and reduce generalizability. Additionally, the current database is not human-curated, which may introduce parsing errors or inaccuracies. Second, our qualitative assessment was based on a limited number of clinical protocols and evaluated only the outputs from GPT-3.5 and GPT-4, with feedback from just a few domain experts. Furthermore, due to computational and API cost constraints, the number of generated samples was limited, potentially affecting the statistical significance of our findings and complicating comparisons between API-based and locally executed models.

References

Jonathan Anderson. 1983. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, 26(6):490–496. Publisher: [Wiley, International Reading Association].

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.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

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.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books. Google-Books-ID: 2nbuAAAAMAAJ.

Meri Coleman and T. L. Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284. Place: US Publisher: American Psychological Association.

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

Ashwin Devaraj, Iain Marshall, Byron Wallace, and Junyi Jessy Li. 2021. Paragraph-level Simplification of Medical Texts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4972–4984, Online. Association for Computational Linguistics.

European Union. 2023. Q&A: Clinical Trial Regulation (EU) No 536/2014 2023. Accessed December 26, 2023.

Rudolph Flesch. 1948. A New Readability Yardstick. *Journal of Applied Psychology*, 32(3):221–233. Place: US Publisher: American Psychological Association.

GDPR. 2023. General Data Protection Regulation (GDPR) - The principle of Transparency. Accessed December 22, 2023.

Tomas Goldsack and Chenghua Lin. 2025. Overview of the biolaysumm 2025 shared task on the lay summarization of biomedical research articles. In *The 24rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making Science Simple: Corpora for the Lay Summarisation of Scientific Literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill. Google-Books-ID: ofI0AAAAMAAJ.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

A. N. Kolmogorov. 1933. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability Controllable Biomedical Document Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Henry B. Mann and Donald R. Whitney. 1947. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.

G. Harry Mc Laughlin. 1969. SMOG Grading: A new readability formula. *Journal of Reading*, 12(8):639–646. Publisher: [Wiley, International Reading Association].

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.

Fatima N. Mirza, Oliver Y. Tang, Ian D. Connolly, Hael A. Abdulrazeq, Rachel K. Lim, G. Dean Roye, Cedric Priebe, Cheryl Chandler, Tiffany J. Libby, Michael W. Groff, John H. Shin, Albert E. Telfeian, Curtis E. Doberstein, Wael F. Asaad, Ziya L. Gokaslan, James Zou, and Rohaid Ali. 2024. Using ChatGPT to Facilitate Truly Informed Medical Consent. *NEJM AI*, 1(2):AIcs2300145. Publisher: Massachusetts Medical Society.

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.

Readability. 2019. Readability 0.3.1. Accessed November 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.

R. J. Senter and E. A. Smith. 1967. Automated readability index. Technical Report AMRL-TR-6620, Wright-Patterson Air Force Base, Ohio, USA.

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.

Oisín Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. MDC at BioLaySumm Task 1: Evaluating GPT Models for Biomedical Lay Summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 611–619, Toronto, Canada. Association for Computational Linguistics.

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.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted Large Language Models Can Outperform Medical Experts in Clinical Text Summarization. *Nature Medicine*, 30(4):1134–1142. ArXiv:2309.07430 [cs].

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. *arXiv preprint*. ArXiv:2305.16739 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

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 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 publishes trusted reviews of biomedical research in two formats: a technical abstract and a plain language summary.	8465 projects (13,922 texts) (*shorter than 250 excluded)	Python libraries: Selenium (2023) (for automated browser interactions) and Beautiful Soup (2023) (for web scraping).
Pfizer Results Plain Language Summaries	Plain	Plain Language Study Results Summaries (RPLS) of Pfizer clinical studies (Pfizer, 2023). Sections containing tables or diagrams were excluded.	125	Specific sections of the 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, provided by sponsors (e.g., AstraZeneca, GSK, Amgen) (Pharma Intelligence UK Limited, 2023).	294	Automatic PDF extraction introduced errors (missing letters, broken words). GPT-3.5 API was used only to correct these errors, ensuring texts matched the original RPLS PDFs.
PLOS + eLife (Luo et al., 2022; Goldsack et al., 2022; Goldsack and Lin, 2025)	Technical and Plain	Dataset from the BioLaySumm competition containing biomedical and life sciences article summaries. We only used the validation sets.	1376 (PLOS) 241 (eLife)	Official data published by Goldsack and Lin (2025)

Table A1: Overview of the data sources used in this study. All texts are available in our GitHub Data Repository[1].

[1] https://github.com/feliperussi/bridging-the-gap-in-health-literacy/tree/main/data_collection_and_processing/Data%20Sources

B Linguistic Features and Readability Indexes

In this study, the readability indexes (items 1–9) were computed using formulas based on variables from Readability (2019), while the linguistic features (items 10–49) were extracted using SpaCy (2023) (model `en_core_web_sm`). The remaining readability features (items 50–62) were obtained with the Readability library. Below is an enumerated list and for a concise overview, Table B1 presents the same variables along with their enumeration.

1. **Flesch-Kincaid Grade Level (FKGL):** Estimates the U.S. school grade level needed to comprehend the text (Flesch, 1948; Kincaid et al., 1975).

2. **Automated Readability Index (ARI):** Computes readability using characters, words, and sentences (Senter and Smith, 1967).

3. **Coleman-Liau Index (CLI):** Measures readability based on letter and word counts per sentence (Coleman and Liau, 1975).

4. **Flesch Reading Ease (FRE):** Produces a score where higher values indicate easier readability (Flesch, 1948; Kincaid et al., 1975).

5. **Gunning Fog Index (GFI):** Estimates the number of years of formal education needed to understand the text (Gunning, 1952).

6. **LIX:** Calculates readability by analyzing the proportion of long words in the text (Anderson, 1983).

7. **SMOG readability formula (SMOGIndex):** Estimates readability by counting polysyllabic (Mc Laughlin, 1969).

8. **RIX:** Computes readability from the number of long words per sentence (Anderson, 1983).

9. **Dale-Chall Readability Score (DCRS):** Assesses readability by comparing text words against a list of familiar words (Chall and Dale, 1995).

10. **total_words:** Total number of words in the text (excluding punctuation), identified by spaCy. e.g., in "Hello, world!", there are 2 words.

11. **total_sentences:** Total number of sentences in the text, based on spaCy's sentence segmentation. e.g., "Hello. World!" yields 2 sentences.

12. **total_characters:** Total number of characters in the text. e.g., "Hello" has 5 characters.

13. **passive_voice:** Frequency of passive voice constructions, determined via verb forms tagged as VBN. e.g., "was given" in "John was given a book by Mary."

14. **active_voice:** Frequency of active voice constructions, counted as verbs (VERB) not tagged as VBN. e.g., "ran" in "Alice quickly ran to the store," or "decided" in "He decided to give up his job."

15. **passive_toks:** Count of tokens in passive constructions, where spaCy marks passive subjects with `nsubjpass`. e.g., "John" in "John was given a book by Mary."

16. **active_toks:** Count of tokens in active constructions, based on the `nsubj` dependency; e.g., "Alice" in "Alice quickly ran to the store."

17. **verbs:** Count of verbs in the text, determined by tokens with the part-of-speech VERB; e.g., "bought" in "Alice bought 3 apples."

18. **nouns:** Count of nouns in the text, determined by tokens with the part-of-speech NOUN; e.g., "book" in "John was given a book."

19. **adjectives:** Count of adjectives in the text, determined by tokens with the part-of-speech ADJ; e.g., "incredible" in "That was incredible."

20. **adverbs:** Count of adverbs in the text, determined by tokens with the part-of-speech ADV; e.g., "quickly" in "Alice quickly ran to the store."

21. **prepositions:** Count of prepositions in the text, determined by tokens with the part-of-speech ADP; e.g., "by" in "the ball was thrown by him."

22. **auxiliaries:** Count of auxiliary verbs in the text, determined by tokens with the part-of-speech AUX; e.g., "was" in "John was given a book by Mary."

Readability Indexes	(1) FKGL, (2) ARI, (3) CLI, (4) FRE, (5) GFI, (6) LIX, (7) SMOGIndex, (8) RIX, (9) DCRS
Linguistic Characteristics	(10) total_words, (11) total_sentences, (12) total_characters, (13) passive_voice, (14) active_voice, (15) passive_toks, (16) active_toks, (17) verbs, (18) nouns, (19) adjectives, (20) adverbs, (21) prepositions, (22) auxiliaries, (23) conjunctions, (24) coord_conjunctions, (25) determiners, (26) numbers, (27) particles, (28) pronouns, (29) proper_nouns, (30) punctuations, (31) subordinating_conjunctions, (32) symbols, (33) other, (34) persons, (35) norp, (36) facilities, (37) organizations, (38) gpe, (39) products, (40) works, (41) dates, (42) times, (43) quantities, (44) ordinals, (45) cardinals, (46) percentages, (47) locations, (48) laws, (49) stopwords (50) characters_per_word, (51) syll_per_word, (52) words_per_sentence, (53) sentences_per_paragraph, (54) type_token_ratio, (55) syllables, (56) paragraphs, (57) long_words, (58) complex_words, (59) complex_words_dc, (60) tobeverb, (61) auxverb, (62) nominalization

Table B1: Variables used to describe the readability and linguistic characteristics of the texts. Items 1–9 (readability indexes) were computed using formulas based on variables from Readability (2019), items 10–49 (linguistic features) were extracted using SpaCy (2023) (model `en_core_web_sm`), and items 50–62 were obtained using Readability (2019).

23. **conjunctions:** Count of conjunctions in the text, determined by tokens tagged as CCONJ or SCONJ; e.g., "because" and "and" in "Alice quickly ran to the store and bought 3 apples because it was late."

24. **coord_conjunctions:** Count of coordinating conjunctions, determined by tokens with the part-of-speech CCONJ; e.g., "and" in the example of conjunctions.

25. **determiners:** Count of determiners in the text, determined by tokens with the part-of-speech DET; e.g., "the" in "the qwerty word is unknown."

26. **numbers:** Count of numerical values in the text, determined by tokens with the part-of-speech NUM; e.g., "3" in "Alice bought 3 apples."

27. **particles:** Count of particles in the text, determined by tokens with the part-of-speech PART; e.g., "to" in "He decided to give up his job."

28. **pronouns:** Count of pronouns in the text, determined by tokens with the part-of-speech PRON; e.g., "him" in "the ball was thrown by him."

29. **proper_nouns:** Count of proper nouns in the text, determined by tokens with the part-of-speech PROPN; e.g., "Google" or "JFK Airport."

30. **punctuations:** Count of punctuation marks in the text, determined by tokens with the part-of-speech PUNCT; e.g., "," in "John was given a book, and the ball was thrown by him."

31. **subordinating_conjunctions:** Count of subordinating conjunctions in the text, determined by tokens with the part-of-speech SCONJ; e.g., "because" in the example of conjunctions.

32. **symbols:** Count of symbols in the text, determined by tokens with the part-of-speech SYM; e.g., "$" in "worth $100,000."

33. **other:** Count of tokens not classified in other categories, determined by tokens with the part-of-speech X (uncategorized).

34. **persons:** Count of person mentions in the text, determined by entities labeled PERSON; e.g., "John" or "Mary."

35. **norp:** Count of references to nationalities, religious or political groups, determined by entities labeled NORP; e.g., "American."

36. **facilities:** Count of facilities (e.g., buildings, airports, roads), determined by entities labeled FAC; e.g., "JFK" and 'Airport" in "JFK Airport."

37. **organizations:** Count of organizations, determined by entities labeled ORG; e.g., "FAA" or "Google."

38. **gpe:** Count of geopolitical entities (countries, cities), determined by entities labeled GPE; e.g., “London.”

39. **products:** Count of products mentioned, determined by entities labeled PRODUCT.

40. **works:** Count of creative works (e.g., art, books, movies), determined by entities labeled WORK_OF_ART; e.g., “Hamlet.”

41. **dates:** Count of dates mentioned, determined by entities labeled DATE; e.g., “March”, “15”, “,” and “2025” in “March 15, 2025.”

42. **times:** Count of time expressions, determined by entities labeled TIME; e.g., “3:00” and “PM” in “3:00 PM.”

43. **quantities:** Count of quantity expressions, determined by entities labeled QUANTITY; e.g., “10” and “kg.” in “10 kg.”

44. **ordinals:** Count of ordinal numbers, determined by entities labeled ORDINAL; e.g., “first” in “She is the first in its field.”

45. **cardinals:** Count of cardinal numbers, determined by entities labeled CARDINAL; e.g., “3” in “Alice bought 3 apples.”

46. **percentages:** Count of percentage expressions, determined by entities labeled PERCENT; e.g., “50” and “%” in “yield 50% discounts.”

47. **locations:** Count of location mentions, determined by entities labeled LOC; e.g., “Alps” in “The Alps are breathtaking.”

48. **laws:** Count of laws mentioned, determined by entities labeled LAW; e.g., “Section” and “2” in “Section 2 of the law applies to this case.”

49. **stopwords:** Count of stopwords in the text, determined by tokens identified as stop words by spaCy; e.g., “was,” “the,” or “and.”

50. **characters_per_word:** Average number of characters per word, computed as total characters divided by total words.

51. **syll_per_word:** Average number of syllables per word, computed as total syllables divided by total words.

52. **words_per_sentence:** Average number of words per sentence, computed as total words divided by total sentences.

53. **sentences_per_paragraph:** Average number of sentences per paragraph, computed as total sentences divided by total paragraphs.

54. **type_token_ratio:** Ratio of unique words to total words, computed as the number of distinct tokens divided by total words.

55. **syllables:** Total number of syllables in the text.

56. **paragraphs:** Total number of paragraphs in the text.

57. **long_words:** Count of long words in the text, defined as words exceeding a specified length threshold (e.g., more than 7 letters).

58. **complex_words:** Count of complex words in the text, defined as words with three or more syllables (e.g., “inconceivable”), indicating text complexity.

59. **complex_words_dc:** Count of complex words according to the Dale–Chall method (i.e., unknown polysyllabic words from a list of basic words).

60. **tobeverb:** Count of occurrences of the verb “to be” in the text (e.g., “is,” “are,” “was”).

61. **auxverb:** Count of auxiliary verbs in the text (e.g., “have,” “will,” “do”).

62. **nominalization:** Count of nominalizations in the text, i.e., instances where verbs, adjectives, or other linguistic elements are transformed into nouns (e.g., “development” from “develop”).

C Prompts

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 C1: Prompt to translate Cochrane technical abstract into a plain language summary.

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 C2: Prompt to translate a protocol into a plain language summary compliant with EU CTR No 536/2014.