A Framework for Fine-Grained Complexity Control in Health Answer Generation

Anonymous ACL submission

Abstract

Effective communication of health information requires adapting complexity to match the target audience's literacy level. However, manually simplifying medical content is both time-consuming and difficult to scale. To address this challenge, we developed a new framework for automatically generating health answers at multiple complexity levels.

We began by collecting 166 linguistic features to quantify text complexity, including traditional readability metrics (e.g., Flesch-Kincaid, SMOG), medical terminology usage (e.g., UMLS coverage, medical entity recognition), syntactic complexity, semantic coherence, and LLM-based measures (e.g., masked language modeling, LLM-as-a-judge). Applying these features to a custom dataset of parallel health texts and external medical benchmarks, we used feature selection to identify 13 key metrics that reliably distinguish between simple and complex text pairs. From these, we derived a complexity scoring formula by combining the metrics with weights learned from a logistic regression model.

Using this formula, we created a large multi-level dataset of health question-answer pairs, ranging from elementary patient-friendly explanations to advanced technical summaries. The initial QA pairs came from established datasets including LiveQA, MedicationQA, and MEDIQA-AnS. We then used LLaMA-based language models with carefully engineered prompts to transform the original answers into five different versions ordered by complexity. Finally, we fine-tuned a large language model on this dataset, incorporating special tokens to control the complexity of the generated text. The resulting model can generate health answers at fine-grained complexity levels, allowing users to select the desired level of detail and technicality.

1 Introduction

Health literacy, which is the ability to obtain, process, and understand basic health information, remains a significant challenge worldwide. A survey conducted by the World Health Organization (WHO) between 2019 and 2021 across 17 European countries found that between 25% and 75% of people struggle with understanding health-related information, with variation depending on country-specific factors like education and healthcare access (Pelikan et al., 2021).

In the United States, approximately 80 million adults had limited health literacy as of 2018, with disproportionately higher rates among older adults, minority groups, and individuals of lower socioeconomic status (Woods et al., 2023). These statistics matter because people with lower health literacy often struggle to understand medical terms, leading to poorer health outcomes and increased healthcare costs (Shahid et al., 2022). This issue becomes even more important as more people turn to online sources for health information. In 2022, 58.5% of U.S. adults searched for health information online (Wang and Cohen, 2022), yet studies show that most health-related content online exceeds recommended readability levels (Szmuda et al., 2020; Mohile et al., 2023).

Large language models (LLMs) like GPT-4 (OpenAI, 2023), Med-PaLM (Singhal et al., 2023), and Claude (Anthropic, 2024) now generate health information and are increasingly used in healthcare contexts. However, these models typically produce text at a fixed complexity level, often too advanced for many readers (Amin et al., 2024). Current approaches to medical text simplification focus on converting complex text into simpler versions (Gondy et al., 2018; Flores et al., 2023; Li et al., 2024) rather than dynamically adjusting complexity based on individual needs.

This gap presents an opportunity to develop lan-

guage models that can generate health answers with adjustable complexity levels, a capability that would make information more accessible to everyone, regardless of their health literacy level.

2 Related Work

2.1 Text Complexity and Readability Assessment

The earliest attempts to measure text complexity used simple formulas based on surface-level features. Smith and Senter (1967) developed the Automated Readability Index (ARI), which counts characters per word and sentence length to estimate reading difficulty. Shortly after, Kincaid et al. (1975) created the Flesch-Kincaid Grade Level formula, which also considers syllable counts and remains widely used today for its simplicity and reliability.

Zheng and Yu (2018) noted that standard formulas failed to capture medical complexity because they ignored specialized terminology and semantic relationships. They developed a ranking system that compared documents relative to each other rather than assigning absolute scores, using both surface-level features and word embeddings to better match human judgments of readability.

Jiang and Xu (2024) created MedReadMe, manually annotating 4,520 medical sentences with readability labels and identifying complex spans within each sentence. They introduced "Google-Easy" and "Google-Hard" categories based on how commonly terms appear in web searches. Their analysis of 650 linguistic features revealed that medical jargon density and syntactic complexity were the strongest predictors of reading difficulty.

Devaraj et al. (2021) proposed using a masked language model (MLM) to differentiate technical and lay medical text. Their method evaluates how accurately a model trained on scientific literature predicts masked tokens, based on the observation that technical terminology is more predictable within domain-specific contexts. Luo et al. (2022) improved this method by focusing on noun phrases, allowing multi-word medical terms like "heart attack" to be treated as single semantic units.

While methods based on masked language modeling have shown promise, they mainly focus on single-word complexity. Lyu and Pergola (2024) addressed this limitation with SciGisPy, a metric rooted in Fuzzy-Trace Theory (FTT) (Reyna, 2012) that evaluates how well simplified texts preserve the core meaning (gist), emphasizing semantic coherence and the ability to form clear mental models.

2.2 Medical Text Simplification

Medical text simplification started with straightforward rule-based systems. For instance, Damay et al. (2006) used techniques like lexical substitution and sentence restructuring to make medical texts easier to understand. Later, Kandula et al. (2010) took this further by combining both semantic and syntactic methods to simplify electronic medical records and patient education materials.

The field progressed significantly with the development of large-scale datasets for training language models. Devaraj et al. (2021) created the Cochrane dataset, which pairs technical abstracts with lay summaries from the Cochrane Database of Systematic Reviews. Using this parallel data, they trained BART models with unlikelihood training, explicitly penalizing the generation of tokens identified as technical language through a bag-of-words classifier. Flores et al. (2023) replaced the bag-of-words classifier with the Flesch-Kincaid readability formula to identify and penalize complex words. To prevent hallucinations that can occur when optimizing solely for simplicity, they also incorporated factual consistency into their loss function and designed a beam search method that weighs both readability and accuracy during decoding.

Basu et al. (2023) created Med-EASi, a finely annotated dataset for simplifying medical texts that identifies four types of textual transformations: elaboration, replacement, deletion, and insertion. With this dataset, they built T5-based models that allow users to select specific medical terms and control exactly how they should be simplified.

Lu et al. (2023) developed NapSS, a two-stage "summarize-then-simplify" method for medical text simplification that first identifies important sentences using a summarizer trained on paired technical abstracts and their human-simplified versions, and then extracts key phrases to create "narrative prompts" that guide the language model during the simplification process, helping preserve the logical flow and medical accuracy of the original text.

Phatak et al. (2022) applied reinforcement learning to medical text simplification by designing reward functions that balance content preservation, Flesch-Kincaid readability scores, and lexical simplicity. Rahman et al. (2024) later created SimpleDC, a dataset of original and simplified texts related to digestive cancers. They fine-tuned LLaMA

models on this dataset and further improved them using reinforcement learning, guided by a binary classifier trained to detect simple language.

2.3 Controllable Text Generation

Recent research has explored ways to control text readability during generation. Ribeiro et al. (2023) developed methods for controllable summarization using instruction-based prompting, reinforcement learning with a Gaussian reward function that penalizes deviations from desired readability scores, and lookahead decoding to anticipate how word choices impact readability.

Luo et al. (2022) focused on readability control specifically for biomedical text summarization. They first tried prepending special tokens as prompts to the input and then tested a multi-head architecture with separate decoders for different readability levels. While the multi-head approach helped create some distinction between technical and plain language outputs, they found that the level of readability control was still very limited.

Tran et al. (2024) introduced ReadCtrl, which instruction-tunes language models to generate text at specific readability scores on an almost continuous scale rather than predefined categories. Meanwhile, Hsu et al. (2024) found that even with clear instructions, language models often produce outputs that do not align with traditional readability metrics. They also showed that readers generally preferred explanations written at high school level, suggesting that there may be a sweet spot of complexity balancing clarity and informative content.

While prior work has focused primarily on binary simplification or relied on traditional readability metrics that fail to capture the unique challenges of medical terminology, we developed a more comprehensive framework that integrates multiple linguistic features to accurately measure the complexity of medical text and generate content at precisely targeted readability levels.

3 Methods

3.1 Data Collection

We used two established datasets containing paired original and simplified medical texts. Though these datasets provide parallel texts at different complexity levels, the "simplified" versions, while less complex than the originals, are not always simple in absolute terms. This relative simplification creates a sliding scale rather than distinct complexity levels, making it difficult to develop a reliable readability formula. To overcome this limitation, we created a synthetic dataset containing pairs of clearly differentiated simple and complex medical texts.

3.1.1 Medical Text Simplification Datasets

We evaluated our metrics using two parallel corpora of medical texts: PLABA (Attal et al., 2023) and Cochrane (Devaraj et al., 2021). Both datasets include original medical texts paired with simplified versions. PLABA contains sentence and paragraph-level simplifications of biomedical abstracts, while Cochrane focuses on paragraph-level simplifications of systematic reviews. More detailed descriptions are available in Appendix A.1.

3.1.2 Claude Dataset

We created a new dataset using Claude 3.5 Sonnet to generate answers to questions from the HealthSearchQA dataset (Singhal et al., 2023), which contains 3,173 commonly searched consumer medical queries. We manually identified and filtered out questions that were not genuinely health-related to ensure the quality and relevance of our dataset. For each valid question, we prompted the model to produce one answer using technical medical language suitable for healthcare professionals, and another using simple language appropriate for patients with limited health literacy. This approach provided clearly differentiated examples of simple and complex medical text covering the same information content.

Table 1 summarizes the key characteristics of the three datasets used in this stage of the project.

Dataset	Source	Size
PLABA-sent	PubMed abstracts	7,643 pairs
PLABA-para	PubMed abstracts	750 pairs
Cochrane	Systematic Reviews Database	4,459 pairs
Claude	HealthSearchQA questions	3,150 pairs

Table 1: Parallel datasets used for text complexity analysis.

3.2 Metrics

We implemented 166 metrics to measure text readability and complexity, covering various linguistic dimensions. The following sections describe each category of metric we used in our analysis.

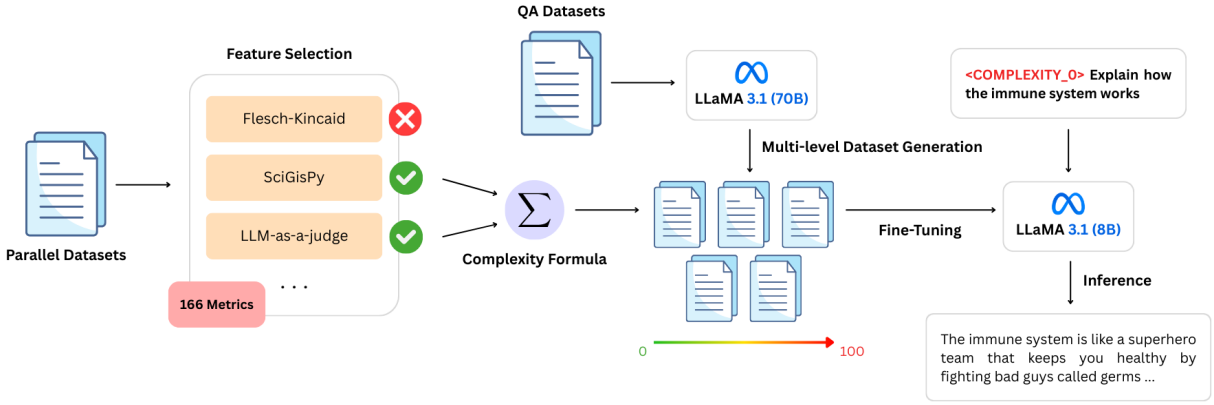

Figure 1: Framework for complexity-controlled health answer generation.

3.2.1 Traditional Metrics

We calculated 20 traditional readability formulas, including Flesch-Kincaid Grade Level (Kincaid et al., 1975), SMOG Index (McLaughlin, 1969), and Coleman-Liau Index (Coleman and Liau, 1975). These metrics estimate text difficulty based on surface-level features like word length, syllable count, and sentence length, working on the general assumption that longer lexical units require more cognitive effort, thereby making the text more complex (Yu et al., 2020). Although not designed for biomedical literature, they can serve as a useful starting point to judge how easy or difficult a text is to read and understand. We supplemented these with 8 statistical measures capturing additional aspects of readability, including the proportion of difficult words from the Dale-Chall list (Dale and Chall, 1948) and lexical diversity metrics such as TTR and MTLD (McCarthy and Jarvis, 2010).

3.2.2 Syntactic Structure

We implemented 16 syntax-based metrics using spaCy (Honnibal et al., 2020) for dependency parsing and part-of-speech tagging, organized into two categories. For lexical distribution, we calculated content-to-function word ratio, which compares meaning-carrying words to grammatical words (Just and Carpenter, 1992), and part-of-speech distributions to identify texts with higher noun density typical of scientific writing (Biber et al., 1999). For structural complexity, we measured dependency distance (Gibson, 2000), passive voice proportion (Ferreira, 2003), noun phrase length (Biber et al., 1999), embedding depth (Gibson, 1998), negation density, and left-right asymmetry (Hawkins, 2004). These metrics capture aspects of syntactic complexity that increase cognitive load, such as deeply embedded clauses and words separated from their grammatical dependents.

3.2.3 Medical Terminology and Jargon

We implemented 19 term-level metrics using the Unified Medical Language System (UMLS) Metathesaurus (National Library of Medicine, 2024) and Consumer Health Vocabulary (CHV) (Zeng and Tse, 2006). For concept identification, we used QuickUMLS (Soldaini, 2016), which performs faster approximate dictionary matching compared to MetaMap (Aronson and Lang, 2010). These metrics include term density, expert-to-lay ratio, semantic type diversity, and CHV familiarity scores that measure how frequently terms appear in consumer health materials (Keselman et al., 2007).

We also built a RoBERTa-large (Liu et al., 2019) sequence tagger with Conditional Random Fields (CRF), trained on the MedReadMe dataset to identify seven distinct categories of medical jargon as defined by Jiang and Xu (2024). These categories include easy and hard medical terms, medical entities, complex terms, multisense words, and medical and general abbreviations. This method enables more fine-grained analysis than dictionary lookups, capturing context-dependent terminology and terms absent from UMLS. From this, we derived 29 other metrics capturing jargon density, distribution across categories, and clustering patterns.

3.2.4 Gist Formation

We adapted GisPy (Hosseini et al., 2022), an open-source tool based on Fuzzy-Trace Theory (Reyna, 2012), which measures how easily readers can understand the essential meaning of a text. GisPy calculates scores for several components that contribute to gist formation, including referential cohesion (connecting ideas between sentences), corefer-

ence resolution (tracking entities throughout text), deep cohesion (presence of causal connectives), and semantic verb overlap (relatedness of actions). We modified the original implementation to use BioSimCSE-BioLinkBERT-BASE (raj Kanakarajan et al., 2022), trained on biomedical literature, making it more suitable for our task. We also implemented SciGisPy (Lyu and Pergola, 2024), which tailors GisPy for biomedical text simplification. SciGisPy introduces domain-specific improvements, such as information content measures derived from biomedical corpora and semantic chunking to measure topic cohesion.

3.2.5 Masked Language Model

We implemented three MLM-based metrics using Bio+ClinicalBERT (Alsentzer et al., 2019), which outperformed other BERT variants in our tests. These metrics measure complexity by calculating how predictable medical terminology is within context. The first metric randomly masks 15% of tokens, the second specifically targets noun phrases, and the third applies a ranking method (RNPTC), which weighs phrases based on their prediction probability (Luo et al., 2022). We found that increasing the number of random masking iterations from 10 to 30 significantly improved reliability by reducing variance. As a result, the simpler random masking approach became more effective than the other two methods in distinguishing between technical and simplified texts.

3.2.6 Semantic Clustering

We built on the method introduced by Cha et al. (2017), which uses word embeddings to measure text complexity. In our implementation, each word is mapped to a BioWordVec embedding (Zhang et al., 2019), and these vectors are grouped using K-means clustering. While the original implementation used 100 clusters, we increased this to 300 to better reflect the distinctions in medical vocabulary. We then create a count vector for how often words fall into each cluster, which serves as a feature vector for predicting readability. We trained two separate Support Vector Regression (SVM) models, one using the CLEAR corpus (Crossley et al., 2023), and another using the MedReadMe dataset (Jiang and Xu, 2024) for medical texts.

3.2.7 ALBERT Transformer

We used the ALBERT-xxlarge model (Lan et al., 2019) from the winning entry in the CommonLit Readability Prize Kaggle competition (Malatinszky et al., 2021). This model processes text through attention layers to capture relationships between words before predicting a readability score. Although the original solution used an ensemble of models, ALBERT-xxlarge was singled out by the winner as especially important, thanks to its parameter-sharing structure, which helps prevent overfitting while still capturing complex language features. The same model was later reused in the REFeREE framework for evaluating text simplification (Huang and Kochmar, 2024).

3.2.8 LLM Expert Evaluation

We created a hybrid method for evaluating text readability using large language models as expert evaluators. Specifically, we prompted three 70 billion-parameter models (Nvidia-Llama-3.1-Nemotron-70B (Wang et al., 2024), Llama3-OpenBioLLM-70B (Pal, 2024), and DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI, 2025)) to evaluate texts on five dimensions: vocabulary complexity, syntactic complexity, conceptual density, required background knowledge, and overall cognitive load. Each model rated texts on a 1–5 scale using few-shot prompting with three calibration examples that we personally annotated. Because running multiple large models is computationally expensive, we trained a smaller and more efficient BioSimCSE-BioLinkBERT-BASE model (raj Kanakarajan et al., 2022) on the averaged LLM scores. This distilled model not only processes texts much faster, but also improves the results by smoothing out inconsistencies in the original LLM judgements.

3.3 Formula Development

After collecting and implementing the linguistic features, we followed a systematic approach to select the most reliable features for our complexity formula. Since we lacked human-annotated readability scores, we developed a data-driven methodology to identify stable features that consistently distinguished simple from expert-level medical texts, using the datasets described in Section 3.1.

The feature selection process began by removing features with absolute pairwise correlations above 0.7 to reduce collinearity and lower the risk of unintentionally excluding important features from the final model. We then applied Lasso logistic regression with bootstrapping, adapting the methodology described by Laurin et al. (2016), which involved the following steps:

1. Creating 1,000 bootstrap samples from our training data using random sampling with replacement.
2. Fitting a Lasso logistic regression model to each bootstrap sample to classify if a text was written for experts or general audience.
3. Calculating the coefficient of variation (CV) for each feature, defined as the standard deviation divided by the mean absolute value of the coefficient, across bootstrap samples.
4. Using the interquartile range (IQR) method to exclude features with unstable coefficients by calculating the upper fence (Q3 + 1.5 × IQR). Features with CV exceeding this threshold were considered outliers and removed.
5. Further filtering features if the 95% confidence interval for the value of the coefficient included zero.

We then trained our final logistic regression model using only the Claude dataset, which contains controlled comparisons of text complexity with a cleaner signal-to-noise ratio. For this purpose, we used ElasticNet regularization to estimate feature weights, as it balances the benefits of both Lasso and Ridge regression and better handles any remaining collinearity among features. This process resulted in a final set of 13 metrics (listed in Appendix A.2) after excluding those that performed exceptionally well in one dataset but poorly or inconsistently in others. These features were likely overfitting to specific data characteristics and were removed to improve generalizability.

3.4 Multi-Level Dataset

After developing and validating our complexity formula, we created a medical dataset containing answers rewritten at multiple levels of complexity to train our controlled text generation model.

3.4.1 Source Datasets

We built our dataset using question-answer pairs from five established medical datasets: LiveQA (Abacha et al., 2017), MedicationQA (Abacha et al., 2019), MEDIQA-AnS (Savery et al., 2020), MedQuAD (Abacha and Demner-Fushman, 2019), and BioASQ Task 13B. After cleaning and filtering for quality, we retained 31,917 question-answer pairs. Table 2 provides a brief overview of these datasets, with detailed descriptions available in Appendix A.3.

Dataset	Source	Size
LiveQA	U.S. NLM	800 pairs
MedicationQA	NIH websites	690 pairs
MEDIQA-AnS	CHiQA-retrieved passages	312 pairs
MedQuAD	NIH websites	16,423 pairs
BioASQ	PubMed/MEDLINE articles	13,692 pairs

Table 2: Source datasets used to create our multi-level medical QA dataset

3.4.2 Dataset Creation

For each question-answer pair in our source datasets, we created five versions of the answer, each written for a different audience, namely young children, middle school students, high school students, college graduates, and biomedical experts. We generated these answers using the models described in Section 3.2.8, with DeepSeek handling 70% of the generation, Nemotron 20%, and OpenBioLLM 10%. This allocation was based on preliminary experiments, which showed that using multiple models helped capture a broader range of writing styles for each education level.

We designed a prompt that generated all five variants simultaneously, with answers becoming progressively more complex (see Appendix A.5). The prompt included three examples to guide the models, descriptions of each target audience, and instructions to keep the answers factually accurate. It also instructed the models to flag any cases where the original answer did not fully address the question, allowing us to filter out problematic samples from the dataset early on.

After generating the variants, we checked the quality of all answers through a two-stage process. First, we used regex patterns to identify and remove samples containing placeholder text instead of proper content. Then we evaluated each variant against its original answer using metrics for content preservation and factual accuracy, including ROUGE (Lin, 2004), BLEURT (Sellam et al., 2020), BERTScore (Zhang et al., 2019), UniEval (Zhong et al., 2022), and SummaC (Laban et al., 2022). The filtering identified relatively few problems and only 2,926 samples (1.56%) were removed from the initial 187,769. This low rejection rate was not surprising, since the variants were created directly from the original answers. Most of the issues found actually stemmed from contradictions or inaccuracies present in the source material.

Each variant was annotated using the complexity formula described in Section 3.3. This gave us raw scores between -34.56 and 31.99, which we converted to a more practical 0-100 scale and then binned into 21 categories labeled 0, 5, 10, and so on up to 100, with each bin containing roughly 8,800 samples. These bins aligned reasonably well with our original five levels, though with some natural overlap between categories. For example, the majority of high school-level variants fell within bins labeled 50-70, while college-level variants typically ranged from 60-80.

The final dataset includes 184,843 answers for 36,969 questions. Each entry has the original question, the reference answer, the variants at different complexity levels, as well as the corresponding evaluation metrics and complexity scores.

3.5 Model Fine-Tuning

After creating our multi-level dataset, we fine-tuned a language model to generate medical text with controlled complexity levels. We experimented with two different methods: natural language instructions and control codes.

For natural language instructions, we used prompts like "Answer the following question with a complexity score of 75 out of 100." For control codes, we added special tokens to the model's vocabulary (e.g., "<COMPLEXITY_75>") and placed them at the beginning of each prompt. These new tokens were initialized by positioning them along a "complexity direction" in the embedding space. We identified simple and complex anchor words in the model's vocabulary, created a vector between them, and placed our tokens along this vector. This gave the tokens semantic meaning before training even began.

We selected Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as our base model and applied LoRA fine-tuning (Hu et al., 2021) with rank 8, alpha 16, and a learning rate of 5e-5, and targeted all projection matrices in the transformer architecture.

During training, we implemented context-aware batching, grouping all answers for the same medical question into a single batch. This helped the model focus on the patterns that actually matter and avoid spurious correlations. For example, if a batch includes both simple and technical answers about asthma, gradient updates adjust the model's weights to preserve important details, such as inflammation and breathing issues, while tailoring the language to match the desired complexity level.

We found that using control codes worked better than using natural language instructions. The training converged faster, and the model generated more consistent responses at each complexity level.

4 Experiments and Results

4.1 Validation of Complexity Scoring Formula

We evaluated our complexity scoring formula using data from the four datasets introduced in Section 3.1. We trained the formula on 80% of the Claude dataset and tested it on the remaining 20%, as well as the complete Cochrane and PLABA datasets. This setup helped us determine how well our formula works for different kinds of text.

For comparison, we used two baselines. The first was the Flesch-Kincaid Grade Level (FKGL), which is the most commonly used method to measure text readability. The second baseline (marked with † in Table 3) represents the top-performing metric for each specific dataset, which naturally varied from one dataset to another. To evaluate performance, we used three statistical measures: Cohen's d gives the standardized difference between the means of the two distributions, Jensen-Shannon Divergence measures the dissimilarity between the distributions, and Area Under the Curve (AUC) measures how well the scoring method distinguishes the two classes, estimating the probability that a randomly chosen complex text gets a higher score than a randomly chosen simple one.

Figure 2 shows the score distributions of simple (green) and complex (red) texts using our formula. The Claude dataset shows the strongest separation, with almost no overlap. Cochrane and PLABA-paragraph are also well separated, although many of the simplified texts still include technical terms, which leads to some overlap. PLABA-sentence has the most overlap, which may be due to the limited context in short texts.

Table 3 compares our formula against the baseline methods. While certain metrics occasionally show slightly better results on specific datasets, their performance fluctuates more from case to case. In contrast, our formula consistently delivers strong results regardless of text length, domain, or simplification strategy. Moreover, perfect numerical separation is not always ideal, as some degree of overlap between distributions may actually reflect genuine ambiguities or edge cases in the data, not necessarily a flaw in the scoring method. In practice, what matters is how well a score captures the

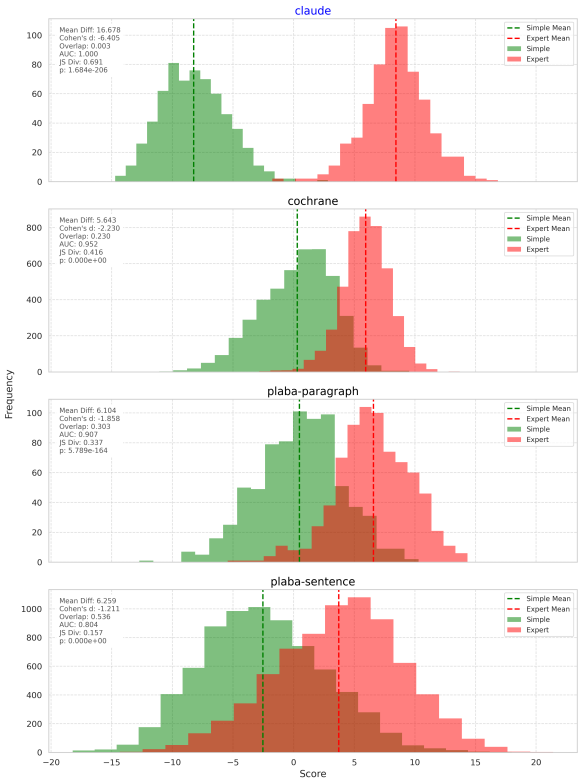

Figure 2: Distribution of complexity scores in the four parallel text datasets.

Dataset	Method	Cohen's *d*	AUC	JS Div.
PLABA-sent	Our formula	**1.21**	**0.80**	**0.16**
	FKGL	0.58	0.67	0.05
	†	0.99	0.76	0.11
PLABA-para	Our formula	1.86	0.91	**0.34**
	FKGL	0.95	0.76	0.12
	†	**1.89**	**0.91**	0.32
Cochrane	Our formula	2.23	0.95	0.42
	FKGL	0.61	0.68	0.06
	†	**2.36**	**0.95**	**0.42**
Claude	Our formula	**6.40**	**1.00**	**0.69**
	FKGL	1.58	0.90	0.31
	†	6.11	1.00	0.67

† Represents the best-performing metric for each dataset.

Table 3: Comparison of readability scoring methods.

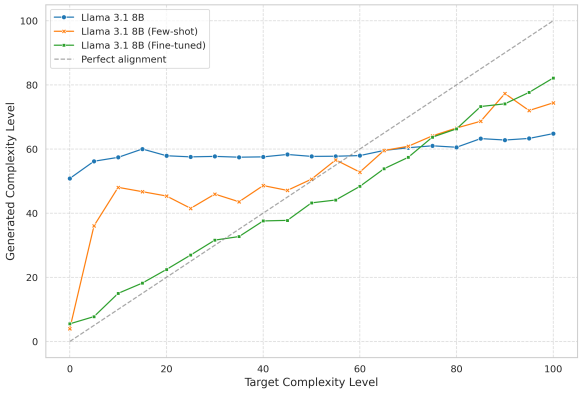

Figure 3: The ability of each model to generate text at the desired complexity level.

perceived reading difficulty experienced by individuals with different levels of health literacy, not just how cleanly it separates two labeled groups in a curated dataset.

4.2 Evaluation of Fine-Tuned Model

We evaluated the ability of our fine-tuned model to generate text at specific complexity levels by comparing it to the original base model and a version using few-shot prompting. Using 100 questions sampled from HealthSearchQA (Singhal et al., 2023), we generated responses at each target complexity level and calculated the difference between the requested complexity and the actual complexity of the generated text.

Figure 3 shows the relationship between the target and the generated complexity levels for each model. The fine-tuned model closely follows the ideal diagonal line, particularly at lower and mid-range levels, with some compression at the highest levels (80-100). The few-shot approach shows a step-like pattern, indicating that it captures general complexity trends but lacks fine-grained control. Meanwhile, the baseline outputs are clustered around a fixed level (~ 60), showing little response to different targets.

5 Conclusions

We introduce a framework for creating medical answers tailored to different health literacy levels. We analyzed 166 linguistic features and defined a scoring formula based on a smaller set of 13, incorporating domain terminology, syntactic complexity, and signals from large language models, to reliably distinguish simple from complex medical text. Using this formula and public resources including LiveQA, MedQuAD, and BioASQ, we created a large dataset of 184,843 medical question-answer pairs rewritten at 21 complexity levels, filling a gap in training materials. We then fine-tuned a language model to generate text at distinct complexity levels, from very simple explanations to highly technical content for medical professionals. This versatility makes it useful in many healthcare settings. It can help create personalized patient education materials, support medical students as they learn more advanced topics, and generate documentation for healthcare providers, such as doctors and nurses.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R. Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. *Studies in Health Technology and Informatics*, 264:25–29.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.

Kanhai S Amin, Linda C Mayes, Pavan Khosla, and Rushabh H Doshi. 2024. Assessing the efficacy of large language models in health literacy: A comprehensive cross-sectional study. *Yale Journal of Biology and Medicine*, 97:17–27.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku anthropic.

Alan R. Aronson and François Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17:229–236.

Kush Attal, Brian Ondov, and Dina Demner-Fushman. 2023. A dataset for plain language adaptation of biomedical abstracts. *Scientific Data*, 10:1–11.

Chandrayee Basu, Rosni Vasu, Michihiro Yasunaga, and Qian Yang. 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, 37:14093–14101.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education.

Miriam Cha, Youngjune Gwon, and H. T. Kung. 2017. Language modeling by clustering with word embeddings for text readability assessment. *International Conference on Information and Knowledge Management, Proceedings*, Part F131841:2003–2006.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Scott Crossley, Aron Heintz, Joon Suh Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2023. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55:491–507.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational Research Bulletin*, 27:37–54.

Jerwin Jan S. Damay, Gerard Jaime D. Lojico, Kimberly Amanda L. Lu, and Dex B. Tarantan. 2006. Simtext: Text simplication of medical literature. *Bachelor's Theses*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association*, 27:194–201.

Ashwin Devaraj, Iain J. Marshall, Byron C. Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 4972–4984.

Fernanda Ferreira. 2003. The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47:164–203.

Lorenzo Jaime Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4859–4873. Association for Computational Linguistics.

Edward Gibson. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68:1–76.

Edward Gibson. 2000. *The dependency locality theory: A distance-based theory of linguistic complexity.*, pages 94–126. The MIT Press.

Gondy, Kauchak David, Gu Yang, Colina Sonia, Yuan Nicole P, Revere Debra Kloehn Nicholas, and Leroy. 2018. Improving consumer understanding of medical text: Development and validation of a new subsimplify algorithm to automatically generate term explanations in english and spanish. *J Med Internet Res*, 20:e10779.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models.

John A Hawkins. 2004. *Efficiency and Complexity in Grammars*. Oxford University Press.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Pedram Hosseini, Christopher Wolfe, Mona Diab, and David Broniatowski. 2022. Gispy: A tool for measuring gist inference score in text. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 38–46. Association for Computational Linguistics.

Yi-Sheng Hsu, Nils Feldhus, and Sherzod Hakimov. 2024. Free-text rationale generation under readability level control.

Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ICLR 2022 - 10th International Conference on Learning Representations*.

Yichen Huang and Ekaterina Kochmar. 2024. Referee: A reference-free model-based metric for text simplification.

Chao Jiang and Wei Xu. 2024. Medreadme: A systematic study for fine-grained sentence readability in medical domain.

Marcel Adam Just and Patricia A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99:122–149.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. *AMIA Annual Symposium Proceedings*, 2010:366.

Alla Keselman, Tony Tse, Jon Crowell, Allen Browne, Long Ngo, and Qing Zeng. 2007. Assessing consumer health vocabulary familiarity: an exploratory study. *Journal of medical Internet research*, 9.

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

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Charles Laurin, Dorret Boomsma, Gitta Lubke, Stat Appl, Genet Mol, and Biol Author. 2016. The use of vector bootstrapping to improve variable selection precision in lasso models. *Statistical applications in genetics and molecular biology*, 15:305.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. Large language models for biomedical text simplification: Promising but not there yet.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries.

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

Junru Lu, Jiazheng Li, Byron C. Wallace, Yulan He, and Gabriele Pergola. 2023. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Findings of EACL 2023*, pages 1049–1061.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022. Readability controllable biomedical document summarization. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680.

Chen Lyu and Gabriele Pergola. 2024. Scigispy: a novel metric for biomedical text simplification via gist inference score. *TSAR 2024 - 3rd Workshop on Text Simplification, Accessibility and Readability, Proceedings of the Workshop*, pages 95–106.

Agnes Malatinszky, Aron Heintz, asiegel, Heather Harris, J S Choi, Maggie, Phil Culliton, and Scott Crossley. 2021. Commonlit readability prize. Kaggle.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60. Association for Computational Linguistics.

Philip M. McCarthy and Scoot Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.

Harry G McLaughlin. 1969. Smog grading - a new readability formula. *Journal of Reading*, pages 639–646.

Juhi M Mohile, Joan B Luzon, Gunjan Agrawal, Neha R Malhotra, and Kathleen M Kan. 2023. Assessment of readability and quality of patient education materials specific to nocturnal enuresis. *Journal of Pediatric Urology*, 19:558.e1–558.e7.

National Library of Medicine. 2024. Umls knowledge sources. Cited 2025 Apr 7.

OpenAI. 2023. Gpt-4 technical report.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences.

Jürgen Pelikan, Christa Straßmayr, Thomas Link, Dominika Miksova, Peter Nowak, Robert Griebler, Christina Dietscher, Stephan den Broucke, Rana Charafeddine, Antoniya Yanakieva, Nygyar Dzhafer, Zdenek Kučera, Alena Šteflová, Henrik Bøggild, Andreas Sørensen, Julien Mancini, Geneviève Chêne, Doris Schaeffer, Alexander Schmidt-Gernig, and Øystein Guttersrud. 2021. International report on the methodology, results, and recommendations of the european health literacy population survey 2019-2021 (hls19) of m-pohl.

Atharva Phatak, David W. Savage, Robert Ohle, Jonathan Smith, and Vijay Mago. 2022. Medical text simplification using reinforcement learning (teslea): Deep learning–based text simplification approach. *JMIR Medical Informatics*, 10.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108. Association for Computational Linguistics.

Md Mushfiqur Rahman, Mohammad Sabik Irbaz, Kai North, Michelle S. Williams, Marcos Zampieri, and Kevin Lybarger. 2024. Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning. *Journal of Biomedical Informatics*, 158:104727.

Kamal raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan Sankarasubbu. 2022. Biosimcse: Biomedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86. Association for Computational Linguistics.

Valerie F. Reyna. 2012. A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and decision making*, 7:332.

Leonardo F.R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 11669–11687.

Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data*, 7:1–9.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Rabia Shahid, Muhammad Shoker, Luan Manh Chu, Ryan Frehlick, Heather Ward, and Punam Pahwa. 2022. Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC Health Services Research*, 22:1–9.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180.

E A Smith and R J Senter. 1967. Automated readability index. Technical report, Aerospace Medical Research Laboratories (U.S.), Wright-Patterson Air Force Base, Ohio. PMID: 5302480.

Luca Soldaini. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction.

T Szmuda, C Özdemir, S Ali, A Singh, M T Syed, and P Słoniewski. 2020. Readability of online patient education material for the novel coronavirus disease (covid-19): a cross-sectional health literacy study. *Public Health*, 185:21–25.

Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. 2024. Readctrl: Personalizing text generation with readability-controlled instruction learning.

Xun Wang and Robin A Cohen. 2022. Health information technology use among adults: United states, july-december 2022 key findings data from the national health interview survey. Technical report.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024. Helpsteer2-preference: Complementing ratings with preferences.

Nikki Keene Woods, Umama Ali, Melissa Medina, Jared Reyes, and Amy K. Chesser. 2023. Health literacy, health outcomes and equity: A trend analysis based on a population survey. *Journal of Primary Care Community Health*, 14.

Biyang Yu, Zhe He, Aiwen Xing, and Mia Liza A. Lustria. 2020. An informatics framework to assess consumer health language complexity differences: Proof-of-concept study. *Journal of medical Internet research*, 22.

Qing T. Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association : JAMIA*, 13:24.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *8th International Conference on Learning Representations, ICLR 2020*.

Jiaping Zheng and Hong Yu. 2018. Assessing the readability of medical documents: A ranking approach. *JMIR Medical Informatics*, 20.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 2023–2038.

A Appendix

A.1 Medical Text Simplification Datasets

PLABA. The Plain Language Adaptation of Biomedical Abstracts dataset (Attal et al., 2023) contains 750 biomedical abstracts paired with their plain language versions, totaling 7,643 sentence pairs. It was created by scraping 75 common medical questions from forums and retrieving relevant paper abstracts from PubMed. Human annotators then adapted these abstracts by replacing technical terminology with common synonyms, splitting complex sentences, and removing content irrelevant to general readers. We used PLABA at both sentence and paragraph levels to evaluate our complexity metrics.

Cochrane Dataset. The Cochrane Simplification dataset (Devaraj et al., 2021) contains 4,459 pairs of technical medical texts and their corresponding simplified versions, sourced from the Cochrane Database of Systematic Reviews. This dataset provides paragraph-level simplifications designed to make medical content more accessible to readers without medical expertise.

A.2 Selected Metrics for Complexity Formula

The following 13 metrics were selected for our final complexity scoring formula, listed here along with their coefficients:

A.2.1 LLM Vocabulary Complexity (3.217)

We use this score to estimate how difficult a piece of text is to understand, based on evaluations from the three 70-billion parameter models we described earlier. Each model rated texts on a scale from 1 to 5, with higher scores indicating more complex language. This feature has the largest positive coefficient in our formula, confirming that vocabulary choice drives most of the perceived difficulty in medical texts.

A.2.2 Dale-Chall Score (1.839)

The Dale-Chall readability formula (Dale and Chall, 1948) estimates how difficult a text is to read based on the average sentence length and the percentage of "difficult" words not found on a predefined list of familiar words. In our implementation, we expanded the original list of 3,000 words by including those from the Spache list. The positive coefficient in the formula shows that texts with longer sentences and more unfamiliar words tend to be significantly more complex.

A.2.3 Type-Token Ratio (0.173)

Type-token ratio (TTR) measures lexical diversity by dividing the number of unique words (types) by the total number of words (tokens) in a text. The positive coefficient confirms that texts with more diverse vocabulary contribute to higher complexity scores, though with less impact than the vocabulary complexity or the Dale-Chall readability formula.

A.2.4 ALBERT Transformer Score (-2.471)

We used the ALBERT-xxlarge model (Lan et al., 2019) from the winning entry in the CommonLit Readability Prize Kaggle competition (Malatinszky et al., 2021). This model processes text through attention layers to capture relationships between words before predicting a readability score. The negative coefficient appears because ALBERT assigns higher scores to texts that are easier to read, which runs in the opposite direction of our scoring system, where higher values indicate lower readability.

A.2.5 Referential Cohesion (0.068)

This feature captures how well a paragraph maintains topical consistency by measuring the semantic similarity between consecutive sentences (Lyu and Pergola, 2024). To compute it, we embed each sentence using BioSimCSE-BioLinkBERT-BASE (raj Kanakarajan et al., 2022) and calculate the cosine similarity between adjacent sentence pairs. A sharp drop in similarity, falling in the bottom 25% of the distribution, marks a potential topic shift, or "breakpoint." We count the number of chunks in each paragraph based on these breakpoints and take the average over the entire text. The small positive coefficient may seem counterintuitive, since texts that are more cohesive are usually easier to read. However, this result suggests that even highly

technical medical texts in our datasets tend to maintain strong internal cohesion despite their complex vocabulary.

A.2.6 Information Content (0.691)

This feature measures how specialized the vocabulary is in a given text, based on how often each word appears in a biomedical corpus (Lyu and Pergola, 2024). The basic idea is that technical terms tend to be rarer and harder to understand. To build our reference corpus, we combined data from biomedical and consumer health sources, including MedQuAD (Abacha and Demner-Fushman, 2019), LiveQA (Abacha et al., 2017), MedicationQA (Abacha et al., 2019), and other medical datasets. We lemmatize each word in the corpus, count how often each lemma appears, and calculate its information content as the negative logarithm of its probability. For any given text, we extract all nouns and verbs, look up their information content values, and calculate the average. The positive coefficient in our model supports the idea that texts with a more technical and less common vocabulary tend to be more complex.

A.2.7 Verb Ratio (-0.330)

Part-of-speech distributions measure the frequency of different grammatical categories relative to the total word count. We calculate separate ratios for nouns, verbs, adjectives, adverbs, conjunctions, and auxiliary verbs. A negative coefficient for verb ratio indicates that texts with fewer verbs relative to other parts of speech are rated as more complex. This is consistent with research showing that academic and scientific writing tends to use more nouns and fewer verbs (Biber et al., 1999).

A.2.8 Function Word Ratio (-0.596)

The content-to-function word ratio calculates the proportion of content words (nouns, verbs, adjectives, adverbs) to function words (auxiliaries, determiners, prepositions, conjunctions) in a text. A negative coefficient means that texts with more content words and fewer function words are seen as more complex. This is because function words help organize sentence structure, so when they are used less frequently, the resulting text can be more syntactically dense and cognitively demanding for readers (Just and Carpenter, 1992).

A.2.9 Masked Probability Score (-0.049)

This metric evaluates how predictable words are in biomedical text using a masked language model Devaraj et al. (2021).. Specifically, we randomly mask 15% of the tokens and run this process 30 times, then measure how accurately Bio+ClinicalBERT can guess the original words. In general, technical or scientific writing tends to have more predictable language patterns, especially due to consistent use of domain-specific terms. The negative weight in the scoring formula helps balance out other vocabulary-based metrics. It prevents penalizing texts that are technically dense but still internally consistent and readable.

A.2.10 MedReadMe Cluster Score (0.295)

This score comes from a clustering-based word embedding model trained on the MedReadMe dataset (Jiang and Xu, 2024). Each word in the text is converted into its BioWordVec embedding, then assigned to one of 300 semantic clusters using K-means clustering. The pattern of these assignments forms a feature vector that represents how the vocabulary is distributed across different semantic categories. A positive coefficient means that texts using vocabulary patterns similar to those found in more complex medical content tend to receive higher complexity scores.

A.2.11 Embedding Depth (-0.161)

Embedding depth measures how deep the hierarchical structure of a sentence goes in its dependency tree. To calculate this, we identify the word with the longest chain of grammatical dependencies leading to the root of the sentence. A sentence with greater embedding depth usually contains more subordinate clauses (introduced by words like "which," "that," "when") and complex phrases embedded within one another. This typically makes text harder to process, as readers must track multiple incomplete grammatical relationships while reading, increasing cognitive effort (Gibson, 1998). However, in our corpus, the expert texts often used more concise, noun-heavy sentences with fewer nested clauses. In contrast, the simpler texts used more explanatory language with embedded clauses to break down complex concepts. This pattern explains the negative coefficient in our formula.

A.2.12 Average Dependency Distance (-0.826)

Dependency distance measures how many words separate a dependent word (object or modifier) from its head word (main verb or noun) in a sentence. Longer distances increase cognitive load, since the reader must keep track of the dependent

word while processing the words in between (Gibson, 2000). We calculate the average dependency distance for each sentence and then find the overall average for the entire text. Although this metric correlates with higher difficulty when used alone, the negative coefficient in our multivariate model suggests an inverse relationship when considered alongside other features.

A.2.13 Coreference Chains (-0.390)

Coreference resolution tracks how entities are referenced throughout a text. When a document refers to the same person, object, or concept using different terms (e.g., pronouns, synonyms, or descriptions), it creates coreference chains that help readers follow who or what is being discussed. For instance, if a text mentions "Dr. Smith" and later refers to her as "she" or "the physician," these references form a continuous link to the same entity. To calculate CoREF, we use FastCoref (Otmazgin et al., 2022) instead of the Stanford CoreNLP implementation previously used in GisPy (Manning et al., 2014; Qi et al., 2020). We decided to make this switch because CoreNLP was causing significant delays in the processing pipeline, especially when working with longer documents. FastCoref, on the other hand, not only performs on par with state-of-the-art models but also runs much faster, completing tasks in seconds that used to take minutes. Following the same methodology as GisPy, we identify all coreference chains in a document, calculate the ratio of chains to sentences for each paragraph, and then compute the final CoREF score as the average of these paragraph-level scores. The negative coefficient indicates that complex medical texts often contain fewer or shorter coreference chains, introducing new entities without established reference patterns, which increases reading difficulty.

A.3 Source Medical QA Datasets

This section provides detailed descriptions of the five source datasets used to create our multi-level medical QA dataset:

LiveQA. The LiveQA dataset (Abacha et al., 2017) includes real-world consumer health questions submitted to the U.S. National Library of Medicine (NLM) during the TREC 2017 LiveQA challenge. The original release had 634 training pairs and 104 test questions, each with multiple reference answers. After cleaning the data, we retained 800 question-answer pairs covering topics such as diseases, treatments, medications, and medical exams.

MedicationQA. The MedicationQA dataset (Abacha et al., 2019) contains 690 consumer questions about medications, each paired with an answer from a trusted medical website, such as MedlinePlus and DailyMed, addressing topics like drug usage, dosage, side effects, and drug interactions.

MediQA-AnS. The MediQA-AnS dataset (Savery et al., 2020), created for the MEDIQA 2021 challenge, includes 156 consumer health questions, each paired with two reference summaries (abstractive and extractive) both written by medical experts based on passages retrieved using the CHiQA system (Demner-Fushman et al., 2020).

MedQuAD. The Medical Question Answering Dataset (Abacha and Demner-Fushman, 2019) consists of 47,457 question-answer pairs sourced from 12 websites managed by the U.S. National Institutes of Health (NIH), including MedlinePlus, cancer.gov, and niddk.nih.gov. Due to copyright restrictions, we had to exclude over 31,000 entries, leaving us with a total of 16,423 samples.

BioASQ. The BioASQ Task 13B dataset, part of the 2025 BioASQ challenge, includes 5,389 biomedical questions. Each question is paired with one or more ideal answers, resulting in a total of 13,692 question-answer pairs. These answers are concise, expert-written summaries that draw from scientific literature, primarily PubMed, and use precise biomedical terminology.

A.4 Content Filtering Rules

To ensure quality in our multilevel dataset, we implemented a two-phase filtering process:

A.4.1 Pattern-Based Filtering

We used regular expressions to identify and remove samples containing:

- Placeholder text (e.g., "[CONTENT_MISMATCH]")
- Empty or extremely short answers (fewer than 15 tokens)
- Formatting issues or broken XML structure
- Repetitive text patterns suggesting generation failures

This phase removed approximately 0.3% of the initially generated samples.

A.4.2 Quality-Based Filtering

We evaluated each variant against its original answer using five metrics:

ROUGE-L (Lin, 2004) measures the longest common subsequence between the generated variant and the original answer, capturing structural similarity and content preservation.

BLEURT (Sellam et al., 2020) evaluates semantic similarity between texts using contextualized embeddings, trained to correlate with human judgments.

BERTScore (Zhang et al., 2019) computes token similarity using contextual embeddings from BERT, providing a more semantically-aware measure of content preservation than n-gram overlap metrics.

UniEval (Zhong et al., 2022) assesses multiple dimensions, including relevance to the original content, factual consistency, and coherence.

SummaC (Laban et al., 2022) evaluates contradiction and factual consistency using natural language inference models to detect potential misrepresentations or logical inconsistencies.

We filtered out variants that failed any of these criteria:

- Low relevance and factual consistency (UniEval-relevance < 0.5 AND UniEval-fact-consistency < 0.5)
- Strong logical inconsistency (SummaC < -0.5)
- Poor semantic similarity and text quality (BERTScore-F1 < 0.8 AND BLEURT < 0.2)

This filtering process identified relatively few problems, removing only 2,926 samples (1.56%) from the initially generated 187,769 variants.

A.5 Prompt for Variant Generation

The prompt used to generate variants at different complexity levels was:

```
You are an expert in creating
    educational content for
    different reading abilities.
    Your task is to generate
    multiple answer variants for
    the given question and
    original answer, each at a
    specified complexity level,
    while preserving all factual
    information.

When generating each variant, you
    must:
1. Preserve ALL factual
    information from the original
    answer and keep it relevant to
    the question.
2. Adjust vocabulary, sentence
    structure, and explanation
    detail to match the complexity
    level.
3. Do not introduce substantively
    new claims that aren't
    reasonably implied by the
    original answer.
4. Ensure the answer is coherent
    and well-structured.
5. If the original answer does
    not directly address the
    question asked, respond with:
    '[CONTENT_MISMATCH]' as the
    answer.

Complexity levels (1 to 5) are
    defined as follows:
- Level 1: For a young child; use
    very simple vocabulary, short
    sentences, and basic concepts
    .
- Level 2: For a middle school
    student; use basic scientific
    terms, clear explanations, and
    moderate detail.
- Level 3: For a high school
    student; use technical
    terminology, longer sentences,
    and detailed explanations.
- Level 4: For a college graduate
    ; use in-depth technical
    details, complex sentence
    structures, and scientific
    language.
- Level 5: For a biomedical
    expert; use advanced
    scientific terminology, assume
    prior knowledge, and provide
    precise details.
```

The prompt also included three examples with

answers at each complexity level and instructions to format responses in structured XML.

A.6 Model Performance Metrics

The quantitative results in Table 4 confirm what we see in Figure 3. The fine-tuned model outperforms both alternatives in every metric, with a mean absolute error (MAE) 23% lower than the few-shot method and nearly 50% lower than the baseline. The strong correlation coefficient (0.84) and high R^2 value (0.66) together validate its ability to consistently generate responses at the intended complexity level.

Model	MAE	RMSE	Correlation	R^2
Baseline	26.07	31.28	0.21	-0.07
Few-shot	17.33	22.03	0.69	0.47
Fine-tuned	**13.30**	**17.63**	**0.84**	**0.66**

Table 4: Comparison of how accurately each model generates text at the desired complexity levels.

B Limitations

While our work has made meaningful progress in simplifying medical texts, it also has some important limitations.

First, we focused only on English. The features we used to measure complexity, especially those tied to medical terms, may not translate well to other languages that have different grammar rules or naming conventions in medicine.

Second, we developed our complexity formula without using human feedback. Instead, we assumed that the best formula is the one that maximizes the gap between simple and complex texts, following a set of heuristics we defined based on our understanding of the data. However, perceived complexity is subjective and can vary depending on a person's background, reading ability, and familiarity with medical concepts. Therefore, testing with real users would be necessary to confirm if the formula aligns with human judgments. Furthermore, because the formula is a simple linear equation, it can be "gamed." For example, shortening sentences or swapping in simpler words could reduce the complexity score without making the text any easier to understand. A possible solution would be to follow the same approach we used in Section 3.2.8, distilling the scores into a language model to smooth errors and reduce the impact of outliers. We chose a simple, interpretable formula here, but future work should explore more flexible, non-linear models.

Another concern is the use of synthetic data for training. Even with filters in place, the dataset might still include outdated or inaccurate medical information carried over from the original sources. We also did not evaluate our model against external benchmarks or state-of-the-art systems like GPT-4 or Claude, and lacked specialized datasets to measure factual accuracy and relevance when controlling the complexity of the answers. More importantly, the model was not tested longitudinally with actual users.

Finally, we did not explore alternative methods for measuring text complexity. For example, training models to predict which of two texts is more complex (as in learning-to-rank frameworks) rather than assigning absolute scores could be an alternative approach to evaluate text readability.