Medical Text Simplification: From Jargon Detection to Jargon-Aware Prompting

Taiki Papandreou and Jan Bakker and Jaap Kamps
Institute for Logic, Language and Computation (ILLC)
University of Amsterdam
Amsterdam, The Netherlands
taiki.papandreou-lazos@student.uva.nl, j.bakker@uva.nl, kamps@uva.nl

Abstract

Jargon identification is critical for improving the accessibility of biomedical texts, yet models are often evaluated on isolated datasets, leaving open questions about generalization. After reproducing MedReadMe's jargon detection results and extending evaluation to the PLABA dataset, we find that transfer learning across datasets yields only modest gains, largely due to divergent annotation objectives. Through manual re-annotation we show that aligning labeling schemes improves cross-dataset performance. Building on these findings, we evaluate several jargon-aware prompting strategies for LLM-based medical text simplification. Explicitly highlighting jargon in prompts does not consistently improve simplification quality. When gains occur, they often trade off against readability and are model-dependent. Human evaluation indicates that simple prompting can be as effective as more complex, jargon-aware instructions. We release code to facilitate further research: https://github.com/taikilazos/thesis_codebase.

1 Introduction

Medical text simplification is crucial for improving health literacy by making technical content accessible to lay readers, with jargon handling being a central challenge. In this work, we define jargon as any term or span of words that may be hard to understand for lay readers who are not in the medical domain, such as technical abbreviations or complex terminology requiring simplification. However, jargon detection models are often evaluated on isolated datasets, leaving significant questions about their generalization capabilities.

Recent resources like MedReadMe (Jiang and Xu, 2024) and PLABA (Attal et al., 2023; Ondov et al., 2025) provide valuable benchmarks for jargon-centric research, but they employ different annotation schemes, MedReadMe categorizes term difficulty for lay readers, while PLABA identifies terms requiring simplification via expert adaptations (see Figure 1). This discrepancy creates a fundamental barrier to cross-dataset evaluation and generalization.

PLABA Dataset
We studied 36 drop seizures in 5 patients with myoclonic astatic epilepsy of early childhood (MAEE) with simultaneous split-screen video recording and polygraph. Sixteen were falling attacks and 20 were either less severe attacks exhibiting only deep head nodding or seizures equivalent to drop attacks in terms of ictal pattern but recorded in the supine position. All seizures except those that occurred in patients in the supine position showed sudden momentary head dropping or collapse of the whole body downward.

MedReadMe Dataset
The long-acting bronchodilator tiotropium and single-inhaler combination therapy of inhaled corticosteroids and long-acting beta 2-agonists (ICS/LABA) are commonly used for maintenance treatment of patients with chronic obstructive pulmonary disease (COPD). Combining these treatments, which have different mechanisms of action, may be more effective than administering the individual components.

Figure 1: Example annotations from PLABA and MedReadMe datasets. Underlined terms indicate identified jargon.

To address this, we first reconstruct the MedReadMe experimental setup and extend evaluation to PLABA to probe cross-dataset generalization. Second, we investigate whether explicitly surfacing detected jargon in prompts improves LLM-based simplification of medical abstracts.

Our contributions are:

- We replicate MedReadMe's jargon identifica-

Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025), pages 36–46
November 4-9, 2025

tion and release our implementation, establishing baselines on PLABA and a relabeled subset for cross-dataset evaluation.

- We assess cross-dataset generalization, showing transfer learning is limited by annotation mismatches and that aligning schemes improves performance.
- We introduce and evaluate jargon-aware prompting strategies for simplification, finding benefits are model-dependent and often trade off against readability.

We release code and data to support reproducibility and future work on jargon-aware medical text simplification: https://github.com/taikilazos/thesis_codebase.

2 Related Work

Text simplification aims to make specialized content accessible without sacrificing meaning, a particular challenge in medicine where technical terminology is dense (Agrawal and Carpuat, 2024). Our work connects four areas: lexical complexity detection, biomedical simplification resources, LLM prompting strategies, and evaluation. Lexical complexity and jargon detection have evolved from surface heuristics to contextual models like BERT (Devlin et al., 2019), with MedReadMe providing fine-grained jargon categories for lay reader difficulty analysis (Jiang and Xu, 2024).

Work on biomedical text simplification leverages datasets such as PLABA, which offers expert-authored adaptations with span-level links to technical terms (Attal et al., 2023; Ondov et al., 2025), though Bakker and Kamps (2024) and others highlight challenges in sentence-level alignment (Devaraj et al., 2021; Goldsack et al., 2022; Guo et al., 2024). There has been limited exploration of LLM prompting strategies that explicitly surface jargon to control simplification (Xia et al., 2025). For evaluation, we assess how automatic metrics like FKGL and SARI (Kincaid et al., 1975; Xu et al., 2016) align with human judgments in this jargon-aware setting.

The PLABA dataset has enabled research into more controlled simplification approaches. Notably, Xia et al. (2025) conducted a study on jargon-aware simplification by using detected jargon spans to structure prompts for large language models. Their findings suggest that while explicitly surfacing jargon can be beneficial, its effectiveness is not consistent across models and often comes at the cost of readability, highlighting the complexity of integrating detection with generation.

Dataset	# Sentences	# Jargon
PLABA Training	1,602	2,586
PLABA Validation	178	296
PLABA Testing	4,500	9,126
MedReadMe Training	2,587	5,207
MedReadMe Validation	784	1,789
MedReadMe Testing	1,140	2,112

Table 1: Number of examples and total jargon terms in the PLABA and MedReadMe datasets.

Metric	PLABA	MedReadMe
FKGL	10.73	14.08
Jargon / sent	1.92	1.76
Jargon Length	2.98	3.35

Table 2: Comparison of metrics between the PLABA and MedReadMe datasets.

3 Methodology

3.1 Dataset Analysis

We study two biomedical datasets with distinct objectives and annotation schemes. MedReadMe comprises 4,520 sentences from 180 complex-simple article pairs sampled from 15 medical simplification resources (Guo et al., 2024; Goldsack et al., 2022; Devaraj et al., 2021) and provides a hierarchical jargon taxonomy (binary/3-class/7-class) annotated by non-experts to approximate lay comprehension (Jiang and Xu, 2024). PLABA consists of PubMed abstracts paired with expert-authored plain-language adaptations and marks spans that require simplification (Attal et al., 2023; Ondov et al., 2025). As shown in Table 1, the datasets differ in the number of examples and annotated jargon terms; we preprocess PLABA to a sentence-level format to match MedReadMe. In difficulty characteristics (Table 2), MedReadMe exhibits higher lexical and grammatical complexity (FKGL 14.08 vs. 10.73). PLABA shows slightly higher jargon density but shorter jargon terms. Only 276 jargon terms exactly overlap across datasets, underscoring divergent annotation goals and target audiences. For examples of the differing annotation focus, see Figure 1.

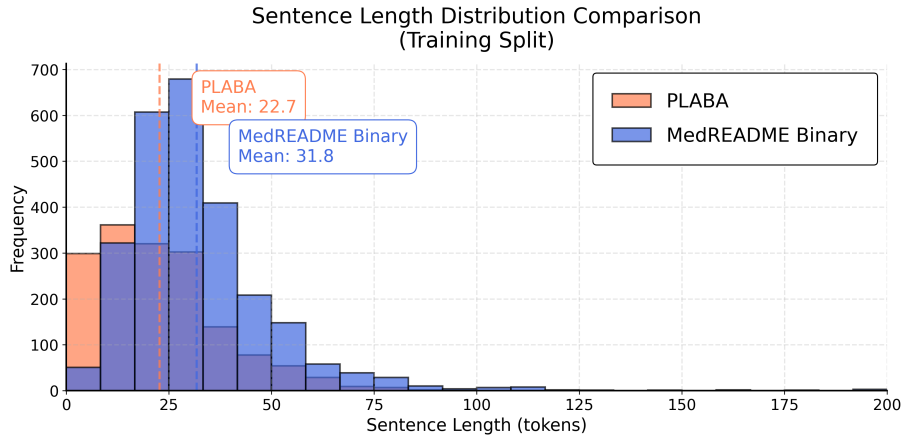

Figure 2: Sentence length distribution for train set: the mean value for MedReadMe is 31.8 and for PLABA 22.7

We also analyzed quantitative metrics to compare the two datasets. Figure 2 illustrates the sentence length distribution in the training splits, with MedReadMe sentences averaging 31.8 tokens compared to PLABA's 22.7 tokens.

3.2 Model Selection and Training

We use standard BIO tagging for span detection: MedReadMe is labeled at binary/3-class/7-class levels, while PLABA is binary-only. Subword tokenization is handled via tokenizer `word_ids()`, assigning B- to the first subword and I- to subsequent subwords; special tokens ([CLS], [SEP], [PAD]) are masked with -100 in the loss. Both datasets are processed at the sentence level with a maximum sequence length of 250 and attention masks to ignore padding.

We reproduced the MedReadMe experiment using BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BioBERT (Lee et al., 2019), and PubMedBERT (Tinn et al., 2023), testing both base and large variants of each model. Since the original paper only referenced the Hugging Face API without specifying model versions for BioBERT and PubMedBERT, our specific choices are documented in Table 3.

We standardized fine-tuning across models: optimizer AdamW (Loshchilov and Hutter, 2019), learning rate 2e-5 (MedReadMe) and 1e-5 (PLABA), batch size 32, up to 20 epochs with early stopping (patience 3) on validation entity-level F1.

3.3 Transfer Learning

We evaluated transferability via two settings: (1) direct transfer (train on MedReadMe $\rightarrow$ evaluate on PLABA; train on PLABA $\rightarrow$ evaluate on MedReadMe) and (2) sequential fine-tuning (PLABA$\rightarrow$MedReadMe evaluated on MedReadMe and PLABA; MedReadMe$\rightarrow$PLABA evaluated on MedReadMe and PLABA).

Family	Type	Model
BERT	Base	bert-base-uncased
RoBERTa	Base	roberta-base
BERT	Large	bert-large-uncased
RoBERTa	Large	roberta-large
BioBERT	Base	biobert-base-v1.1†
PubMedBERT	Base	biomed-base-uncased‡
BioBERT	Large	biobert-large-v1.1†
PubMedBERT	Large	biomed-large-uncased‡

Table 3: HuggingFace models used in experiments: generic model architectures (top half) and biomedical variants (bottom half). †From dmis-lab, ‡From microsoft/BiomedNLP.

All models were based on RoBERTa-large, fine-tuned with a learning rate of 1×10^{-5}, batch size of 16, and early stopping with max epoch of 20. Since MedReadMe contains multi-class annotations, we standardized both datasets to a binary classification setting (jargon vs. non-jargon) for consistency. Performance was primarily measured using entity-level F1.

3.4 Manual Annotation of PLABA Sentences Using the MedReadMe Scheme

To ensure that any observed differences in model performance when transferring between MedReadMe and PLABA are not solely due to mismatches in annotation schemes, we manually re-annotated 100 PLABA sentences using the

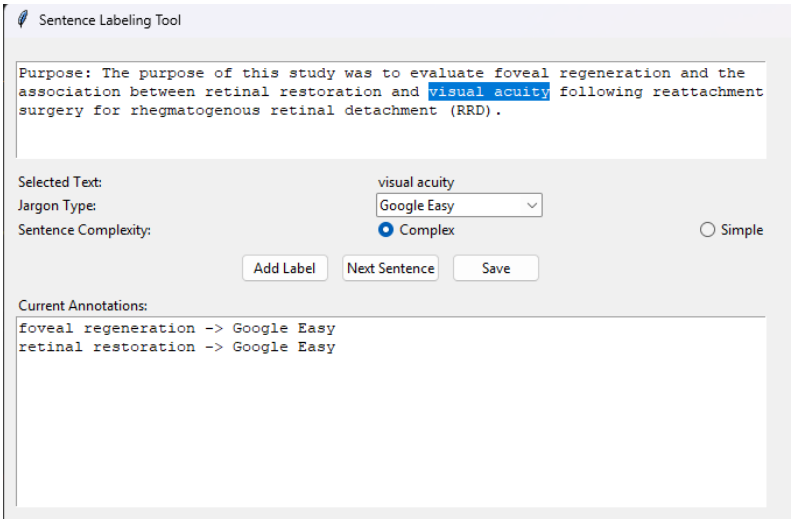

Figure 3: Screenshot of the custom sentence labeling tool. The tool allows the annotator to highlight spans corresponding to jargon terms and assign one of seven MedReadMe classes (e.g., Google Easy, Google Hard, Medical Name, etc.). The annotator can also specify whether a sentence is considered complex or simple, with the latter defined as sentences containing no jargon or only a single Google Easy term.

MedReadMe seven-class taxonomy: Google-Easy, Google-Hard, Medical Named Entity, Medical Abbreviation, General Abbreviation, General Complex Term, and Multi-sense Word (Jiang and Xu, 2024). This approach allows us to directly assess the impact of annotation scheme alignment on model performance.

A custom annotation tool (see Figure 3) was developed to facilitate this process, allowing the annotator to highlight jargon spans and assign the appropriate class. Sentences were also labeled as "complex" or "simple," with "simple" defined as containing no jargon or only a single Google Easy term, and all the other cases as "complex."

All annotations were performed by a single annotator (the main author), following MedReadMe guidelines (Jiang and Xu, 2024).

This relabeled subset allows for a fairer evaluation of model transfer: if model performance improves on the MedReadMe-labeled PLABA data, it suggests that the original drop in transfer performance was primarily due to annotation scheme mismatch and label distribution differences, rather than a fundamental inability of the model to generalize. Sentences were classified as "simple" if they contained no jargon or only a single Google Easy term (aligned with MedReadMe's lower CEFR levels); "complex" sentences included any with additional jargon (e.g., multiple Google Easy or Google Hard/Medical terms), though this feature was not used in classification. The class distribution of the relabeled data can be seen in the Table 4.

Class	Count
Google Easy	203
Google Hard	187
Medical Name	15
Medical Abbreviation	16
General Abbreviation	0
General Complex	16
Multisense	0

Table 4: True label distribution (token-level) for each class in the evaluation set.

3.5 PLABA Test Set for Simplification

The PLABA test set comprises 300 medical abstracts with 3,315 sentences, of which 3,041 (91.7%) contain at least one jargon term (Attal et al., 2023). Sentences contain between 1 and 18 jargon spans, with most (64.6%) containing 1–4. Each span is annotated with one or more recommended simplification actions (e.g., substitute, explain, generalize, omit, exemplify). On average, abstracts contain 26.57 jargon terms. Action distribution is skewed toward substitution (65.62%), followed by explanation (17.59%), omission (10.25%), generalization (6.12%), and exemplification (0.43%). Average jargon length is 1.79 words.

Reference simplifications were 40% shorter at the sentence level (26.18 to 15.94 words) and 6.5 grade levels easier (FKGL 13.55 to 7.04; Kincaid

Original Text:
The patient exhibited tachycardia and dyspnea during examination.

Base Instructions (applied to all prompts):
1) Write a clear sentence; 2) Preserve distinctions and numbers; 3) Replace medical terms only if meaning stays exact; 4) No notes or multiple versions.

1. Simple Prompt: Provide one simplified sentence for the input (focusing on lexical simplification of jargon).

2. Jargon-aware Prompt: Highlight detected terms (e.g., tachycardia, dyspnea) and simplify cautiously.

3. Ground Truth Jargons Prompt (GT): Use ground truth jargon terms as guidance.

4. Ground Truth Actions Prompt (GT action): Specify per-term actions (e.g., substitute, explain).

Figure 4: Prompting strategies for sentence-level medical simplification. See Appendix B for detailed information.

et al., 1975), yet they contained more sentences (19 vs. 11.05), indicating frequent sentence splitting.

3.6 Prompt Design

We evaluate four prompting strategies, from a simple instruction baseline to prompts that explicitly surface jargon terms and, in the most guided variant, specify actions per term. Jargon spans are obtained from our PLABA jargon detector. Simplification operates at the sentence level; simplified sentences are concatenated for abstract-level evaluation.

Ground-truth variants estimate the upper bound of jargon-aware prompting: if gold-guided prompts outperform detected-jargon prompts, the bottleneck lies in detection rather than prompting.

3.7 LLM Models for Simplification

We compare a general-purpose model, Llama-3.1-8B-Instruct[1], with a domain-specialized alternative, Medicine-Llama3-8B[2]. We standardize output cleaning to remove prefixes and meta-commentary, retaining only the simplified sentence for evaluation.

3.8 Evaluation Methods for Simplification

For Jargon Detection tasks, we report F1 due to class imbalance in medical texts. Token-level F1 measures correct classification of individual tokens while ignoring padding tokens (-100) and non-entity (O) predictions. Entity-level F1 requires exact matches between predicted and gold entities in both span boundaries and type. We run generation with fixed decoding settings: temperature = 0.2, top_p = 0.9, repetition penalty = 1.3, no_repeat_ngram_size = 3, and max tokens = 512. Experiments use NVIDIA A100 GPUs; models are loaded from shared storage for throughput. Evaluation is computed at the abstract level by concatenating sentence-level outputs.

As for the simplification task, we report readability (FKGL) (Kincaid et al., 1975) and semantic similarity with BERTScore[3], and we use SARI and BLEU (via EASSE)[4] to assess add/keep/delete operations relative to original and reference.

To validate our automatic metrics, we conducted a human evaluation study[5]. We evaluated at the abstract level rather than sentence level to better reflect real-world reading, where users consume full abstracts; complex sentences aggregated at this level provide a fairer assessment of overall difficulty. Each original medical text was presented alongside five simplified versions: four generated by our models and one gold-standard reference (PLABA's expert-authored simplifications), randomly ordered to prevent bias. We recruited N=5 annotators, all fluent English speakers with at least a graduate-level background in Computer Science-related fields from the University of Amsterdam. Each annotator evaluated three different medical abstracts. For each abstract, they rated the five simplified versions on a 1–5 scale along three aspects: meaning preservation (accuracy of medical information), simplicity (lexical accessibility to non-experts, focusing on jargon reduction), and fluency (natural and coherent writing).

For an action-based perspective, annotators also evaluated the model's ability to perform specific

[1] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[2] https://huggingface.co/instruction-pretrain/medicine-Llama3-8B

[3] https://github.com/feralvam/easse/blob/master/easse/bertscore.py

[4] https://github.com/feralvam/easse

[5] https://qualitativeexpthesis-biomed.streamlit.app/

	Model	Token-Level			Entity-Level		
		Bin	3Cls	7Cls	Bin	3Cls	7Cls
Base	BERT	88.12 (*85.4*)	86.61 (*80.4*)	75.02 (*66.3*)	70.85 (*77.0*)	67.74 (*72.5*)	56.47 (*63.3*)
	RoBERTa	**89.89** (*86.2*)	**88.72** (*81.7*)	**76.72** (*66.7*)	57.83 (*79.7*)	69.29 (*75.2*)	53.23 (*66.6*)
	BioBERT	87.83 (*84.2*)	87.39 (*79.6*)	76.10 (*66.4*)	68.51 (*77.1*)	67.12 (*72.8*)	58.22 (*64.1*)
	PubMedBERT	84.98 (*85.2*)	84.72 (*81.2*)	76.71 (*67.7*)	**71.57** (*75.8*)	**72.50** (*74.8*)	**63.68** (*66.3*)
Large	BERT	88.05 (*86.1*)	87.18 (*80.9*)	76.25 (*67.9*)	67.70 (*78.5*)	68.93 (*74.1*)	58.71 (*43.9*)
	RoBERTa	**89.73** (*86.8*)	**88.72** (*82.3*)	**78.65** (*68.6*)	**73.42** (*80.2*)	68.87 (*75.9*)	**62.63** (*67.9*)
	BioBERT	87.80 (*85.3*)	86.33 (*80.7*)	75.98 (*67.0*)	73.40 (*78.4*)	**70.51** (*72.2*)	60.19 (*64.9*)
	PubMedBERT	86.39 (*85.7*)	85.67 (*82.3*)	75.31 (*68.3*)	72.32 (*79.0*)	69.70 (*75.2*)	61.67 (*66.5*)

Table 5: F1 scores (%) on the MedReadMe dataset. Our results are shown with original results in parentheses. The highest value per column is bolded.

text transformation operations (substitute, generalize) informed by PLABA action annotations (Attal et al., 2023; Ondov et al., 2025). In this setting, each annotator rated three action types across five randomly selected sentences, using the same 1–5 scale. Detailed examples are provided in Appendix B.

4 Results and Analysis

4.1 Jargon Detection Performance

We successfully replicated the MedReadMe (MRM) experiment, though with notable differences. As shown in Table 5, our implementation achieved higher token-level F1 scores (e.g., 89.89% vs. 86.8% for RoBERTa-base) but lower entity-level F1 scores (e.g., 73.42% vs. 80.2% for RoBERTa-large) compared to the original study. This suggests our models were better at classifying individual tokens but worse at identifying exact span boundaries, potentially due to differences in the evaluation pipeline.

Performance varied significantly across jargon categories (Table 6). The RoBERTa model excelled at identifying medical abbreviations (F1=0.869) but struggled with nuanced distinctions, such as differentiating Google-Hard from Google-Easy terms (F1=0.514). It failed completely on rare classes like multisense terms, highlighting the impact of severe class imbalance.

On the PLABA dataset, all models performed worse than on MRM, with RoBERTa-large achieving the highest entity-level F1 of 46.70% (Table 7). Surprisingly, domain-specific models like BioBERT showed no clear advantage. This performance gap is likely due to PLABA's smaller size and, crucially, its different annotation objective.

Class	Prec.	Rec.	F1	Supp.
G_EASY	0.697	0.828	0.756	3,939
G_HARD	0.748	0.391	0.514	1,178
MED_ABBR	0.831	0.910	0.869	933
MED_NAME	0.506	0.701	0.588	455
GEN_CPLX	0.695	0.628	0.660	489
GEN_ABBR	0.866	0.792	0.827	130
MULTI	0.000	0.000	0.000	28

Table 6: 7-class performance for RoBERTa-large on MedReadMe (MRM) dataset. G: Google, MED: Medical, GEN: General, ABBR: Abbreviation, NAME: Name Entity, CPLX: Complex, MULTI: Multisense.

Model	F1	Precision	Recall
BERT	44.17	39.74	49.70
RoBERTa	**46.70**	46.06	47.36
BioBERT	43.42	46.99	40.35
PubMedBERT	45.43	43.19	47.92

Table 7: Entity-level performance metrics across different language models (large version) on the PLABA dataset.

4.2 Transfer Learning and the Impact of Annotation Schema

Cross-dataset transfer learning yielded only modest gains, underscoring the challenge of generalizing across differently annotated resources (Table 8). For instance, a model trained on MedReadMe achieved only 33.71% entity F1 when evaluated directly on PLABA.

To test if this was due to annotation mismatch,

Experiment	Token F1	Entity F1
MRM → PLABA	61.22	33.71
PLABA+MRM → PLABA	62.94	37.01
MRM+PLABA → PLABA	66.84	37.71
PLABA → MRM	59.01	25.03
PLABA+MRM → MRM	**89.80**	**73.84**
MRM+PLABA → MRM	73.96	46.64

Table 8: F1 scores (%) for transfer learning experiments. Sequential transfer refers to fine-tuning on a second dataset after initial training. MRM stands for MedReadMe.

Setting	SARI↑	BERTScore↑	FKGL↓	BLEU↑
Simple	29.87	19.91	**13.53**	2.34
Jargon	29.92	19.07	14.16	2.40
GT	30.62	**20.51**	14.16	3.03
GT action	**32.26**	11.55	15.36	**4.06**

Table 9: Performance metrics for Llama-3.1-8B-Instruct across different prompts.

we manually relabeled a 100-sentence PLABA subset with the MedReadMe schema. When evaluated on this aligned data, the MedReadMe-trained model's performance improved markedly from 33.71% to 42.00% entity F1. This confirms that the performance drop was primarily due to divergent annotation schemes rather than a model limitation. When labels are aligned, models generalize effectively.

4.3 Jargon-Aware Text Simplification

We next investigated whether explicitly highlighting jargon in prompts improves LLM-based simplification. We evaluated four prompting strategies of increasing complexity on both a general-purpose (Llama-3.1, Table 9) and a domain-specialized (Medicine-Llama3, Table 10) model.

The results were model-dependent and revealed a consistent trade-off. For Llama-3.1, more explicit guidance (e.g., providing ground-truth actions) led to the best performance on operation-based metrics like SARI (32.26) but at the cost of readability, yielding the highest FKGL (15.36). In contrast, the simple prompt achieved the best readability (FKGL=13.53).

Contrary to expectations, the Medicine-Llama3 model performed best across all metrics with the simple prompt and its performance degraded with more complex, jargon-aware instructions. This suggests that domain-specific pre-training does not automatically translate into an ability to effectively leverage explicit jargon instructions.

Setting	SARI↑	BERTScore↑	FKGL↓	BLEU↑
Simple	**28.81**	**13.59**	**12.69**	**1.69**
Jargon	28.64	12.66	12.76	1.65
GT	28.70	11.00	13.67	1.60
GT action	28.57	8.90	13.87	1.42

Table 10: Performance metrics for Medicine-Llama3-8B across different prompts.

Version/Task Type	Mean Rating	SD
Ground Truth	5.00	0.00
Simple	3.33	1.41
Jargon	2.34	0.94
GT	2.84	1.18
GT action	2.67	0.94
Action-based Tasks	4.60	0.57

Table 11: Average Ratings and Standard Deviations by Version and Task Type

4.4 Human Evaluation

A qualitative human evaluation (Table 11) revealed that while the reference simplifications received perfect scores, all model outputs were perceived as lower quality. The simple prompt was competitive (Mean=3.33), while jargon-aware prompts did not reliably improve perceived quality. Notably, the high standard deviations indicate substantial disagreement among raters. A key observation was that sentence-level simplification often led to a loss of context and information across the abstract, limiting overall coherence. See Appendix C for examples of generated outputs for different prompting strategies.

5 Discussion and Conclusions

In this work, we thoroughly evaluated the automatic jargon detection methods for biomedical texts. We reproduced MedReadMe experiments, established PLABA baselines, and showed that cross-dataset transfer is limited primarily by annotation mismatches. We then experimented with jargon-aware prompting strategies for the automatic simplification of these texts.

On jargon detection, our replications achieved

higher token-level but lower entity-level F1 than the original report, highlighting remaining challenges in precise span boundary modeling. Category-wise analyses showed strong performance on frequent, well-formed classes (e.g., medical abbreviations) and weaknesses on rarer or nuanced classes (e.g., Google-Hard, multisense), reflecting severe class imbalance. Importantly, evaluating on a PLABA subset re-annotated with the MedReadMe scheme (100 sentences) improved entity-level F1 from 33.71% to 42.00%, demonstrating that schema alignment substantially boosts transferability.

Turning to simplification, our experiments show that the effect of jargon-aware prompting is model-dependent rather than uniformly beneficial. The general-purpose Llama-3.1-8B-Instruct benefited from more explicit guidance (best SARI/BLEU with ground-truth actions), but with reduced readability (higher FKGL). In contrast, the domain-specialized Medicine-Llama3-8B performed best with simple prompts, suggesting that domain pretraining does not automatically translate into better handling of explicit jargon instructions. This dependency may arise from how models process prompts: general models require explicit jargon surfacing to prioritize medical terms, while specialized models implicitly handle them, making simple instructions sufficient.

Across models, we observed a consistent trade-off: more detailed prompting can improve operation- and overlap-based metrics (SARI, BLEU) while harming readability (FKGL). Qualitative judgments echoed this tension: references set a clear upper bound; simple prompts were competitive, whereas jargon-aware prompts did not reliably improve perceived quality, and sentence-level processing likely contributed to information loss across abstracts.

Thus, explicitly including identified jargon in prompts does not consistently improve LLM medical text simplification. Jargon matters, but surfacing terms alone is insufficient; benefits depend on the model and come with readability trade-offs.

Future work should (i) improve span boundary modeling and mitigate class imbalance in detection, (ii) explore schema-aware or multi-task training for cross-dataset robustness, and (iii) couple detection with controllable, document-level generation and evaluation that jointly captures medical fidelity and accessibility. We release code and data to support further research.

6 Limitations

The lack of multiple annotators for manual re-annotation is a limitation and should be addressed in future work to improve reliability. The small-scale human evaluation (only 5 annotators from a computer science background, evaluating just 3 abstracts each) and resulting high rater disagreement may limit generalizability of perceived quality. Automatic metrics like SARI may not fully capture jargon-specific changes, and the low BLEU scores in simplification experiments indicate challenges in generating high-quality outputs. While the evaluation was at the document level, generation was at the sentence level, losing the global context. Future work could explore better prompt engineering, larger-scale evaluations with diverse annotators, and document-level generation to address these issues.

7 Lay Summary

Medical texts are full of complex terms that can confuse people without a scientific background. This makes it hard for patients and the general public to understand health information. Our research focuses on two key areas: identifying these difficult terms (called "jargon") and simplifying medical texts so they're easier to read.

First, we studied how well computer models can spot jargon in medical writing. We compared two datasets: MedReadMe, which labels terms by how hard they are for lay people to understand, and PLABA, which marks terms that experts think need simplifying. We found that models trained on one dataset don't work as well on the other because the datasets have different goals. But when we manually relabeled some PLABA data to match MedReadMe's style, the models improved a lot, showing that aligning how we define jargon helps cross-dataset learning.

Second, we tested ways to make large language models (like AI chatbots) simplify medical texts. We tried simple prompts and more complex ones that highlight detected jargon. Surprisingly, the simple prompts often worked just as well or better than the jargon-focused ones. Results depended on the model—general-purpose models liked more guidance, but specialized medical models did better with basics. This suggests that just telling an AI to simplify might be enough, without needing to point out every jargon term.

Our work shows that making medical info ac-

cessible is tricky, but better data alignment and smarter prompting can help. We hope this leads to tools that make health communication clearer for everyone, improving patient understanding and outcomes. All our code and data are publicly available to support future research.

Acknowledgments

Experiments in this paper were carried out on the National Supercomputer Snellius, supported by SURF and the HPC Board of the University of Amsterdam. Jan Bakker and Jaap Kamps are partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is also partly funded by the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). We thank the annotators for their contributions to the human evaluation study. Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

References

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10(1):8.

Jan Bakker and Jaap Kamps. 2024. Cochrane-auto: An aligned dataset for the simplification of biomedical abstracts. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 41–51, Miami, Florida, USA. Association for Computational Linguistics.

Jan Bakker, Taiki Papandreou-Lazos, and Jaap Kamps. 2024. Biomedical text simplification models trained on aligned abstracts and lay summaries. In *The Thirty-Third Text REtrieval Conference Proceedings (TREC 2024), Gaithersburg, MD, USA, November 15-18, 2024*, volume 1329 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Chao Jiang and Wei Xu. 2024. MedReadMe: A systematic study for fine-grained sentence readability in medical domain. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17293–17319, Miami, Florida, USA. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, 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. Technical report, Naval Technical Training Command Millington TN Research Branch.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Brian Ondov, William Xia, Kush Attal, Ishita Unde, Jerry He, and Dina Demner-Fushman. 2025. Lessons from the trec plain language adaptation of biomedical abstracts (plaba) track. *arXiv preprint arXiv:2507.14096*.

Taiki Papandreou, Jan Bakker, and Jaap Kamps. 2025. University of Amsterdam at the CLEF 2025 SimpleText Track. In *Working Notes of CLEF 2025:*

Conference and Labs of the Evaluation Forum, volume 4038 of *CEUR Workshop Proceedings*, pages 4356–4362. CEUR-WS.org.

Panagiotis Taiki Papandreou-Lazos. 2025. Medical text simplification: From jargon detection to automated simplification. Master's thesis, University of Amsterdam.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2023. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4):100729.

William Xia, Ishita Unde, Brian David Ondov, and Dina Demner-Fushman. 2025. JEBS: A fine-grained biomedical lexical simplification task. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17654–17666, Vienna, Austria. Association for Computational Linguistics.

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

A Data, code, and models

We release code and data to support reproducibility and future work on jargon-aware medical text simplification: `https://github.com/taikilazos/thesis_codebase`.

Extensive further documentation can be found in (Papandreou-Lazos, 2025).

Related experiments were reported at the TREC 2024 PLABA track (Bakker et al., 2024) and at the CLEF 2025 SimpleText Track (Papandreou et al., 2025).

B Prompt Design

Original Text

```
The patient exhibited tachycardia and dyspnea
↪  during examination.
```

Base Instructions (applied to all prompts)

```
IMPORTANT: Follow these rules exactly:
1. Write a clear sentence
2. Keep ALL medical distinctions and patterns
3. Keep exact numbers and measurements
4. Replace medical terms with plain words ONLY if
↪  meaning stays exactly the same
5. Keep medical terms if simplifying would lose
↪  precision
6. No explanations or notes
7. No multiple versions
```

1. Simple Prompt

```
Text to simplify: The patient exhibited
↪  tachycardia and dyspnea during examination.
Write one simplified sentence.
```

2. Jargon-aware Prompt

```
Replace these terms ONLY if you can keep their
↪  exact medical meaning:
- tachycardia
- dyspnea
Text to simplify: The patient exhibited
↪  tachycardia and dyspnea during examination.
Write one simplified sentence.
```

3. Ground Truth Jargons Prompt

```
You must simplify the following terms in the
↪  sentence below (if present):
- tachycardia
- dyspnea
Text to simplify: The patient exhibited
↪  tachycardia and dyspnea during examination.
Write one simplified sentence.
```

4. Ground Truth Actions Prompt

```
For each marked term in the sentence below, take
↪  the specified action:
- tachycardia: substitute
- dyspnea: explain
Text to simplify: The patient exhibited
↪  tachycardia and dyspnea during examination.
Write one simplified sentence:
```

C Generated Text Examples

To illustrate the differences in simplification quality across prompting strategies, we provide excerpts from one test abstract. We selected two key sentences with medical jargon and show how each version handles them. Figure 5 shows the source, reference (Ground truth human simplification), and prediction based on either the Simple prompt, the Jargon-aware prompt, the Ground truth jargons prompt, and the Ground truth actions prompt

These examples highlight how simpler prompts can be fluent but sometimes inaccurate, while jargon-aware prompts vary in readability and fidelity.

Original Sentence 1

Background: The purpose of the study was to examine whether early repeated infections due to Trichomonas vaginalis among human immunuodeficiency virus (HIV)-positive and HIV-negative women are reinfections, new infections, or cases of treatment failure.

Ground Truth human simplification

For the study's background, the study's aim was to find whether early repeated infections of Trichomonas vaginalis (a parasite that can cause a sexually transmitted infection or STI) among women with human immunuodeficiency virus (HIV) and without HIV are reinfections, new infections, or cases of treatment failing.

Simple prompt

The goal of this research is to figure out what happens when many people get infected again by an STI called trichomoniasis in both those who have AIDS and do not have it.

Jargon-aware prompt

The goal is to figure out what happens when people get infected again by this type of bacteria called trachomona's vaginallis, which can happen even after they have gotten better because either it came back through there sexual partener as a re-infection from them,or got another infeciton on its own, or maybe the cure didnt work for some reason so now we need to find out why.

Ground Truth Jargons Prompt

The goal of this research is to determine why some people infected by Trichamonias get reinfected again after being treated for it.

Ground Truth Actions Prompt

The goal of this research is to determine what causes frequent repeat occurrences of an infection caused by Trichamonias vaginallis among both HIV-positive and negative females - either they get it again after being treated previously, contract it for the first time despite previous treatments, or their current condition does not respond well to medication.

Original Sentence 1

Methods: Women attending an HIV outpatient clinic and a family planning clinic in New Orleans, Louisiana, who had culture results positive for T. vaginalis were treated with 2 g of metronidazole under directly observed therapy.

Ground Truth human simplification

For the study's methods, women attending an HIV clinic and family planning clinic in New Orleans, Louisiana, who had tests detecting T. vaginalis, were treated with 2 grams (g) of metronidazole (an antibiotic for infections) under directly observed treatment.

Simple prompt

Women from two clinics received treatment at this place because they tested positive for Trichomonas vaginalis using cultures; each woman was given 2000 milligrams of Metronidazol while being watched by healthcare workers during their visit.

Jargon-aware prompt

Women from clinics attended by people living with AIDS and those seeking birth control services received treatment after testing positive for a parasite causing STDs; this included taking two grams of antibiotics while being watched as you swallow them.

Ground Truth Jargons Prompt

Women from two clinics received treatment after testing positive for Trichomonas infection; they took 2000 milligrams of antibiotic medication while being watched by healthcare workers during their visits.

Ground Truth Actions Prompt

Women from two clinics in New York received treatment after testing positive for Trichomonas infection; they took 2000 milligrams of antibiotic medication while being closely monitored by healthcare workers during their visits.

Figure 5: Examples: Effect of prompt variation on the output.