README: Bridging Medical Jargon and Lay Understanding for Patient Education through Data-Centric NLP

Zonghai Yao [1], **Nandyala Siddharth Kantu** [1], **Guanghao Wei** [1], **Hieu Tran** [1],
Zhangqi Duan [1], **Sunjae Kwon** [1], **Zhichao Yang**[1], **README annotation team**[2], **Hong Yu**[1,2]

University of Massachusetts, Amherst[1], University of Massachusetts, Lowell[2]
zonghaiyao@umass.edu

Abstract

The advancement in healthcare has shifted focus toward patient-centric approaches, particularly in self-care and patient education, facilitated by access to Electronic Health Records (EHR). However, medical jargon in EHRs poses significant challenges in patient comprehension. To address this, we introduce a new task of automatically generating lay definitions, aiming to simplify complex medical terms into patient-friendly lay language. We first created the README dataset, an extensive collection of over 50,000 unique (medical term, lay definition) pairs and 300,000 mentions, each offering context-aware lay definitions manually annotated by domain experts. We have also engineered a data-centric Human-AI pipeline that synergizes data filtering, augmentation, and selection to improve data quality. We then used README as the training data for models and leveraged a Retrieval-Augmented Generation method to reduce hallucinations and improve the quality of model outputs. Our extensive automatic and human evaluations demonstrate that open-source mobile-friendly models, when fine-tuned with high-quality data, are capable of matching or even surpassing the performance of state-of-the-art closed-source large language models like ChatGPT. This research represents a significant stride in closing the knowledge gap in patient education and advancing patient-centric healthcare solutions [1].

1 Introduction

Throughout the extensive history of natural language processing (NLP), enhancing individuals' ability to read and comprehend complex texts has been a significant research goal that remains only partially achieved (Zeng et al., 2020). Among the various tasks in this field, improving the comprehension of medical electronic health records
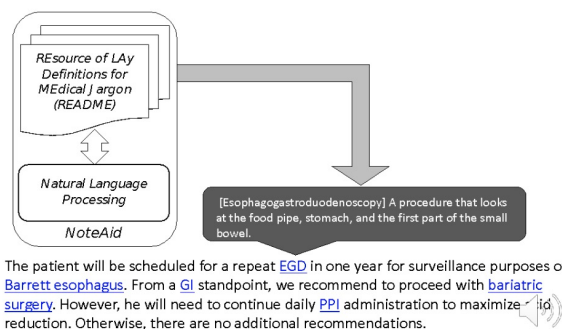

Figure 1: A visualization of the NoteAid pipeline, where NLP tools first identify jargon that may be challenging for patients to understand. The lay definitions corresponding to these jargon terms are then retrieved from relevant dictionaries and presented to the patients, enhancing their comprehension and engagement with their health information.

(EHRs) stands out due to the complexity and specificity of medical terminology (Nutbeam, 2023). Efforts to enhance the comprehension of EHRs can not only advance NLP's goal of aiding the understanding of complex texts but also hold substantial social significance by increasing efficiency and reducing errors for both patients and healthcare professionals (Nanna et al., 2009; Boling, 2009; Spooner et al., 2019; Baldry et al., 1986; Schillinger et al., 2009).

However, despite advancements in EHRs management (Walker et al., 2019), one significant barrier persists in the form of medical jargon [2] in EHRs, impeding patient understanding and self-care (Kujala et al., 2022; Choudhry et al., 2016; Khasawneh et al., 2022). As shown in Figure 1, tools like NoteAid (Chen et al., 2018; Kwon et al., 2022), which employs NLP to demystify complex

---

[2]A "medical jargon term" is the specialized language healthcare professionals use that can be complex for non-medical individuals. A "lay definition" translates this jargon into accessible language for the general public, aiming to bridge the understanding gap and enhance the communication of health-related information.

medical terms, have been instrumental in bridging the communication gap between healthcare professionals and patients (Lalor et al., 2018, 2021, 2023). However, lay language dictionary resources like the Consumer Health Vocabulary (CHV) (Zeng and Tse, 2006; He et al., 2017; Ibrahim et al., 2020) are limited in scale, posing a challenge to NoteAid. For instance, only a fraction (about 4%) of the medical terms in NoteAid have been annotated with lay definitions, highlighting the need for a more scalable solution to address this knowledge gap effectively.

Addressing this issue requires shifting our focus to online health education materials such as Unified Medical Language System (UMLS) (Bodenreider, 2004), MedlinePlus (Patrias and Wendling, 2007), Wikipedia, and Google. However, as indicated in Table 4, these resources often present too difficult information for the average patient to understand. For example, these resources' average readability measured by the Flesch Kincaid Grade Level (Solnyshkina et al., 2017) was post-secondary or higher education, while the average readability of a US adult was 7-8th grade level (Doak et al., 1996, 1998; Eltorai et al., 2014). To bridge this gap, we have engaged medical experts to meticulously curate lay definitions for jargon terms found in NoteAid-MedJEx (Kwon et al., 2022), targeting a comprehension level suitable for individuals with a 7th to 8th-grade education. Each jargon term has been redefined across various contexts, ensuring their applicability in diverse clinical scenarios. This effort led to our creation of the REsource of lAy Definitions for MEdical jargon (README) dataset, an expansive resource containing over 51,623 (medical jargon term, lay definition) pairs. The README dataset comprises an impressive 308,242 data points, each consisting of a clinical note context, a medical jargon term, and its corresponding lay definition. Thus, the dataset significantly enhances the accessibility and comprehensibility of medical information for patients.

Yet, the critical aspect of generating lay definitions remains largely unexplored. As patients gain more access to their EHRs, the demand for lay definition resources is escalating. Despite our efforts to expand them, they are inevitably destined to surpass the capacity of current expert-annotated resources. Moreover, the dynamic nature of "jargon" based on individual and context makes pre-annotated expert resources less adaptable to real-life scenarios. The model-driven automatic genera-

tion of lay definitions from medical jargon emerges as a viable solution. Recent research highlighted ChatGPT's potential in its integration with the field of medicine (Brown et al., 2020; OpenAI, 2023; Yang et al., 2023), including generating human-readable definitions for biomedical terms (Remy and Demeester, 2023). Nonetheless, our evaluation of open-source models (refer to Figure 3) shows a significant performance degradation compared to ChatGPT. Using open-source large language models like Llama2 (Touvron et al., 2023) and small language models such as GPT-2 (Radford et al., 2019) is crucial because proprietary LLMs accessed via third-party APIs may not always be feasible, especially in fields like healthcare with strict privacy requirements and economic constraints. Open-source models offer the necessary privacy, while smaller models provide economic and infrastructural benefits, addressing distinct concerns about effectively deploying NLP tools in healthcare scenarios.

To bridge this gap, we aim to train an in-house system using open-source models for automatic lay definition generation to provide reliable lay definitions for jargon in patient education tools like NoteAid. Inspired by research on Retrieval-Augmented Generation (RAG) in general and medical domains (Lewis et al., 2020; Asai et al., 2023; Xiong et al., 2024; Wang et al., 2024; Guo et al., 2024), we designed to use external resources to overcome the limitations of these open-source models in medical knowledge (Sung et al., 2021; Yao et al., 2022, 2023; Chen et al., 2023). We are positioning automatic lay definition generation as a form of text simplification, where language models are prompted to generate context-aware, jargon-specific, and layperson-friendly definitions based on general definitions retrieved from external knowledge resources. Specifically, in this work, we use the UMLS to retrieve general definitions of jargon terms and construct a dataset upon the README that includes context, jargon terms, general definitions, and lay definitions.

To improve the initial README dataset's data quality for model training, we developed a data-focused process called Examiner-Augmenter-Examiner (EAE), as illustrated in Figure 2. Drawing inspiration from the human-in-the-loop concept (Monarch, 2021), we employed human experts to guide AI in both the examiner and augmenter stages. The examiner filters high-quality training data, which may come from expert an-

notations before the augmenter stage or from AI-generated content after. The augmenter generates potentially high-quality synthesized data to increase data points. In the end of EAE, Human annotators review filtered data to ensure its quality. After obtaining high-quality expert-annotated and AI-synthesized datasets, we employed the AI-synthesized dataset to augment the expert-annotated dataset, aiming to explore the effectiveness of AI-synthesized data in training. We implemented a range of heuristic data selection strategies to integrate AI synthetic data, allowing us to incorporate suitable data points into our training process.

In summary, our contributions are as follows:
- Introduced a new task of automatically generating lay definitions for medical jargon. We created a substantial expert-annotated dataset of 308,242 data points that can be used directly as a detailed lay-language dictionary for patient education tools such as NoteAid and as training data for this new task.
- Developed a robust, data-centric pipeline that effectively integrates data filtering, augmentation, and the selection of synthetic data. This approach enhances the quality of README datasets, merging the strengths of AI with human expertise to achieve optimal results.
- Our extensive automatic and human evaluations reveal that when trained with high-quality data, open-source, mobile-friendly small models can achieve or even exceed the performance of cutting-edge closed-source large language models, such as ChatGPT.

2 Problem Statement

Consider a dataset $D = \{X, Y, Z_+\}$ comprising t EHRs, where $X = \{x^1, x^2, \ldots, x^t\}$ represents the contexts of these EHRs, $Y = \{y^1, y^2, \ldots, y^t\}$ denotes the corresponding jargon terms, and $Z_+ = \{z_+^1, z_+^2, \ldots, z_+^t\}$ are the ground truth expert lay definitions. Each EHR context x^i is a sequence of n tokens, expressed as $x^i = \{x_1^i, x_2^i, \ldots, x_n^i\}$, and each lay definition z_+^i consists of m tokens, given by $z_+^i = \{z_{+,1}^i, z_{+,2}^i, \ldots, z_{+,m}^i\}$. The README lay definition generation task T aims to train a reference model M_{ref} such that $M_{ref}(z_+^i \mid x^i, y^i)$ is optimized. The standard approach for fine-tuning M_{ref} on T involves using the cross-entropy loss $\ell_{ce}(z_+^i, M_{ref}(x^i, y^i))$ over the dataset D. To enhance the training of M_{ref}, we introduce an additional set of general definitions $Z_- =$ $\{z_-^1, z_-^2, \ldots, z_-^t\}$, where each z_-^i corresponds to the general definition of the jargon term y^i, generated using openly available data sources (UMLS) or GPT-3.5-turbo. Our proposed EAE pipeline is designed to acquire high-quality general definition data Z_-, culminating in the augmented dataset $D_{simp} = \{X, Y, Z_+, Z_-\}$. The README lay definition generation task T is then formalized as a text simplification task, where M_{ref} is trained to produce Z_+ based on X, Y, and Z_-. This process utilizes a selected subset $D_{SEL} \subseteq D_{simp}$, chosen according to one of the selection criteria: RANDOM, SYNTAX, SEMANTIC, or MODEL.

3 Method

3.1 README Data Collection

The dataset source is a collection of publicly available deidentified EHR notes from hospitals affiliated with an anonymized institution. Herein, 18,178 sentences were randomly sampled, and domain experts then annotated the sentences for medical jargon and corresponding lay definition.

3.2 Lay Definition Annotation

Domain-experts read each sentence and identified as medical jargon terms that would be considered difficult to comprehend for anyone no greater than a 7th-grade education [3]. Overall, 51,623 unique (medical jargon term, lay definition) pairs with 308,242 mentions [4] in the EHR repository have been annotated by complying with the annotation guidelines presented in Appendix A.

3.3 General Definition Retrieval

We then employed the Scispacy library [5] to retrieve the corresponding UMLS definitions (Lindberg et al., 1993) of these annotated medical jargon terms as the general definitions. We follow the retrieval and preprocessing steps in Appendix C to filter the valid general definitions for the README dataset [6]. This preliminary cleaning results in 308,242 data points in the README-exp (e.g., expert annotated) dataset, each consisting of a clinical

---

[3] The rule of thumb is that if a term has a lay definition comprehensible to a 4-7th grader as judged by FKGL (Solnyshkina et al., 2017), this term is included as a jargon term.

[4] So there are 308,242 data points with (EHR context, medical jargon term, lay definition) format.

[5] We used the Scispacy *en_core_sci_lg* model to obtain the data. https://github.com/allenai/scispacy

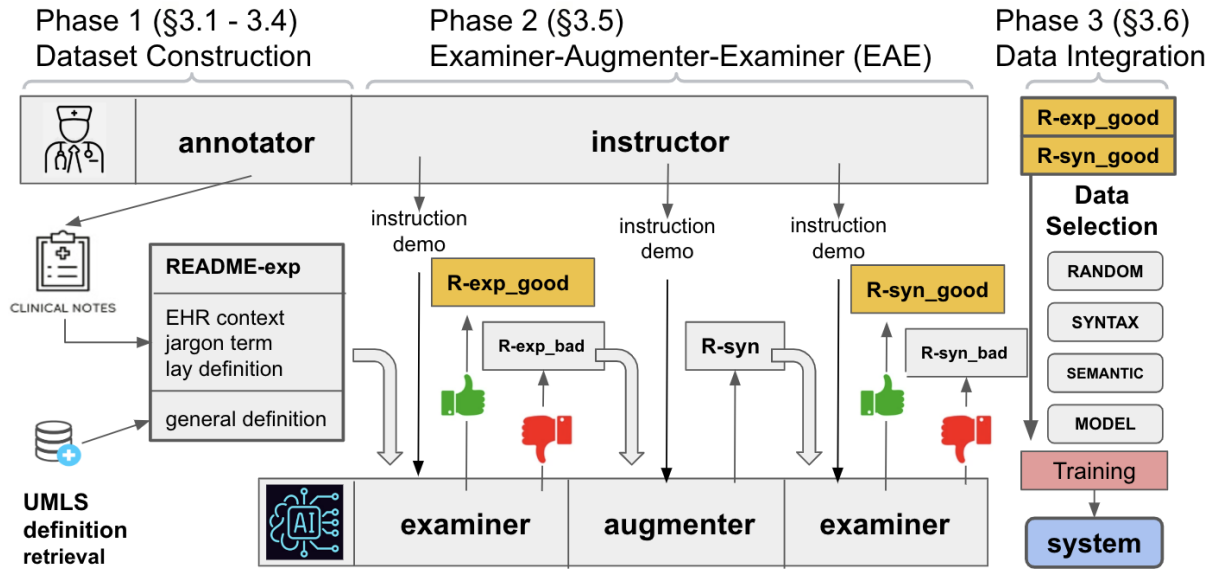[6] We discuss the concept ambiguity issue in Appendix D

Figure 2: Our Data-Centric NLP pipeline, comprising the Examiner-Augmenter-Examiner (EAE) framework and different data selection methods. EAE shows how humans (physicians) and AI (LLM, e.g. ChatGPT) cooperate to make a high-quality README dataset. We collect general definitions for every jargon term from external knowledge resources such as UMLS. "R" is "README". "exp" is "expert annotation version", "syn" is "AI synthetic version". "instruction" and "demo" (examples for in-context learning) are combined into the prompt for LLM. In the pipeline, the human duties at different stages are annotator (labeling the initial dataset) and instructor (providing suitable prompts to guide AI at every stage). The AI duties at different stages are examiner (filter high-quality data) and augmenter (improve the quality of low-quality data). Appendix Table 5 describes the number of different versions of the README dataset in each step. After we get two high-quality datasets, R-exp_good and R-syn_good, from the EAE pipeline, we then deploy 4 different data selection strategies to combine high-quality expert-annotated data R-exp_good and high-quality AI-synthetic data R-syn_bad for in-house system training.

note context, a medical jargon term, a corresponding lay definition, and a corresponding general definition, as shown in Figure 2.

3.4 Examiner-Augmenter-Examiner (EAE)

Examiner (expert-annotated data) Initially, basic data cleaning, as outlined in Appendix C, was applied. To enhance this, we employed GPT-3.5-turbo, using a few-shot learning approach with seven examples - four demonstrating acceptable data points and three showing unacceptable ones. These prompts served as the 'Human' element in our Human-AI-in-the-loop model, as depicted in Figure 2 and detailed in Algorithm 1. The prompts are detailed in Table 9. We choose GPT-3.5-turbo here because our evaluation (Section 4.3 and Appendix F) shows that the definitions it generates for medical terms can reach a human-satisfying level. Post-cleaning [7], approximately 39% of UMLS general definitions were deemed suitable by the Examiner (e.g., GPT-3.5-turbo). The suitable 113,659 data points were archived in R-exp_good, while the

unsuitable 177,140 ones were stored in R-exp_bad.

Augmenter Given the low yield of usable UMLS definitions, we employed GPT-3.5-turbo to augment our dataset. The augmentation process, a critical part of our Data-centric pipeline, involved the system prompt: "*Generate a general definition of the term.*" This step, accompanied by two examples (Table 9), aimed to create correct general definitions but may not be suitable for laypeople (similar to the UMLS definitions). The outcome of this process was R-syn (e.g., AI synthetic), containing 171,831 newly generated definitions.

Examiner (AI-synthetic data) The ChatGPT-generated definitions underwent a second cleaning round using the same methodology as in Section 3.4 Examiner (expert-annotated data). Here, approximately 56% of the ChatGPT definitions were found suitable for model training, with the remaining being either contextually inappropriate or incompatible with the expert-provided lay definitions. The final tally was 96,668 'good' and 75,175 'bad' general definitions, stored in R-syn_good

---

[7]More GPT Running Details are in Appendix B.4

and R-syn_bad, respectively. We discuss the EAE pipeline efficacy in Appendix E

Qaulty Checking After the end of the EAE, we sampled 500 medical jargon terms each from the R-exp_good and R-syn_good datasets. We conducted human verification [8] on corresponding data points for these 1000 jargon terms. We obtained a high human agreement for both the R-exp_good and R-syn_good datasets' quality. After correcting individual invalid data, we used this portion of the data as evaluation and test data (in a 1:1 ratio). We used the remaining R-exp_good and R-syn_good datasets as training data, ensuring that these 1,000 medical jargon terms would not appear in the training data. Table 5 shows the overall dataset statistics of all the README versions across the pipeline.

3.5 Integration of Synthetic and Expert Data

We adopted four distinct sampling strategies to integrate the AI-synthetic training data (e.g., R-syn_good) into the expert-annotated training data (e.g., R-exp_good):

- **RANDOM:** This approach randomly selected N entries from the R-syn_good dataset. This is the baseline for our subsequent three heuristic methods.
- **SYNTAX:** For the syntax-based sampling approach, the ROUGE_L metrics F1 score in Section 4.1 was used as a key evaluative tool. ROUGE_L focuses on the longest common subsequence, which measures the longest string of words that occurs in both the predicted and reference texts. By using this metric, we could rank the synthetic definitions according to their syntactic closeness to the human-written definitions, which helped us select samples that would potentially be more understandable and natural-sounding.
- **SEMANTIC:** For semantic-based sampling, we utilized SentenceTransformers [9]. Renowned for its text semantic analysis efficiency, this model enabled us to measure the semantic similarity between lay definitions in R-exp_good and R-syn_good datasets. We ranked synthetic data based on these scores, considering higher scores as indicative of greater semantic closeness to expert annotations.

- **MODEL:** In model-based sampling, we used models initially trained on the R-exp_good dataset to generate definitions for the R-syn_good dataset. We employed the ROUGE_L F1 score to evaluate the alignment between model-generated and actual R-syn_good lay definitions. This technique aids in mitigating training challenges associated with data heterogeneity. It enriches the dataset with examples that enhance the model's convergence towards the desired distribution (e.g., expert-annotated lay definitions).

4 Experiments

4.1 Automatic Evaluation Metrics

We evaluate the efficacy of our model in producing lay definitions by contrasting them with the ground-truth reference lay definitions, utilizing the ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) metrics. However, they are based on the exact word overlap and therefore provide insight into the informativeness of the generated lay definitions but do not necessarily reflect their factual accuracy (Maynez et al., 2020). We also employ Scispacy to extract medical concepts from the model-generated and the reference lay definitions. We then compute the F1 Scores for these concept lists, referred to as UMLS-F1, to specifically measure the factuality of the generated content [10].

4.2 Experimental Setting

We use the following base models: GPT-2, Distil-GPT2, BioGPT, and Llama2 in our experiments [11]. We use the following symbols:

1. jargon2lay(J2L): Directly generates a lay definition for a given jargon term.

2. jargon+context2lay(J+C2L): Generates a lay definition for a given jargon term based on the context information from clinical documents.

3. jargon+gen2lay(J+G2L): Generates a lay definition for a given jargon term based on the general definition provided by UMLS.

4. jargon+context+gen2lay(J+C+G2L): Generates a lay definition for a given jargon term based on both the context information from clinical documents and the general definition provided by UMLS.

We use one jargon, 'EDG', as an example and show the input prompt of different settings J2L,

---

J+C2L, J+G2L, J+C+G2L [12] on Table 8.

Our experiments were divided into five distinct parts. In **Set-1**, we aimed to evaluate the performance gap on the J2L task between open-source base models and GPT3.5/4 models. We do one-shot prompting for Llama2 [13], GPT-3.5-turbo, and GPT-4 with prompts in Table 6. **Set-2** explored the varying data quality across different versions within our EAE pipeline, where we fine-tuned the base models on J+G2L task. **Set-3** focused on evaluating the effects of various data selection strategies on data augmentation outcomes. To do this, we ranked AI-synthetic data (R-syn_good) using different methods in Section 3.5, selecting either the top N entries with the highest scores (e.g., 'SEMANTIC' in Table 2) or the bottom N entries with the lowest scores (e.g., 'SEMANTIC_r'). These selections were then evenly combined with expert-annotated data (R-exp_good) at a one-to-one ratio. Following this, we fine-tuned the base models using these diverse, mixed datasets to determine the impact of each selection method on model performance. In **Set-4**, we investigated the effects of incorporating different types of information (EHR context and UMLS-retrieved general definition) into the model inputs. In **Set-5**, we fine-tuned models of varying sizes (ranging from DistilGPT2-88M to Llama2-7B) and compared the outcomes with GPT-3.5-turbo using best settings learned from previous Set-2, Set-3, and Set-4. More experiment designs and results can be found in Appendix I
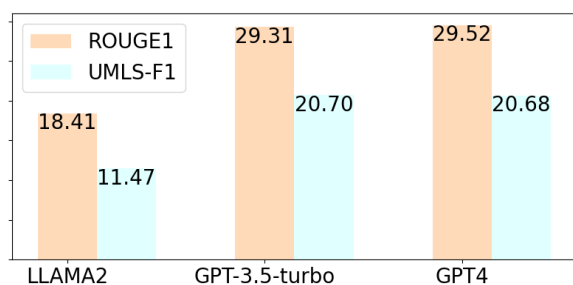
4.3 Results

Figure 3: One-shot performances on jargon2lay.

---

[12]We use "context" and "C" interchangeably to refer to the EHR context in which the jargon occurs. Similarly, we use "general definition" and "G" interchangeably to refer to the general definition retrieved from UMLS.

[13]Because smaller open source language models Distil-GPT2/GPT2/BioGPT do not have the ability to complete instruction following under zero-shot or few-shot settings, we only compare the results of Llama2 as a representative of open source language models with GPT3.5/4 to see the gap.

J+G2L	ROUGE1	ROUGE2	ROUGEL	METEOR	UMLS-F1	Rank
R-exp	23.94	8.95	22.79	17.75	12.83	4
R-exp_good	26.99	10.76	25.57	21.19	16.88	2
R-exp_bad	22.25	7.46	21.09	17.18	11.72	5
R-syn_good	25.71	9.75	24.41	20.23	16.01	3
R-exp+syn_good	29.82	13.14	28.47	24.42	20.09	1

Table 1: Various README versions data performance.

We start with our experiments with some base models' performance on the J2L dataset with **Set-1**. In Appendix F, we showed a high level of human evaluators' agreement for the definitions generated by GPT-3.5-turbo, which underlines its efficacy in crafting human-readable explanations for biomedical concepts. Consequently, GPT-3.5-turbo serves as a strong baseline for quality in README lay definition generation, indicative of a standard that meets human satisfaction. Despite this, our analysis of Llama2, illustrated in Figure 3, reveals a significant performance gap. This discrepancy underscores the critical necessity for enhancing the capabilities of Llama2 and other open-source models to achieve high-quality output in lay definition generation tasks, thereby improving patient education together with systems like NoteAid. Additionally, we observed that GPT-3.5-turbo and GPT4 [14] exhibited comparable proficiency in this task.

4.3.1 Effectiveness of EAE pipeline

In **Set-2**, we focus on the efficacy of different data versions when finetuning with the GPT-2 model. The results, as reflected in Table 1, indicate that high-quality expert data (R-exp_good) demonstrates clear superiority over unexamined expert data (R-exp), emphasizing the crucial role of ECE-examiner in enhancing data quality. Furthermore, high-quality synthetic data (R-syn_good) outperforms the unexamined expert data (R-exp), underscoring the significant value of ECE-augmenter. Notably, the combination of R-exp_good and R-syn_good shows improved performance over R-exp_good alone, suggesting that including synthetic data is beneficial. This composite approach of R-exp_good+R-syn_good leading the rank underscores the efficacy of our EAE pipeline.

4.3.2 Expert and Synthetic Data Integration

In **Set-3**, we explored the effects of various data selection strategies on data augmentation outcomes. We found the results of SEMANTIC are closer to

---

[14]All GPT experiments in our paper were conducted using Microsoft Azure, which can be used HIPAA-compliantly, ensuring the ethical handling of sensitive data.

ROUGE1	J2L	J+C2L	J+G2L	J+C+G2L
RANDOM	19.03	19.65	26.21	26.97
SYNTAX(R)	19.82(-2.28)	20.65(-2.78)	27.56(-8.36)	28.33(-8.84)
SEMANTIC(R)	19.74(-1.11)	19.94(-1.39)	26.42(-0.92)	27.19(-1.96)
MODEL(R)	20.18(-0.8)	20.46(-1.88)	27.9(-2.37)	28.65(-4.22)
UMLS-f	J2L	J+C2L	J+G2L	J+C+G2L
RANDOM	8.05	9.09	15.78	16.72
SYNTAX(R)	8.53(-0.11)	9.99(-1.00)	17.76(-5.11)	17.42(-5.54)
SEMANTIC(R)	8.98(-0.38)	9.07(-0.69)	15.93(-0.02)	16.74(-0.93)
MODEL(R)	9.18(-0.63)	9.4(-0.78)	18.34(-2.5)	18.45(-2.55)

Table 2: Different data selection methods performance. The values in parentheses represent the difference between the corresponding X_r (bottom N) and X (top N), i.e., X_r - X. The smaller this value, the more it indicates that the X method can select higher quality synthetic data for data augmentation; conversely, the closer this value is to 0, the more it suggests that the X method cannot identify the higher-quality synthetic data for training.

RANDOM than SYNTAX and MODEL, and SEMANTIC_r - SEMANTIC is close to 0. However, significant differences can be observed in SYNTAX and SYNTAX_r, MODEL and MODEL_r. Table 2 highlights two main findings: firstly, data selection is crucial, as all methods have better performance with higher-ranked data (e.g., SYNTAX, SEMANTIC, MODEL) over lower (e.g., SYNTAX_r, SEMANTIC_r, MODEL_r). All three methods, SYNTAX, SEMANTIC, and MODEL, have better results than the RANDOM baseline. Secondly, SYNTAX and MODEL are more effective in selecting higher-quality synthetic data for data augmentation than SEMANTIC and RANDOM. We selected SYNTAX as the default data augmentation method for subsequent experiments.

4.3.3 Retrieval-augmented Generation

	ROUGE1	ROUGE2	ROUGEL	METEOR	UMLS-F1
Without data augmentation (R-exp_good)					
J2L	19.47	6.06	18.53	14.74	8.76
J+C2L	19.40	6.40	18.38	15.12	9.24
J+G2L	26.99	10.76	25.57	21.19	16.88
J+C+G2L	27.58	11.31	26.35	21.72	17.12
Data augmentation (R-exp+syn_good with SYNTAX)					
J2L	21.98	7.42	20.88	16.98	10.95
J+C2L	22.13	7.90	21.04	17.54	10.71
J+G2L	29.82	13.14	28.47	24.42	20.09
J+C+G2L	29.89	13.48	28.49	24.65	20.27

Table 3: Efficacy of incorporating EHR context and general definition in input data. The retrieved general definitions significantly aid the overall performance of the model (ROUGE and METEOR) and also reduce hallucinations (UMLS factuality score).

Set-4 results underscore the significant improvement in model performance when input data is enriched with UMLS-retrieved general definitions.

As illustrated in Table 3, regardless of whether we utilize only expert-annotated data or data augmentation with AI-synthetic data, including general definitions consistently enhances effectiveness. This finding confirms the value of RAG with the general definition in the lay definition generation task. Meanwhile, adding EHR context to the input data yields a moderate impact on model performance.
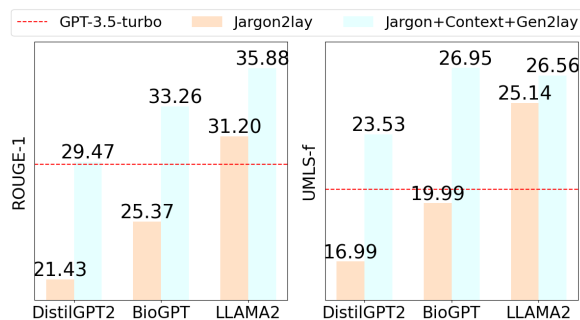
4.3.4 Model Performances Against ChatGPT

Figure 4: Comparative performance analysis of DistilGPT2, BioGPT, and LLAMA2 against GPT-3.5-turbo.

In **Set-5**, we observed that LLAMA2-7B's ROUGE-1 and UMLS-F1 metrics surpassed GPT-3.5-turbo in the J2L task post-training. For the J+G2L setting, DistilGPT2-88M demonstrated equivalent results to GPT-3.5-turbo, while BioGPT's performance exceeded it, and LLAMA2-7B significantly outperformed the GPT-3.5-turbo. These findings, as depicted in Figure 4, emphasize the effectiveness of open-source, mobile-adapted small models when appropriately fine-tuned with high-quality datasets, offering a promising avenue for deploying lightweight yet powerful NLP tools in mobile healthcare applications to help patient education.

5 Human Evaluation

5.1 Human Evaluation settings

Our human evaluation was conducted by 5 human evaluators [15]. We randomly selected 50 pairs of (jargon, generated lay definitions) from the test dataset for this human evaluation. The task for evaluators was to reference the expert definitions and choose a binary preference among the following four groups of definitions: 1) DistilGPT2-J2L vs.

---

[15]Since the generated lay definition is provided to lay people, five people without medical background were found to conduct the human evaluation here.

GPT-3.5-turbo, 2) DistilGPT2-J+C+G2L vs. GPT-3.5-turbo, 3) LLAMA2-J2L vs. GPT-3.5-turbo, 4) LLAMA2-J+C+G2L vs. GPT-3.5-turbo. After we get judgments from multiple people per instance, we do not aggregate their labels before calculating the win rate but count them individually.
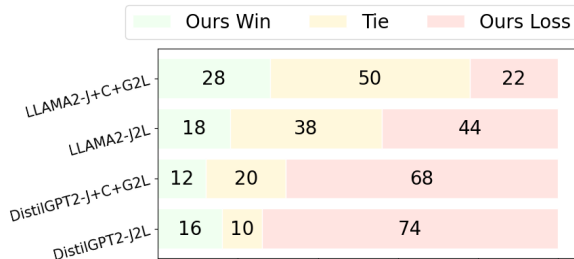
5.2 Human Evaluation Results

Figure 5: Human evaluation results (win rate %).

As shown in Figure 5, although the result of adding EHR context and general definition (DistilGPT2-J+C+G2L) is better than DistilGPT2-J2L, the win rate of the two DistilGPT2 models' output is still significantly behind the result of GPT-3.5-turbo. For LLAMA-2, generating lay definitions directly from jargon is still not as good as GPT-3.5-turbo, but adding context and general definitions is of great help. Human evaluators prefer LLAMA2-J+C+G2L more than GPT-3.5-turbo. There are some inconsistencies between the results of human evaluation and automatic evaluation. We further interviewed our medical experts about the reasons for our system win or loss cases and concluded the following conclusions to help future improvements: 1. While all our in-house systems perform satisfactorily, GPT-3.5-turbo stands out for its flexibility and user-friendliness. It excels at elaborating complex medical terms, offering detailed explanations and practical examples to improve comprehension. 2. Recent advancements (Cai et al., 2023; Zhang et al., 2023b) reveal ChatGPT's role in enhancing patient education through interactive formats like NoteAid-interactive (Zhang et al., 2023a). It enables patients to actively ask questions and seek clarifications while the AI tailors responses to aid their understanding. This interactive approach, absent in traditional dictionary-style definitions like our README dataset, calls for next-step model distillation work or further refinement in aligning the in-house system's outputs with patient preferences. 3. Additionally, developing automatic metrics aligning closely with human evaluation is another critical next step.

6 Related Work

The evolution of patient-centric healthcare necessitates simplified patient access to medical information. Tools like NoteAid and MedJEx have initiated efforts to make EHR content more comprehensible (Chen et al., 2018; Kwon et al., 2022). However, sentence-level text simplification efforts have been expanded to larger datasets that capture broader biomedical contexts (Jonnalagadda et al., 2010; Guo et al., 2021; Devaraj et al., 2021). Recent research has similarly focused on the development of datasets and methods for text simplification, primarily at the sentence level, with the expansion into datasets capturing broader contexts of biomedical abstracts (Cao et al., 2020; Lu et al., 2023; Goldsack et al., 2022; Luo et al., 2022b). Notably, efforts such as the CELLS, PLABA, and AGCT datasets have contributed significantly to this domain, providing extensive resources for training models capable of translating scientific discourse into lay language (Guo et al., 2024; Attal et al., 2023; Remy and Demeester, 2023). Our work diverges from these existing efforts by introducing the README dataset, an expansive collection explicitly designed for context-aware lay definitions, addressing the nuanced task of generating patient-friendly definitions directly from medical terms, filling a critical gap in patient education resources.

In line with advancing the quality of generated texts, we have embraced Retrieval-Augmented Generation (RAG) to mitigate common issues in natural language generation, such as "hallucinations" (Karpukhin et al., 2020; Shuster et al., 2021). Two main categories of information retrieval methods have been used to augment the generation of biomedical natural language generation, definition-based and embedding-based retrieval techniques (Guo et al., 2024; Alambo et al., 2022; Moradi and Ghadiri, 2018; Xiong et al., 2024). Our RAG belongs to the definition-based retrieval technique.

Finally, our contributions distinctly highlight the integration of a robust, data-centric Human-AI pipeline that improves data quality and the efficiency of models trained on the README dataset. This innovative pipeline leverages the Data-centric AI framework, navigating through phases of collection, labeling, preparation, reduction, and augmentation to build a dataset that is both expansive and representative (Zha et al., 2023; Ng et al., 2021;

Whang et al., 2023). The process begins with meticulous data collection and expert labeling, ensuring a foundation of high-quality, domain-specific data (Section 3.1 - 3.3). During the preparation phase, raw data undergoes rigorous cleaning and transformation, readying it for effective model training (Krishnan and Wu, 2019). The dataset is then enriched through strategic data augmentation techniques, incorporating verified quality AI-synthetic data to expand its scope and utility (Section 3.4). Furthermore, data reduction strategies are employed to select the most suitable instances for integration, enhancing the dataset's overall effectiveness (Section 3.5). Through these meticulous stages, the README dataset not only supports but significantly enhances the capabilities of smaller, open-source models, allowing them to match or even exceed the performance of larger proprietary models like ChatGPT in specific healthcare applications.

7 Conclusions

Our study underscores the potential of NLP to democratize medical knowledge, enabling patient-centric care by simplifying complex medical terminology. Developing the README dataset and implementing a data-centric pipeline has improved dataset quality and expanded AI training possibilities. Our experiments show that small open-source models can match advanced close-source LLMs like ChatGPT with well-curated data. README will be open to the community as an important lay dictionary for patient education. We hope our work can help the innovative patient education community advance toward a future where all patients can easily understand their health information.

8 Limitations and Ethical Considerations

This study provides valuable insights, but experimental results evaluated in humans demonstrate limitations of the current work and some future directions. First, better automatic evaluation metrics need to be explored to be closer to human evaluation results. Secondly, in this paper, we have only explored some heuristic data selection methods, and we need to explore more sophisticated methods in the future. In addition, the next step of the in-house system is to collect patient preferences for human alignment, which can help us generate a more user-friendly or customized lay definition. Also, we can use ChatGPT or LLAMA2-J-C-G2L to serve as the teacher and use DistilGPT2-

based systems to serve as the students, performing distillation to improve the performance of the small models post current supervisor-fine-tuning on README. Finally, more interactive ways need to be considered in the future to make the in-house system more user-friendly and patient-centric.

Consider Privacy Implications, LLMs (especially third-party APIs like ChatGPT) may raise privacy concerns when conducting patient education, which may violate HIPAA regulations. In this study, we manually annotated lay definitions on publicly available MedJEx jargon terms and obtained general definitions from accessible UMLS. We also make AI-synthetic data to help training since synthetic data generation is an active field in the clinical domain especially to overcome privacy concerns (Pereira et al., 2022; Shafquat et al., 2022; Mishra et al., 2023). The trained in-house system can be deployed on the patient's mobile to avoid patient data leaving the local area, which can better protect the patient's privacy and security. Consider Biases, LLMs trained on large amounts of text data may inadvertently capture and reproduce biases present in the data. Therefore, an in-house system trained on our data (whether expert annotation or AI synthetic) may perpetuate incorrect information or provide inaccurate answers. Finally, although we used UMLS-based RAG to reduce hallucinations, LLMs may still generate factual errors when conducting patient education.

References

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2023. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. *arXiv preprint arXiv:2303.03948*.

Herman Aguinis, Isabel Villamor, and Ravi S Ramani. 2021. Mturk research: Review and recommendations. *Journal of Management*, 47(4):823–837.

Amanuel Alambo, Tanvi Banerjee, Krishnaprasad Thirunarayan, and Michael Raymer. 2022. Entity-driven fact-aware abstractive summarization of biomedical literature. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 613–620. IEEE.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

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

Molly Baldry, Carol Cheal, Brian Fisher, Myra Gillett, and Val Huet. 1986. Giving patients their own records in general practice: experience of patients and staff. *Br Med J (Clin Res Ed)*, 292(6520):596–598.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Peter A. Boling. 2009. Care transitions and home health care. *Clinics in Geriatric Medicine*, 25(1):135–148. The Past, Present and Future of Home Health Care.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *arXiv preprint arXiv:2308.03253*.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. *arXiv preprint arXiv:2005.00701*.

Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, and Hong Yu. 2018. A natural language processing system that links medical terms in electronic health record notes to lay definitions: system development using physician reviews. *Journal of medical Internet research*, 20(1):e26.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Asad J Choudhry, Yaser MK Baghdadi, Amy E Wagie, Elizabeth B Habermann, Stephanie F Heller, Donald H Jenkins, Daniel C Cullinane, and Martin D Zielinski. 2016. Readability of discharge summaries: with what level of information are we dismissing our patients? *The American Journal of Surgery*, 211(3):631–636.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Cecilia Conrath Doak, Leonard G Doak, Gilbert H Friedell, and Cathy D Meade. 1998. Improving comprehension for cancer patients with low literacy skills: strategies for clinicians. *CA: A Cancer Journal for Clinicians*, 48(3):151–162.

Cecilia Conrath Doak, Leonard G Doak, and Jane H Root. 1996. Teaching patients with low literacy skills. *AJN The American Journal of Nursing*, 96(12):16M.

Adam EM Eltorai, Soha Ghanian, Charles A Adams Jr, Christopher T Born, and Alan H Daniels. 2014. Readability of patient education materials on the american association for surgery of trauma website. *Archives of trauma research*, 3(2).

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. *arXiv preprint arXiv:2210.09932*.

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.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

Zhe He, Zhiwei Chen, Sanghee Oh, Jinghui Hou, and Jiang Bian. 2017. Enriching consumer health vocabulary through mining a social q&a site: A similarity-based approach. *Journal of biomedical informatics*, 69:75–85.

Mohammed Ibrahim, Susan Gauch, Omar Salman, and Mohammed Alqahatani. 2020. Enriching consumer health vocabulary using enhanced glove word embedding. *arXiv preprint arXiv:2004.00150*.

Siddhartha Jonnalagadda, Luis Tari, Jorg Hakenberg, Chitta Baral, and Graciela Gonzalez. 2010. Towards effective sentence simplification for automatic processing of biomedical text. *arXiv preprint arXiv:1001.4277*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Amro Khasawneh, Ian Kratzke, Karthik Adapa, Lawrence Marks, and Lukasz Mazur. 2022. Effect of notes' access and complexity on opennotes' utility. *Applied Clinical Informatics*, 13(05):1015–1023.

Sanjay Krishnan and Eugene Wu. 2019. Alphaclean: Automatic generation of data cleaning pipelines. *arXiv preprint arXiv:1904.11827*.

Sari Kujala, Iiris Hörhammer, Akseli Väyrynen, Mari Holmroos, Mirva Nättiaho-Rönnholm, Maria Hägglund, and Monika Alise Johansen. 2022. Patients' experiences of web-based access to electronic health records in finland: Cross-sectional survey. *Journal of Medical Internet Research*, 24(6):e37438.

Sunjae Kwon, Zonghai Yao, Harmon S Jordan, David A Levy, Brian Corner, and Hong Yu. 2022. Medjex: A medical jargon extraction model with wiki's hyperlink span and contextualized masked language model score. *arXiv preprint arXiv:2210.05875*.

John P Lalor, Wen Hu, Matthew Tran, Hao Wu, Kathleen M Mazor, and Hong Yu. 2021. Evaluating the effectiveness of noteaid in a community hospital setting: Randomized trial of electronic health record note comprehension interventions with patients. *Journal of medical Internet research*, 23(5):e26354.

John P Lalor, Hao Wu, Li Chen, Kathleen M Mazor, and Hong Yu. 2018. Comprehenotes, an instrument to assess patient reading comprehension of electronic health record notes: development and validation. *Journal of medical Internet research*, 20(4):e139.

John P Lalor, Hao Wu, Kathleen M Mazor, and Hong Yu. 2023. Evaluating the efficacy of noteaid on ehr note comprehension among us veterans through amazon mechanical turk. *International Journal of Medical Informatics*, page 105006.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of Medical Informatics*, 2(01):41–51.

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. *arXiv preprint arXiv:2302.05574*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022a. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Bbac409.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022b. Readability controllable biomedical document summarization. *arXiv preprint arXiv:2210.04705*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Prakamya Mishra, Zonghai Yao, Shuwei Chen, Beining Wang, Rohan Mittal, and Hong Yu. 2023. Synthetic imitation edit feedback for factual alignment in clinical summarization. *arXiv preprint arXiv:2310.20033*.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Milad Moradi and Nasser Ghadiri. 2018. Different approaches for identifying important concepts in probabilistic biomedical text summarization. *Artificial intelligence in medicine*, 84:101–116.

Kevin M Nanna et al. 2009. Health literacy: Challenges and strategies. *Online Journal of Issues in Nursing*, 14(3):E1.

Andrew Ng, Dillon Laird, and Lynn He. 2021. Data-centric ai competition. *DeepLearning AI*.

Don Nutbeam. 2023. Artificial intelligence and health literacy—proceed with caution. *Health Literacy and Communication Open*, 1(1):2263355.

OpenAI. 2023. Gpt-4 technical report.

Karen Patrias and Dan Wendling. 2007. *Citing Medicine:*. Department of Health and Human Services, National Institutes of Health, US . . . .

Mayana Pereira, Sikha Pentyala, Anderson Nascimento, Rafael T de Sousa Jr, and Martine De Cock. 2022. Secure multiparty computation for synthetic data generation from distributed data. *arXiv preprint arXiv:2210.07332*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

François Remy and Thomas Demeester. 2023. Automatic glossary of clinical terminology: a large-scale dictionary of biomedical definitions generated from ontological knowledge. *arXiv preprint arXiv:2306.00665*.

Dean Schillinger, Margaret Handley, Frances Wang, and Hali Hammer. 2009. Effects of self-management support on structure, process, and outcomes among vulnerable patients with diabetes: a three-arm practical clinical trial. *Diabetes care*, 32(4):559–566.

Afrah Shafquat, Jason Mezey, Mandis Beigi, Jimeng Sun, and Jacob W Aptekar. 2022. A source data privacy framework for synthetic clinical trial data. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*.

Marina Solnyshkina, Radif Zamaletdinov, Ludmila Gorodetskaya, and Azat Gabitov. 2017. Evaluating text complexity and flesch-kincaid grade level. *Journal of social studies education research*, 8(3):238–248.

Amy J. Spooner, Natasha Booth, Tai-Rae Downer, Louisa Gordon, Adrienne P. Hudson, Natalie K. Bradford, Chris O'Donnell, Alanna Geary, Robyn Henderson, Cherie Franks, Aaron Conway, Patsy Yates, and Raymond J. Chan. 2019. Advanced practice profiles and work activities of nurse navigators: An early-stage evaluation. *Collegian*, 26(1):103–109.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. 2024. Readctrl: Personalizing text generation with readability-controlled instruction learning. *arXiv preprint arXiv:2406.09205*.

Jan Walker, Suzanne Leveille, Sigall Bell, Hannah Chimowitz, Zhiyong Dong, Joann G Elmore, Leonor Fernandez, Alan Fossa, Macda Gerard, Patricia Fitzgerald, et al. 2019. Opennotes after 7 years: patient experiences with ongoing access to their clinicians' outpatient visit notes. *Journal of medical Internet research*, 21(5):e13876.

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. *The VLDB Journal*, 32(4):791–813.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Zhichao Yang, Zonghai Yao, Mahbuba Tasmin, Parth Vashisht, Won Seok Jang, Feiyun Ouyang, Beining Wang, Dan Berlowitz, and Hong Yu. 2023. Performance of multimodal gpt-4v on usmle with image: Potential for imaging diagnostic support with explanations. *medRxiv*, pages 2023–10.

Zonghai Yao, Yi Cao, Zhichao Yang, Vijeta Deshpande, and Hong Yu. 2022. Extracting biomedical factual knowledge using pretrained language model and electronic health record context. In *AMIA Annual Symposium Proceedings*, volume 2022, page 1188. American Medical Informatics Association.

Zonghai Yao, Yi Cao, Zhichao Yang, and Hong Yu. 2023. Context variance evaluation of pretrained language models for prompt-based biomedical knowledge probing. *AMIA Summits on Translational Science Proceedings*, 2023:592.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Applied Sciences*, 10(21):7640.

Qing T Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *arXiv preprint arXiv:2303.10158*.

Xiaocheng Zhang, Zonghai Yao, and Hong Yu. 2023a. Ehr interaction between patients and ai: Noteaid ehr interaction. *arXiv preprint arXiv:2312.17475*.

Zihao Zhang, Zonghai Yao, Huixue Zhou, Hong Yu, et al. 2023b. Ehrtutor: Enhancing patient understanding of discharge instructions. *arXiv preprint arXiv:2310.19212*.

12620

A Annotation Guideline

The dataset was annotated for medical jargon and lay definition by six domain experts from medicine, nursing, biostatistics, biochemistry, and biomedical literature curation [16]. Herein, the annotators applied the following rules for identifying what was jargon and how to write a suitable lay definition:

Rule 1. Medical terms that would **not be recognized by about 4 to 7th graders**, or that **have a different meaning in the medical context than in the lay context (homonym)** were labeled. For example:

- accommodate: When the eye changes focus from far to near.
- antagonize: A drug or substance that stops the action or effect of another substance.
- resident: A doctor who has finished medical school and is receiving more training.
- formed: Stool that is solid.

Rule 2. Terms that are not strictly medical, but are **frequently used in medicine**. For example:

- "aberrant", "acute", "ammonia", "tender", "intact", "negative", "evidence"

Rule 3. When jargon words are **commonly used together, or together they mean something distinct or are difficult to understand from the individual parts quickly** were labeled. For example:

- vascular surgery: Medical specialty that performs surgery on blood vessels.
- airway protection: Inserting a tube into the windpipe to keep it wide open and prevent vomit or other material from getting into the lungs.
- posterior capsule: The thin layer of tissue behind the lens of the eye. It can become cloudy and blur vision.
- right heart: The side of the heart that pumps blood from the body into the lungs.
- intracerebral hemorrhage: A stroke.

Rule 4. Terms whose **definitions are widely known** (e.g., by a 3rd grader) do NOT need to be labeled. For example:

- "muscle", "heart", "pain", "rib", "hospital"

Rule 4.1 When in doubt, label the term. For example:

- "colon", "immune system"

B Evaluation of the Annotation

An observational study was performed to evaluate the annotators' reliability in identifying jargon and providing lay definitions, and assess the agreement of the dataset annotators with each other and laypeople.

B.1 Data Collection and Setting

For evaluation, twenty sentences were randomly selected from deidentified inpatient EHR notes in the EHR repository of one hospital affiliated with an anonymized institution. Sentences consisting only of administrative data, sentences less than ten words long, and sentences substantially indistinguishable from another sentence were filtered out.

Note that the annotators had never seen the sampled sentences. The twenty sentences were made up of 904 words in total. Common words were discarded so as not to inflate the calculated agreement. These consisted of all pronouns, conjunctions, prepositions, numerals, articles, contractions, months, punctuation, and the most common 25 verbs, nouns, adverbs, and adjectives. Terms occurring more than one time in a sentence were counted only once. Furthermore, multi-word terms were counted as single terms to ameliorate the double-counting issue. Two members of the research team determined multi-word terms by consensus. In this work, multi-word terms were defined as adjacent words that represented a distinct medical entity (examples: "PR interval", "internal capsule", "acute intermittent porphyria"), were commonly used together (examples: "hemodynamically stable", "status post", "past medical history") and terms that were modified by a minor word (examples: "trace perihepatic fluid", "mild mitral regurgitation", "rare positive cells", "deep pelvis"). After applying these rules, 325 candidate medical jargon terms and their lay definition were utilized. The laypeople comprised 270 individuals recruited from Amazon Mechanical Turk (MTurk) (Aguinis et al., 2021).

B.2 Annotation Reliability

The results showed that there was good agreement among annotators (Fleiss' kappa = 0.781). The annotators had high sensitivity (91.7%) and specificity (88.2%) in identifying jargon terms and providing suitable lay definitions as determined by the laypeople (the gold standard).

---

[16]The annotator agreement scores can be found in Appendix B.

Source	Definition	FKGL
UMLS	An endoscopic procedure that visualizes the upper part of the gastrointestinal tract up to the duodenum.	13.5
MedlinePlus	Esophagogastroduodenoscopy (EGD) is a test to examine the lining of the esophagus, stomach, and first part of the small intestine (the duodenum).	16.1
Wikipedia	Esophagogastroduodenoscopy (EGD), also called by various other names, is a diagnostic endoscopic procedure that visualizes the upper part of the gastrointestinal tract down to the duodenum. It is considered a minimally invasive procedure since it does not require an incision into one of the major body cavities and does not require any significant recovery after the procedure (unless sedation or anesthesia has been used).	20.9
Google	An EGD is a procedure in which a thin scope with a light and camera at its tip is used to look inside the upper digestive tract – the esophagus, stomach, and first part of the small intestine, called the duodenum. It's also called an upper endoscopy, or an esophagogastroduodenoscopy.	13.2
README	[Esophagogastroduodenoscopy] A procedure that looks at the food pipe, stomach, and the first part of the small bowel.	5.6

Table 4: Definitions of Esophagogastroduodenoscopy from various sources.

README-version	Dataset Description	DataPoints
README-exp	(ehr context, jargon, lay def, general definition)	308,242
README-exp	(jargon, lay def, general definition)	51,623
README-exp_good	(ehr context, jargon, lay def, general definition)	113,659
README-exp_good	(jargon, lay def, general definition)	11,765
README-exp_bad	(ehr context, jargon, lay def, general definition)	177,140
README-exp_bad	(jargon, lay def, general definition)	39,856
README-syn	(ehr context, jargon, lay def, general definition)	177,140
README-syn	(jargon, lay def, general definition)	39,856
README-syn_good	(ehr context, jargon, lay def, general definition)	96,668
README-syn_good	(jargon, lay def, general definition)	96,668
README-syn_bad	(ehr context, jargon, lay def, general definition)	75,157
README-syn_bad	(jargon, lay def, general definition)	75,157

Table 5: The Dataset Statistics of Different README versions.

B.3 Details about Jargons and Lay Definitions Statistics in README-exp

B.4 GPT Running Details in Examiner step

To optimize computing resources, we streamlined our dataset by removing EHR context and focusing solely on unique data points with (medical jargon term, lay definition, general definition) format in this Examiner step [17]. This reduced our dataset from 308,242 to 51,623 data points. Upon processing these through Examiner, 11,765 were classified as 'good' quality general definitions, and 39,856

as 'bad'. We subsequently performed an SQL join operation, integrating the 'good' and 'bad' datasets with the previously removed EHR context data, which resulted in 113,659 'good' and 194,580 'bad' data points with (EHR context, medical jargon terms, lay definition, general definition) format. After eliminating duplicates, the final count for bad data points in R-exp_bad was 177,140.

C General Definition Retrieval and Preprocessing

UMLS (Lindberg et al., 1993) is a set of files and software that combines many health and biomedical vocabularies and standards to enable interoperability between computer systems. Given the
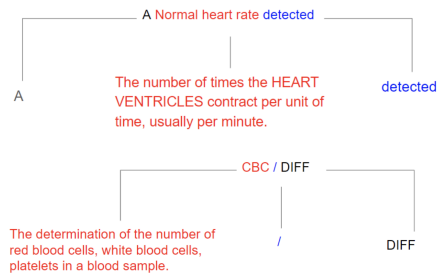
---
[17]This was done because, in many cases, we have the same jargon term, general definition from UMLS, and lay definition but different EHR contexts. We can reduce the amount of GPT running if we ignore the EHR context difference.

Figure 6: General definitions of two jargon examples.

complexity of medical terminology, it's noteworthy that some terms in UMLS are associated with multiple definitions. This reflects the reality that a medical term's meaning can vary depending on its contextual use. To accurately select the most appropriate general definition for each context, we utilized the SentenceBERT (Reimers and Gurevych, 2019) similarity score between lay definitions and all possible UMLS definitions available for a jargon term to identify the most fitting definition.

Jargon terms in our dataset varied in length and composition, with not all words necessarily having corresponding UMLS definitions. This variability necessitated the development of distinct approaches for different scenarios:

1. **Single-word Terms:** In cases where the jargon term comprised a single word, we either found a UMLS definition or none. Data points lacking a UMLS definition in this category were excluded.

2. **Two-word Terms:** For jargon terms composed of two words (word1 and word2), we considered several subcases:

   (a) Only word1 has a UMLS definition.
   (b) Only word2 has a UMLS definition.
   (c) Both word1 and word2 have UMLS definitions.
   (d) The phrase "word1 word2" has a collective UMLS definition.

3. **Terms with More than Two Words:** Similar to the case with two words, we can have cases where few words have UMLS definitions in the jargon term and few do not.

Our solution to address these scenarios involved a unified approach, given the impracticality of tackling each case individually. As illustrated in Figure 6, we extracted the UMLS definitions for all phrases within a jargon term. This process yielded a list of strings, providing a comprehensive general definition for each term. We then concatenate these strings in a comma-separated manner to get one

string used as a general definition. There might not be a UMLS definition for all the words in the jargon term. In that case, we are concatenating the word directly.

The initial README dataset has a lot of unusable data. There are many reasons why a data point might not be useful. The main reasons are (i) the lay definition provided by the annotator is not correct or missing. (ii) The jargon term is missing. (iii)The general definition obtained from UMLS is too scientific or not relevant to the jargon term. This was done by making sure there were no empty column values and removing parsing issues of comma-separated files. Using this data for training would be bad for the training as these data points will make the model worse. So these data points are not considered. This preliminary cleaning brings out data from 350K to 308K.

D Discussion on Concept Ambiguity

An important aspect that merits discussion is the potential ambiguity of medical concepts and its impact on generating lay user explanations. The inherent complexity and context-dependency of medical terms can create challenges in crafting universally understandable definitions, potentially leading to patient misinterpretations. Our approach to addressing the ambiguity of medical concepts is rooted in the Word-Sense Disambiguation (WSD) phase, as elaborated in related works such as (Kwon et al., 2022). The WSD phase links ambiguous medical terms to accurate, disambiguated definitions from medical concept dictionaries. These definitions serve as the foundation for generating lay definitions suitable for patient understanding, especially when dictionary definitions lack readability. In our study, the WSD phase is implicitly managed during the general definition retrieval stage using Scispacy and UMLS tools.

E Discussion on EAE Pipeline Efficacy

The primary objective of the EAE pipeline is to validate the effectiveness of Scispacy + UMLS for Word Sense Disambiguation (WSD), rather than addressing quality issues in the expert-annotated lay definitions. While quality checks and filtering for (jargon, context, lay definition) do not necessitate an LLM, they are essential for (jargon, context, lay definition, general definition) due to the noise introduced by Scispacy + UMLS tools. It is crucial to note that this noise is outside the scope of this

paper.

An important aspect that merits discussion is the efficacy of the EAE pipeline design, particularly regarding the high number of data points categorized as R-exp_bad or R-syn_bad in Table 5. Ideally, a more efficient pipeline would achieve a higher number of R-exp_good data points, thereby reducing the need for additional verification rounds. However, our current settings prioritize reducing false positives, even if this results in an increased number of false negatives. Strict AI verification is essential to mitigate false positives, which could lead to patient misunderstandings if incorrect definitions are generated. Given the large size of the original README dataset, the harm of false negatives is more acceptable.

Therefore, we have intentionally set a stricter standard for AI during prompting (as shown in Table 9). This approach may classify some good-quality data as bad, but it ensures that any data passing the Examiner stage is of high quality. This is corroborated by the high human agreement rates observed in Appendix F. Consequently, the final small human verification step is manageable without significantly increasing the workload.

The EAE pipeline and related prompts effectively detect and filter (jargon, context, lay definition, general definition) with minimal human effort, ensuring a sufficient quantity of valid data for subsequent training. In scenarios where the dataset is smaller or where more false positives are acceptable, the sensitivity of the AI examiner may need to be adjusted. Nevertheless, the three stages of the E→A→E pipeline are crucial for maintaining data quality in our task, as highlighted in Table 1, and can be extended to other similar scenarios.

F   Data Quality Checking and Train/Eval/Test Split after EAE Pipeline

After the end of the EAE, we use R-exp_good and R-syn_good as high-quality data for our system. The dataset was split into two categories: human examination data (which will also be used as final evaluation and test data since medical experts examine this split), and training data, where we ensure the medical jargon in the human examination will not appear in the training split. We sampled 500 medical jargon terms each from the R-exp_good and R-syn_good datasets. Therefore the human examination split consisted of 1000 unique terms,

each accompanied by general and lay definitions to be rated based on two criteria:

1. Hard Correlation: Marked 'Yes' if the lay definition closely rephrases or shares significant wording with the general definition, implying comprehensibility without advanced medical knowledge.
2. Soft Correlation: Marked 'Yes' if the general definition accurately represents the term but is slightly contextually misaligned; marked 'No' if the definition is incorrect or overly verbose, complicating the derivation of a lay definition.

Here is one example for **Term** 'von Willebrand disease':

Expert Definition(lay definition): A bleeding disorder. It affects the blood's ability to clot

General definition: Hereditary or acquired coagulation disorder characterized by a qualitative or quantitative deficiency of the von Willebrand factor. The latter plays an important role in platelet adhesion. Signs and symptoms include bruises, nose bleeding, gum bleeding following a dental procedure, heavy menstrual bleeding, and gastrointestinal bleeding.

Although the lay definition is not incorrect, it is very wordy and complex and matches less to the lay definition. We consider this soft correlated and not hard correlated. A Hard Correlation automatically implies a Soft Correlation.

Two medical students [18] help us finish this human examination. Our findings revealed that 88% of R-exp_good and 100% of R-syn_good met the Hard Correlation criteria. There is a 100% soft Correlation criteria for both R-exp_good and R-syn_good. After correcting individual invalid data (e.g., those cases where Soft is not satisfied), we used this human examination dataset as evaluation and test data (in a 1:1 ratio).

G   Factuality metrics: UMLS-F1

The assessment of factual accuracy in generated lay definition leverages the UMLS concept overlap metric. The Unified Medical Language System (UMLS), established by (Bodenreider, 2004), significantly contributes to the biomedical domain's interoperability. It achieves this by amalgamating and disseminating a comprehensive collection of biomedical terminologies, classification systems, and coding standards from many sources. By doing

---

[18]Both have hospital internship experience

In this task, we ask for your expertise in generating the corresponding lay definition from the medical jargon. Mainly, we provide the target medical jargon term. We need you to generate a lay definition for this jargon term.

Example:
jargon term: [TERM]
lay definition: [DEFINITION]

jargon term: [TERM]
lay definition:

Table 6: One shot prompt for experiment set-1.

so, UMLS aids in reconciling semantic variances and representational disparities found across different biomedical concept repositories.

For the identification and alignment of medical named entities within texts to their corresponding biomedical concepts in UMLS, we employed the Scispacy library [19]. Scispacy excels in identifying and clarifying entities, thus facilitating the accurate association of named entities found in lay definitions with the relevant UMLS concepts. This capability is critical for evaluating the lay definitions' factual accuracy and is used by recent related work (Adams et al., 2023).

The analytical process for lay definitions utilizes metrics of precision and recall. Precision represents the ratio of concepts present in both the generated and reference lay definitions, serving as a measure of the generated lay definition's factual correctness. In contrast, recall evaluates how well the information in the generated lay definition matches the intended content, reflecting the relevance of the presented information.

To calculate these metrics, we consider the concept sets from both the reference lay definition (C_{ref}) and the generated lay definition (C_{gen}). The formulas for recall and precision are as follows:

$$\text{Recall} = \frac{|C_{ref} \cap C_{gen}|}{|C_{ref}|}$$

$$\text{Precision} = \frac{|C_{ref} \cap C_{gen}|}{|C_{gen}|}.$$

The F1 score, derived from the above precision and recall values, is reported to provide a balanced measure of the generated lay definition's accuracy and relevance.

[19]We used the Scispacy *en_core_sci_lg* model.

H   More Experimental Settings

We use the following base models: GPT-2 [20] (Radford et al., 2019), DistilGPT2 [21], BioGPT [22] (Luo et al., 2022a), and Llama2 [23] (Touvron et al., 2023) in our experiments. We trained the base models on the different README dataset variants with Supervised Fine-tuning for 100000 steps (batch size 8) [24]. In all our evaluations, we used a beam size of 4, no-repeat-ngram-size=2, and minimum length and maximum length of sentences were set as (10, 100). We used five different random seeds to sample training data for all our experiments, and the scores reported in the tables are the average of these random seeds.

Here, we provide an example for jargon2lay(J2L), jargon+context2lay(J+C2L), jargon+gen2lay(J+G2L), and jargon+context+gen2lay(J+C+G2L) for easier understanding. Let's assume we one data point in README dataset with (jargon, EHR context, lay def, general def) format:

- **jargon:** EGD

- **lay def:** [esophagogastroduodenoscopy] A procedure that looks at the food pipe, stomach, and the first part of the small bowel.

- **EHR context:** [* * 11 - 22 * *] EGD Grade I varices - ablated [* * 11 - 22 * *] sigmoidoscopy friability , reythema , congest and abnormal vasularity in a small 5 mm area of distal rectum .

- **general def:** ['An endoscopic procedure that visualizes the upper part of the gastrointestinal tract up to the duodenum.']

[20]https://huggingface.co/gpt2
[21]https://huggingface.co/distilgpt2
[22]https://huggingface.co/microsoft/biogpt
[23]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[24]We did all the experiments with 1 NVIDIA Tesla RTX 8000 GPU - 40 GB memory, with Adam optimizer – betas=(0.9,0.999), epsilon=1e-08, learning rate=5e-04.

The input of different settings J2L, J+C2L, J+G2L, J+C+G2L can be found on Table 8.

I More Experimental Designs and Results

In Section 4, our experimental design focuses on evaluating the quality of outputs generated by different systems. Here, "quality" is measured by two main criteria: the overall similarity between the system's output and the ground truth lay definition (e.g., ROUGE or METEOR), and the presence or absence of factual inaccuracies in the generated lay definition (UMLS-F1). In this appendix section, we will focus more on readability. Specifically, we follow the ReadCtrl (Tran et al., 2024) to explore how the README dataset aids models in generating outputs with more controllable readability. Therefore, we conducted instruction-following experiments with GPT-3.5 (few-shot), GPT-4 (few-shot), Claude3-opus (few-shot), Llama2-chat (few-shot), and Llama2-README-finetuning (few-shot). The prompt used was: *"Given an input jargon term and general definition, please output a lay definition with a readability score around target readability [X]."*

where [X] was replaced by FKGL 1-12. A good instruction following should output lay definitions with readability scores similar to the target FKGL.

[X]	GPT-3.5	GPT-4	Claude3	Llama2	Ours
1	7.1410	6.0820	7.5364	10.4919	3.8800
2	6.7836	6.5907	7.2024	10.4365	4.5106
3	6.7412	7.4916	7.9721	10.4571	5.5185
4	7.5948	7.7300	8.4284	10.9103	6.1644
5	7.9722	8.1104	9.7814	10.6538	6.6462
6	8.7160	8.4608	10.9537	10.3240	6.9269
7	9.0761	8.9479	11.1111	10.3477	7.4499
8	10.0191	9.6390	13.3369	10.6044	8.2328
9	11.3319	11.0364	14.7280	10.5953	8.9487
10	12.4661	11.9267	16.5011	10.0969	9.5266
11	13.4467	12.4227	16.8663	10.6457	10.0348
12	13.2357	13.3720	17.2713	10.2263	10.5039

Table 7: Mean FKGL Scores for Each Model

As illustrated in Table 7 and Figure 7, our investigation across a range of state-of-the-art LLMs shows varying degrees of compliance with readability-controlled instructions. Mainstream models like GPT-3.5, GPT-4, and Claude3 are upward but far from the perfect curve, indicating they can follow instructions but not precisely. Llama2 does not show an upward trend, suggesting it cannot follow instructions. In contrast, Llama2-README closely follows the perfect curve, indicating precise instruction-following capability.
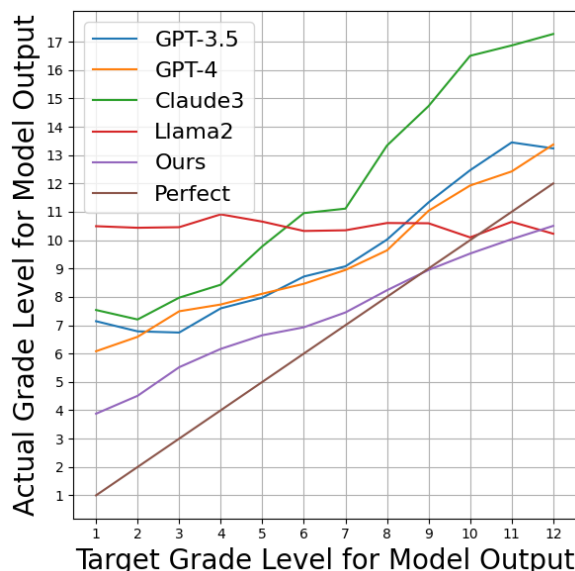

Figure 7: ReadCtrl (Tran et al., 2024) instruction following ability using README dataset.

These results suggest that the README dataset contains sufficiently diverse readability information, making it highly useful for controllable text generation, particularly in readability control. This capability has significant potential for personalized patient education and represents a promising future research direction.

jargon2lay(J2L):

In this task, we ask for your expertise in generating the corresponding lay definition from the medical jargon. Mainly, we provide the target medical jargon term. We need you to generate a lay definition for this jargon term.

jargon term: EGD

lay definition:

jargon+context2lay(J+C2L):

In this task, we ask for your expertise in generating the corresponding lay definition from the medical jargon. Mainly, we provide the target medical jargon term along with the contextual snippets in which they appear in the text. We need you to generate a lay definition for this jargon term.

jargon term: EGD

context: [* * 11 - 22 * *] EGD Grade I varices - ablated [* * 11 - 22 * *] sigmoidoscopy friability , reythema , congest and abnormal vasularity in a small 5 mm area of distal rectum .

lay definition:

jargon+gen2lay(J+G2L):

In this task, we ask for your expertise in generating the corresponding lay definition from the medical jargon. Mainly, we provide the target medical jargon term. In addition, we also provide a definition from the dictionary for reference. We need you to generate a lay definition for this jargon term.

jargon term: EGD

dictionary definition: ['An endoscopic procedure that visualizes the upper part of the gastrointestinal tract up to the duodenum.']

lay definition:

jargon+context+gen2lay(J+C+G2L):

In this task, we ask for your expertise in generating the corresponding lay definition from the medical jargon. Mainly, we provide the target medical jargon term along with the contextual snippets in which they appear in the text. In addition, we also provide a definition from the dictionary for reference. We need you to generate a lay definition for this jargon term.

jargon term: EGD

context: [* * 11 - 22 * *] EGD Grade I varices - ablated [* * 11 - 22 * *] sigmoidoscopy friability , reythema , congest and abnormal vasularity in a small 5 mm area of distal rectum .

dictionary definition: ['An endoscopic procedure that visualizes the upper part of the gastrointestinal tract up to the duodenum.']

lay definition:

Table 8: The prompt of different settings J2L, J+C2L, J+G2L, J+C+G2L.

README-EAE

> **inputs :** README-exp
>
> **output :** Final README dataset with best general definitions and lay definitions
>
> $Initialize :$
> $exam_p = examinerpromptforgpt3;$
>
> $aug_p =$ $augmenterpromptsforChatGPT;$
>
> **foreach** $datapoint : README - exp$ **do**
>> **if** $exam_p(datapoint) == yes$ **then**
>>> R-exp_good.add(datapoint);
>>
>> **end**
>> **else**
>>> R-exp_bad.add(datapoint);
>>
>> **end**
>
> **end**
> **foreach** $datapoint : R - exp\_bad$ **do**
>> $temp = aug_p(datapoint);$
>> R-syn.add(temp);
>> **if** $exam_p(temp) == yes$ **then**
>>> R-syn_good.add(temp);
>>
>> **end**
>> **else**
>>> R-syn_bad.add(temp);
>>
>> **end**
>
> **end**
> **return** *R-syn_bad*;

Algorithm 1: Algorithm for Data Cleaning

Prompt

```
Examiner - 1

model = "gpt-3.5-turbo(ChatGPT API)"
[This examiner prompt does not use context as discussed in section 3.2.1]

Decide whether the general definition is correct.

If we can generate the lay definition from the general definition then answer is yes.

term : mg
general definition : this is short for milligram which is 1/1000 of a gram usually considered a small amount.
lay definition : A tiny amount of something, usually a drug.
answer : yes

term : vitamin c
general definition : [`A nutrient that the body needs in small amounts to function and stay healthy. Vitamin C helps fight infections, heal wounds,
    ↪  and keep tissues healthy. It is an antioxidant that helps prevent cell damage caused by free radicals (highly reactive chemicals).
    ↪  Vitamin C is found in all fruits and vegetables, especially citrus fruits, strawberries, cantaloupe, green peppers, tomatoes, broccoli,
    ↪  leafy greens, and potatoes. It is water-soluble (can dissolve in water) and must be taken in every day. Vitamin C is being studied in the
    ↪   prevention and treatment of some types of cancer.']
lay definition : A nutrient needed by the body to form and maintain bones, blood vessels, and skin.
answer : yes

term : nodule
general definition : [`A small lump, swelling or collection of tissue.']
lay definition : A growth or lump that may be cancerous or not.
answer : yes

term : qd
general definition : [`Occurring or done each day.']
lay definition : Every day.
answer : yes

If the general definition contains many words from the term then answer is no.

term : prochlorperzine
general definition : [`prochlorperzine', ` ']
lay definition : A drug used to prevent or reduce nausea and vomiting.
answer : no

term : mg
general definition : [`mg']
lay definition : A tiny amount of something, usually a drug.
answer : no

If the lay definition can not be generated by the general definition then answer is no.

term : Virt - Vite
general definition : [`Virt', ` - ', `The determination of the amount of Vitamin E present in a sample.']
lay definition : A mix of vitamins. It provides vitamin B-6, vitamin B-12 and folic acid to people who do not have enough of these for good health.
answer : no
```

```
Augmenter

system_prompt = "your job is to generate a general definition of the term."
model="gpt-3.5-turbo(ChatGPT API)",
messages=[
      {"role": "system", "content": system_prompt},
      {"role": "user", "content": ""},
      {"role": "user", "content": "term : incisional."},
      {"role": "assistant", "content": "general definition : An intentional cut made to an individual's body with the intent of performing a
             ↪ diagnostic or therapeutic intervention."},
      {"role": "user", "content": "term : PO"},
      {"role": "assistant", "content": "general definition : Of, or relating to, or affecting, or for use in the mouth.."},
      {"role": "user", "content": prompt_t}
   ]
```

```
Examiner - 2

model = "gpt-3.5-turbo(ChatGPT API)"
exactly same as that of Examiner - 1
```

Table 9: All Examiner-Augmenter-Examine prompts.