Making Health Knowledge Accessible Through Personalized Language Processing

Yue Guo

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Trevor A. Cohen, Chair

Gondy Leroy

Sheng Wang

Lucy Lu Wang

Program Authorized to Offer Degree:
Department of Medical Education and Biomedical Informatics

University of Washington

Abstract

Making Health Knowledge Accessible Through Personalized Language Processing

Yue Guo

Chair of the Supervisory Committee:
Trevor A. Cohen
Department of Medical Education and Biomedical Informatics

The 2019 COVID pandemic exposed the difficulties the general public faces when attempting to use scientific information to guide their health-related decisions. **Though widely *available* in scientific papers, the information required to guide these decisions is often not *accessible***: medical jargon, scientific writing styles, and insufficient background explanations make this information opaque to non-experts. Consequently, there is a pressing need to deliver scientific knowledge in lay language, which has motivated my research on automated plain language summary generation to make health information more accessible.

The main challenges addressed in this thesis are limited data, generating background knowledge, lack of evaluation metrics, and the need for personalization. To tackle the limited data challenge, I introduce the task of automated generation of plain language summaries (PLSs) of biomedical scientific reviews and construct the Corpus for Enhancement of Lay Language Synthesis (CELLS), the largest and most diverse dataset for PLS in the medical domain. For generating background knowledge, I explore methods for Retrieval-Augmented Lay Language (RALL) generation, augmenting state-of-the-art text generation models with information retrieval from various sources.

A key part of this process has been evaluating existing metrics to see if they effectively measure performance for this task, and considering if there might be better options. To address the lack of evaluation metrics, I present APPLS, the first granular testbed for analyzing evaluation metric performance for PLS, and introduce POMME, a new metric that

employs language model perplexity to assess text simplicity.

Finally, I broaden the discussion beyond health information - exploring how we can personalize and improve communication across different domains. Grounded in the real-world setting of interdisciplinary reading, this research offers insights into features and methods for the novel task of integrating personal data into scientific jargon identification.

In conclusion, my thesis provides a comprehensive approach to making biomedical literature more accessible and understandable for health consumers by addressing key challenges in developing automated PLS generation systems. The contributions span data collection, method development, evaluation metric design, and personalization, paving the way for more effective communication of health information to the general public.

ACKNOWLEDGMENTS

I would like to express my heartfelt gratitude to Trevor Cohen for his unwavering support and encouragement throughout my research. Your guidance and patience have been invaluable, and your mentorship has inspired me to pursue a career in academia, aspiring to be an advisor like you.

I extend my sincere thanks to my committee members: Lucy Lu Wang, Gondy Leroy, and Sheng Wang. Your collaboration and mentorship have been instrumental in my growth as a researcher. I am grateful for your valuable suggestions and advice, not only in my research but also in my career development. A special thanks to Lucy for going above and beyond in providing me with everything I needed and for always thinking ahead.

I also want to acknowledge the professors at UW, particularly Anne Turner, for her emotional support during the challenging times of the pandemic. To Todd McNutt and Harry Quon from JHU Radiation Oncology, thank you for introducing me to the world of AI applications in medicine and nurturing my interest in this field. I am grateful to professor Youlin Qiao from the Chinese Academy of Medical Sciences and Peking Union Medical College for being a role model and teaching me the responsibilities of a physician-scientist in saving lives. And to Hao Zhang from the Red Cross Society of China, thank you for introducing me to the International Red Cross and Red Crescent, showing me that I can make a difference in the world. Lastly, I want to express my gratitude to Tal August, Joseph Chee Chang, and Maria Antoniak for the fantastic internship experience last summer.

I am grateful to my collaborators: Qiu Wei, Yizhong Wang, Changye Li, Liwei Jiang, Xiruo Ding, and George Xu. A special thanks to Jingyi Xie and Shihan Xu for their constant encouragement.

I acknowledge the funding provided by the US National Library of Medicine [grant number R21LM013934], which supported the works contained in this thesis.

Finally, I want to thank my parents, Lianhe Guo and Shuyun Fu, my sister Lei Guo, and my beloved furry friends Taotao, Loki, and Sylvie. Most importantly, I am deeply grateful to my husband, Hao Peng, for his love and support throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

Table 1: List of Abbreviations

Abbreviation	**Definition**
PLS	Plain Language Summary
CELLS	Corpus for Enhancement of Lay Language Synthesis
RALL	Retrieval-Augmented Lay Language
UMLS	Unified Medical Language System
APPLS	Testbed for plain language summary
POMME	Evaluation metric for plain language summary simplicity
LLM	Large Language Model
NLP	Natural Language Processing

Chapter 1

INTRODUCTION: PLAIN LANGUAGE SUMMARY GENERATION

1.1 Motivation

Health literacy, the ability to comprehend scientific concepts and research in the medical domain, is essential for making informed health decisions and ensuring positive treatment outcomes (Parker et al., 1999). The internet has greatly expanded access to health information for the general public. However, the public faces challenges in understanding this information due to difficulties identifying credible sources (Howes et al., 2004), unfamiliar medical jargon (Korsch et al., 1968), and the complex structure of professional language (Friedman et al., 2002). The evolving nature of health knowledge further complicates the task of identifying the most current information. The COVID-19 pandemic has highlighted these challenges, particularly in interpreting biomedical literature without specialized training. This thesis aims to address the gap between the increasing availability of health information and the public's difficulty in understanding it by translating biomedical literature into plain language (example in Figure 1.1).

Scientific Abstract: The therapeutic and economic benefits of continuous positive airway pressure (CPAP) for the treatment of obstructive sleep apnoea syndrome (OSAS) have been established in middle-aged people…
Plain Langauge Summary: Obstructive sleep apnoea (OSA) is a condition in which the walls of the throat relax during sleep, repeatedly blocking the airway for a few seconds. OSA affects up to one in five older people, so as more people get older the best treatment needs to be found. OSA can be treated with continuous positive airway pressure (CPAP), such as nasal mask or full face mask. CPAP is already known to help middle-aged people with OSA…

Figure 1.1: Example scientific abstract and plain language summary pair. Besides text simplification, plain language summarization involves background explanation, such as definitions, motivation, and examples.

1.2 Task Definition

To increase the availability of health information, I define the task of **plain language summary (PLS) generation**: automated generation of lay language summaries of biomedical scientific reviews. It is a type of translation problem between the language of healthcare professionals, and that of the general public. In my characterization of this task, I have identified four key challenges for its automation.

1.2.1 Challenge 1. Limited Data

Automated summarization of scientific documentation has been a long-standing research topic. Advances have been achieved in tasks including abstract generation for academic papers (Cohan et al., 2018a), citation sentence generation (Luu et al., 2020), or extreme summarization of the entire document (Cachola et al., 2020b). Despite these advances, limited progress has been made in the medical domain, primarily due to challenges with limited paired data consisting of biomedical scientific reviews and their corresponding PLSs, an issue addressed by this doctoral work. Paired data is crucial for this task, as it demonstrates complex health information mapped to simpler alternatives. With a substantial amount of paired training data, language models can learn to effectively translate technical health information into content that is more easily understood by a broader audience.

1.2.2 Challenge 2. Generating Background Explanations

Prior work has framed PLS generation as a summarization or simplification task. However, adapting biomedical text for the lay public includes the distinct task of *background explanation*: adding external content in the form of definitions, motivations, or examples to enhance comprehensibility. In Figure 1.1, the continuous positive airway pressure provides a concrete example so that readers can link it with a more familiar concept (mask). The external content is often not included in scientific abstracts, as it is usually already known to those in the field.

1.2.3 Challenge 3. Lack of Tailored Evaluation Metrics

Reliable and validated metrics are crucial for evaluating improvements in PLS generation systems. While prior studies relied on readability score or costly human evaluation, these have clear gaps. Readability scores using word and sentence length do not capture key aspects of quality, like understandibility, coherence, or fluency. Human assessment is inconsistent and not scalable. Automated scores are needed as they are fast and accurate. However, those automated scores were not designed specifically for PLS generation. And no previous study has validated whether they can effectively measure summary quality for this task.

1.2.4 Challenge 4. Personalization

While models trained on population data have seen some promising healthcare applications, personalized AI is key to unlocking healthcare's full potential. An important challenge to communicating knowledge across domains is aligning on a shared vocabulary. As science becomes more specialized, so too does its terminology raising the barrier of learning across disciplines. I envision systems that can identify whether specialized terminology will be unfamiliar to an individual reader. NLP techniques have been developed to identify and simplify scholarly jargon. The majority of these techniques use a corpus of documents as a proxy for what a reader knows (e.g., Wikipedia contains words known to a general audience). However, an individual's specific background knowledge also plays a role in determining their familiarity with a word. For example, a long-term diabetes patient might know what insulin (a treatment for diabetes) is, but a newly-diagnosed patient might need an explanation of it. Information on reader's background could help determine what they know, and what they need explained.

1.3 Overview of Chapters

This dissertation focuses on addressing key challenges I have identified in making health knowledge accessible through personalized language processing. The four main challenges addressed are limited data, generating background explanations, lack of evaluation metrics,

and the need for personalization.

To address the challenge of limited data, in Chapter 2 (Guo et al., 2021), I introduce the novel task of automated generation of PLSs of biomedical scientific reviews. The chapter describes my construction of a dataset of 7,805 summaries, and a qualitative analysis of the inherent NLP challenges in PLS generation, including identification of the previously unrecognized sub-task of background explanation generation. Furthermore, I investigate the hypotheses 1) that language models are capable of generating PLSs that are comparable to human-authored summaries; and 2) further in-domain pre-training will enhance generation performance.

The problem of limited data is further addressed by the work described in Chapter 3 (Guo et al., 2022a), in which I construct the Corpus for Enhancement of Lay Language Synthesis (CELLS), the largest and most diverse dataset for PLS in the medical domain, aggregating 62,886 pairs of scientific abstracts with corresponding PLSs from 12 journals spanning various biomedical domains. Additionally, this chapter also addresses the challenge of generating background knowledge. I hypothesize that the Retrieval-Augmented Lay Language (RALL) method would improve background explanation generation. I augment state-of-the-art text generation models with information retrieval of term definitions from the UMLS and Wikipedia, or embeddings of explanations from Wikipedia documents. I also test the ability of two Large Language Models (LLMs), open-source Llama 2 and closed-source GPT-4 , for background explanation, with and without external knowledge from Wikipedia.

To address the lack of evaluation metrics, in Chapter 4 (Guo et al., 2023a), I present APPLS, the first granular testbed for analyzing evaluation metric performance for PLS. To assess how well existing metrics capture the key aspects of PLS, I define four criteria that a PLS metrics should be sensitive to: informativeness, coherence, faithfulness, and simplification. I then use tailored perturbation procedures to analyze the responsiveness of 15 metrics to changes affecting these criteria, including the most widely used metrics in text simplification and summarization, and recently-proposed prompt-based evaluation. Lastly, I introduce POMME, a new metric that employs language model perplexity to assess text simplicity, and validate its performance in our testbed and two additional simplification

datasets.

For the challenge of personalization, in Chapter 5 (Guo et al., 2024), I introduce the task of *personalized scholarly jargon identification*, grounded in the setting of interdisciplinary reading. I validate our setting with an initial study on interdisciplinary reading, collecting a dataset of over 10K individual familiarity ratings and information needs from 11 computer science researchers about terms drawn from 100 out-of-domain abstracts. I enumerate features representing an individual's background knowledge based on papers they have written and read. Using these features and our dataset, I investigate baselines for estimating term familiarity, including regression models and prompt-based approaches using LLMs.

In conclusion, my thesis addresses the key challenges in developing automated PLS generation systems for biomedical scientific literature. The contributions span data collection, method development, evaluation metric design, and personalization, providing a comprehensive approach to making biomedical literature more accessible and understandable for health consumers.

Chapter 2

CHALLENGE 1. LIMITED DATA

To address the limited data in plain language summarization, in the work described in this chapter, I start with constructing the dataset and conducting analyses of the various challenges in performing the PLS generation task. I experiment with state-of-the-art summarization models as well as several data augmentation techniques, and evaluate their performance using both automated metrics and human assessment.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. "Automated lay language summarization of biomedical scientific reviews." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 1, pp. 160-168. 2021.

2.1 Introduction

The ability to understand scientific concepts, content, and research in the medical domain is defined as health literacy (Parker et al., 1999), which is crucial to making appropriate health decisions and ensuring treatment outcomes. The development of the internet has enabled the general population to access health information, greatly expanding the volume of available health education materials. However, challenges arise in reading and understanding these materials because of the inability to identify credible resources (Howes et al., 2004), unfamiliarity with medical jargon (Korsch et al., 1968), and the complex structure of professional language (Friedman et al., 2002). Furthermore, knowledge in the health domain evolves over time, presenting laypeople with the additional challenge of discerning the most up-to-date information. The COVID-19 pandemic has cast a spotlight on the challenges in the general public's ability to obtain, interpret, and apply knowledge to guide their health-related behavior. These challenges are exemplified by difficulties in interpreting ar-

ticles from the biomedical literature, for those without specific training in this domain. Our project aims to bridge the gap between the increasing availability of health information and the difficulty the public has understanding it. To do so, we confront the task of rendering the biomedical literature comprehensible to the lay public, which can be framed as a type of translation problem: from the language of healthcare professionals to plain language.

A systematic review, such as those in the widely-used Cochrane Database of Systematic Reviews[1] (CDSR), is a type of biomedical scientific publication that synthesizes current empirical evidence for a research question, to identify the strongest evidence to inform decision making while minimizing bias. Of importance for the current research, reviews in the CDSR include lay language summaries, written by review authors or Cochrane staff. In this paper, we introduce the novel task of automated generation of lay language summaries of biomedical scientific reviews. We introduce a dataset constructed by extracting 7,805 high-quality abstract pairs that consist of both abstracts intended for professional readers, and plain language versions written by domain experts. The *source* in the training dataset is the healthcare professional version of an abstract, with an average length of 714 words. The *target* is the corresponding plain language version, with an average length of 371 words.

This dataset is the first corpus developed to facilitate development of methods to improve document-level understandability of biomedical narratives for the general public, and covering a broad range healthcare topics. The closest parallel in the literature may be the manually annotated dataset, MSD, which was developed recently to improve bi-directional communication between clinicians and patients (Cao et al., 2020). MSD is focused on communication of clinical topics at the sentence level, and the accompanying work approached the problem using text style transfer and simplification algorithms. However, constraining the task to sentence level prohibits methods from considering the broader context in which a sentence occurs. While the text summarization community has developed various corpora (Allahyari et al., 2017) for document-level summarization tasks, current resources for the biomedical scientific domain are limited (Moradi and Ghadiri, 2019). Furthermore, the proposed plain language summarization task imposes additional challenges, such as termi-

[1] http:///www.cochranelibrary.com

nology explanation and sentence structure simplification, that are not required for general domain summarization (see Table 2.1 for an analysis of five additional task components). One important goal of our work is to meet the need for a dataset to support research into the task of generating summaries of biomedical professional literature that are comprehensible to the lay public.

To approach this task, we implemented several state-of-the art extractive and abstractive summarization models and evaluated them on the collected CDSR dataset. On account of the limited size of this dataset, we also applied intermediate pre-training (both out-of-domain on-task, and in-domain off-task) to the best abstractive model. We evaluated the utility of pre-training this model on CNN/DM (Nallapati et al., 2016), a much larger general domain summarization dataset. To provide the model with more domain-specific biomedical language, we pre-trained it on an unlabeled biomedical corpus of 300K abstracts from the PubMed database.

Standard automated metrics of summarization and readability were adopted to evaluate model performance. In addition, we used ratings of human evaluators to assess the generated summaries from several perspectives. The results suggest that the best-performing model can generate lay language summaries with promising quality and readability.

Our main contributions can be summarized as follows:

- We introduce the novel task of automated generation of lay language summaries of biomedical scientific reviews.

- We construct a dataset of 7,805 summaries and a qualitative analysis of the NLP challenges inherent in this task.

- We evaluate performance of state-of-the-art summarization models leveraging neural machine translation architectures, with and without data augmentation techniques, on this task.

- We conduct automated and human evaluation from multiple perspectives, showing that machine generated summaries can achieve promising quality and readability as

compared with reference summaries developed for the lay public by domain experts.[2]

[2]We release our code at https://github.com/qiuweipku/Plain_language_summarization

Category	Source	Target	BART+CNN/DM+PubMed	BART+PubMed
Removing Unnecessary Details	. . . A complete case analysis (i.e. participants who completed the study) among trials investigating CDAD (31 trials, 8672 participants) suggests that probiotics reduce the risk of CDAD by 60%. . . . (Goldenberg et al., 2017)	. . . Our results suggest that when probiotics are given with antibiotics the risk of developing CDAD is reduced by 60% on average. . . .	. . . We found that probiotics reduce the risk of developing CDAD by 60% . . .	. . . A complete case analysis (i.e. participants who completed the study) among trials investigating CDAD (31 trials, 8672 participants) suggests that probiotics reduce the risk of developing CDAD by 60%. . . .
Relevant Background Explanation	Hepatitis C is a major cause of liver-related morbidity and mortality. Standard therapy is ribavirin plus pegylated interferon to achieve undetectable level of virus in the blood, but the effect on clinical outcomes is controversial. . . . (Brok et al., 2010)	Globally about 170 million people are chronically infected with hepatitis C virus. Hepatitis C is a blood-borne virus and routes of transmission include intravenous drug use, mother-to-infant transmission, unsafe medical practices, high-risk sexual behavior, and blood transfusion.	Hepatitis C is a major cause of liver-related morbidity and mortality. Standard therapy is ribavirin plus pegylated interferons to achieve undetectable level of virus in the blood, but the effect on clinical outcomes is controversial. . . .	Hepatitis C is a major cause of liver-related morbidity and mortality. Standard therapy is ribavirin plus pegylated interferons to achieve undetectable level of virus in the blood, but the effect on clinical outcomes is controversial. . . .
Jargon Explanation	. . . We considered all randomised controlled trials (RCTs) comparing EVLA, endovenous RFA or UGFS with conventional surgery in the treatment of SSV varices for inclusion. . . . (Paravastu et al., 2016)	. . . We found three randomised controlled trials (clinical studies where people are randomly put into one of two or more treatment groups) that compared endovenous lasers (EVLA) with surgery. . . .	. . . We found three randomised controlled trials (clinical studies where people are randomly put into one of two or more treatment groups) that compared endovenous lasers (EVLA) with surgery. . . .	. . . We found three randomised controlled trials (clinical studies where people are randomly put into one of two or more treatment groups) that compared endvenous laser Ablation with surgery. . . .
Converting to Interrogative Sentence	. . . Non-resection versus resection for an asymptomatic primary tumour in patients with unresectable stage IV colorectal cancer . . . (Cirocchi et al., 2012)	. . . Should the primary cancer be surgically removed in asymptomatic patients with unresectable stage IV colorectal cancer? . . .	. . . Primary tumour resection versus no resection in asymptomatic patients with unresectable stage IV colorectal cancer who are treated with palliative chemo/radiotherapy	Primary tumour resection in asymptomatic patients with unresectable stage IV colorectal cancer who are treated with palliative chemo/radiotherapy . . .
Sentence Structure Simplification	. . . Abnormal blood flow patterns in fetal circulation detected by Doppler ultrasound may indicate poor fetal prognosis. . . . (Alfirevic et al., 2017)	. . . Doppler ultrasound detects changes in the pattern of blood flow through the baby's circulation. These changes may identify babies who have problems. . . .	. . . Abnormal blood flow patterns in the blood vessels of the unborn baby may indicate poor outcomes for the baby. . . .	. . . The aim of the review was to find out if using Doppler ultrasonography (ultrasound of the baby's heart and blood vessels) during pregnancy can improve outcomes for babies. . . .

Table 2.1: Typical transformation phenomena from *source* to *target*, and the corresponding generated summaries using two best-performing models. We only show part of the long text here for brevity.

2.2 Related Work

2.2.1 Scientific Document Summarization

Automated summarization of scientific documents has been a long-standing research topic (Paice, 1980; Teufel and Moens, 2002). Advances have been achieved through the development of high-quality datasets and evaluation tasks, including but not limited to abstract generation of academic papers (Cohan et al., 2018a), citation sentence generation (Luu et al., 2020) or extreme summarization of the entire document (Cachola et al., 2020b). In the medical field, Sarkar et al. (2011) explored extractive summarization of medical news articles. The Text Analysis Conference (TAC) 2014 Biomedical Summarization track introduced several subtasks to evaluate citation-based summaries. Our proposed task differs from this prior work in two ways: 1) our target summaries are in plain language, so the overall task requires other capabilities beyond summarization; 2) the target summaries in our dataset are considerably longer (see Table 2.2), which poses further challenges.

Existing summarization methods can be broadly categorized into extractive and abstractive approaches (Das and Martins, 2007). The former (Erkan and Radev, 2004; Cheng and Lapata, 2016a) select sentences from the original text, while the latter (Rush et al., 2015; Nallapati et al., 2016) can generate summaries using words that are not found in the original document. Our task is abstractive by nature. For scientific document summarization, various features (e.g. citation networks) can be used to improve performance. While augmentation of this sort is beyond the scope of the current work, we refer the interested reader to Altmami and Menai (2020) and Moradi and Ghadiri (2019) for a comprehensive overview of the relevant approaches.

2.2.2 Text Simplification

Text simplification (Shardlow, 2014) modifies the content or structure of a text to make it easier to understand. Unlike summarization, text simplification approximates the meaning of the original sentence without necessarily shortening it. Simplification techniques have been used in the biomedical domain for generating patient-centered radiology reports (Qenam et al., 2017), name entity recognition (Habibi et al., 2017), preprocessing for biomed-

ical interaction recognition (Baumgartner et al., 2008), syntactic parsing (Jonnalagadda et al., 2010), and simplification of medical journal text (Jonnalagadda et al., 2010). For lexical simplification, WordNet (Miller, 1995), the UMLS (Bodenreider, 2004) and Wiktionary (Zesch et al., 2008) are widely used as synonym resources to find and replace medical terms. Toward the goal of producing summaries of abstracts that are understandable to a lay audience, our task requires simplification at the document rather than the single sentence level, and combines this with summarization to achieve both shorter and more readily understandable summaries.

2.2.3 Pre-training and Transfer Learning

Large pre-trained neural networks have led to recent advances in performance across a broad range of NLP tasks (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019) - especially when the available labeled data are limited (Brown et al., 2020b), or there is a shift in domains involved (Hendrycks et al., 2020). Due to the complexity of our task and the relatively small size of our data, we posit that pre-training is a prerequisite to strong performance. Furthermore, recent work shows that adaptive pre-training with domain-relevant unlabeled data (Gururangan et al., 2020a) or task-relevant labeled data (Pruksachatkun et al., 2020) can further improve the performance of pre-trained models. The intermediate pre-training strategies we evaluated in our experiments are inspired by these findings, and we encourage efforts in finding additional useful data to further improve performance on this task.

2.2.4 Biomedical Domain Summarization

The most common document types used for summarization tasks in the biomedical domain are clinical notes, with the aim to reduce information overload for health practitioners (Pivovarov and Elhadad, 2015; Feblowitz et al., 2011; Mollá et al., 2011). This aim differs from our main objective: generating lay language summaries of biomedical scientific reviews for health consumers. Another related area of work concerns information retrieval from the internet, where the goal is to help consumers find (rather than interpret) health information (Goeuriot et al., 2020).

2.3 Dataset Description

2.3.1 The CDSR

The Cochrane Database of Systematic Reviews (CDSR) includes high-quality systematic reviews (Uman, 2011) in various health care domains that facilitate evidence-based medical decision making. For a systematic review, two independent reviewers will review eligible peer-reviewed papers, registered clinical trials, conference papers, or "grey literature"[3]; search for evidence on a clearly formulated question; extract data from the studies; and grade the quality of available data. According to the hierarchy of scientific evidence (Murad et al., 2016), systematic review is the most robust evidence supporting an argument. Of particular importance for the current work, a plain language version of each abstract accompanies its professional language counterpart. Of note, plain language summaries have been required from authors submitting a review since 2015. Prior to this, they were written by Cochrane staff with specialized training.

2.3.2 Dataset construction

We extracted 7,805 abstracts (*source*), paired with their plain language versions (*target*) from CDSR reviews available up to March 29, 2020. The original data is downloadable via the official API [4]. We only retained examples with source length between 300 to 1,000 words, and target length between 100 to 700 words. This resulted in a set of 5,195 source-target pairs which constitutes our training set, a further 500 abstract pairs as the validation set, and 1000 more as the test set.

2.3.3 Data analysis

Table 2.2 shows the characteristics of the dataset. Methods of calculating the readability score are detailed in the Evaluation Metrics Section 2.4. The readability of both source and target texts are at undergraduate level (e.g. 13th grade = 1st year, 15th grade = 3rd year).

[3]This can be loosely defined as literature that is disseminated outside the usual publishing channels.

[4]https://www.cochranelibrary.com/cdsr/reviews

	Train		Validation		Test	
	Source	**Target**	**Source**	**Target**	**Source**	**Target**
# abstracts	5,195	5,195	500	500	1,000	1,000
Average length (words)	714	374	713	368	727	378
Vocabulary size	57,685	34,175	16,574	10,596	23,938	15,407
Flesch-Kincaid	14.68	13.25	14.93	13.57	14.70	13.23
Gunning	14.57	13.54	14.69	13.79	14.57	13.49
Coleman-Liau	15.40	14.37	15.57	14.51	15.39	14.43

Table 2.2: Dataset statistics across the different splits.

Notably, the average lengths of abstracts from the source sets are larger than those from the target sets, and the target sets have lower readability scores (i.e. more readable) on average. Since the length of the abstracts and the readability scores from different subset splits are similar, the dataset splits can be considered to be comparable.

In order to understand how experts translate scientific biomedical abstracts into plain language versions that target the general population, we identified five typical transformation phenomena on the basis of our observations when constructing the current dataset.[5] These transformation categories with examples are presented in Table 2.1.

The most common transformation to make the paragraph more straightforward is to remove unnecessary details. Although some details such as experimental settings, control experiments or quantitative results are informative for professionals and may indicate the quality of a scientific review, such information may confuse laypeople and obscure the key findings from the review. The critical message for laypeople is the general association between an intervention and a health condition, rather than the precise details of the scientific evidence used to support this conclusion.

Explaining relevant background information, including the prevalence, mortality, risk

[5]More comprehensive guidelines for writing Cochrane plain language summaries can be found in McIlwain et al. (2014).

factors and outcome of a condition, enables readers to establish whether or not the topic under discussion meets their information needs. Jargon (or even some standard medical terms) presents another challenge that prevents laypeople from referring to peer-reviewed papers for answers to their health-related questions. Providing definitions for technical terms (such as "randomized control trials" in Table 2.1) can make professionally-authored text more understandable to a lay audience. Restating the sentence, especially title or headings, in an interrogative sentence makes the scientific content more engaging, and highlights the clinical question under consideration in the review. Sometimes, it is difficult for laypeople to understand the importance of the study question.

Finally, many cases in our dataset require sentence structure simplification. Rephrasing lengthy, convoluted sentences as shorter ones can divide complex information into smaller, easier-to-process units. We also identified other less frequent transformations, such as avoiding passive voice, using "must" to indicate requirements, and minimizing abbreviations. As it would be intractable to exhaustively identify and categorize all of the transformation types in our dataset, we provide only the most commonly encountered ones in this paper.

2.4 Evaluation Metrics

The various phenomena in our task present a challenge for comprehensive and fair evaluation. Therefore, we adopt several automatic evaluation metrics, as well as human evaluation to assess different aspects of model performance.

2.4.1 Automated evaluation

Summarization evaluation

We first use ROUGE (Lin, 2004) to evaluate the summarization performance. ROUGE-n measures overlap of n-grams between the model-generated summary and the human-generated reference summary, and ROUGE-L measures the longest matching sequence of words using the longest common subsequence. In this task, we report the F1 scores of ROUGE-1, ROUGE-2, and ROUGE-L as the summarization performance measures.

ROUGE scores were computed using `pyrouge`[6].

Readability evaluation

Other than how much information is retained in the summary, we are also interested in assessing the ease with which a reader can understand a passage, defined as readability. We use three standard metrics to evaluate readability: Flesch-Kincaid grade level (Kincaid et al., 1975), Gunning fog index (Gunning et al., 1952), and Coleman-Liau index (Coleman and Liau, 1975). These scores are computed using `textstat`[7], and their formulae are as follows:

- **Flesch-Kincaid grade level**:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59,$$

- **Gunning fog index**:

$$0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right)+100\left(\frac{\text{complex words}}{\text{words}}\right)\right],$$

  where complex words are those words with three or more syllables.

- **Coleman-Liau index**:

$$0.0588L-0.296S-15.8,$$

  where L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

These readability evaluation metrics all estimate the years of education generally required to understand the text. Lower scores indicate that the text is easier to read. More specifically, scores of 13-16 correspond to college-level reading ability in the United States education system. Table 2.2 shows the readability scores of the source and target sets of our dataset. Although all the scores indicate a college level of education is needed to read even the target

[6]https://pypi.org/project/pyrouge/

[7]https://pypi.org/project/textstat/

summary, we do see a stable difference between the scores of the source and target. This indicates that these scores are useful for reflecting the different level of readability for text in our dataset.

2.4.2 Human evaluation

While we have adopted the most commonly used metrics for assessing summarization and simplification performance, many aspects of the generated text, such as fluency, grammaticality, and coherence, are not captured by them. Of particular importance, factual correctness of the generated text is crucial in the medical domain. To consider these desirable properties, we developed a method for further assessment of summary quality by human evaluators.

Specifically, we presented an evaluator with two biomedical abstracts followed by four questions. Evaluators were recruited if they: (1) were able to read and write in English; and (2) had at least 12 years education (as the education level required for the training dataset we preprocessed is college level). Evaluators were excluded if they (1) had participated in medical training or shadowing in a hospital; or (2) had completed advanced (graduate level) biology courses. These criteria were selected to ensure that our evaluators were representative of the college-educated lay public. Participants were recruited by convenience sampling, and the study was considered exempt upon institutional IRB review. The estimated time for completion of the human evaluation was 30 minutes for each participant. No compensation was provided for participating in this study. We recruited 8 human raters. The average age of these evaluators was 23.5 years old. Four of them were female, and all had more than 12 years of formal education. Each of the abstract/summary pairs was assigned to two independent evaluators.

Two versions of each biomedical abstract were presented: SOURCE refers to the original professional language version, and SUMMARY refers to the version to be evaluated. This was either the professionally written target, or the version generated by our best-performing automated summarization model, BART pre-trained on both CNN/DM and PubMed. Evaluators were blinded to the authorship of the summary (BART vs. human expert).

Two biomedical abstracts (A and B) were randomly selected from the test set. Evaluators were required to read through the two pairs of abstracts and compare the SUMMARY to SOURCE considering the following aspects on a 1-5 Likert scale (1 - very poor; 5 - very good):

- **Grammaticality** Do you think the SUMMARY is grammatically correct?
- **Meaning preservation** Does the SUMMARY provide all the useful information you think is important from the source?
- **Understandability** Is the SUMMARY easier to understand than the source?
- **Correctness of key information** How do you judge the overall quality of the SUMMARY in terms of its correctness of the key information compared to the source?

2.5 Experiments

Model	ROUGE-1	ROUGE-2	ROUGE-L	Flesch-Kincaid	Gunning	Coleman-Liau
Oracle extractive	**53.56**$_{\pm 0.58}$	**25.54** $_{\pm 0.78}$	**49.56**$_{\pm 0.65}$	14.85	13.45	16.13
BERT extractive	26.60$_{\pm 0.51}$	11.11$_{\pm 0.41}$	24.59$_{\pm 0.47}$	**13.44**	**13.26**	**14.40**
Pointer generator	38.33$_{\pm 0.61}$	14.11$_{\pm 0.46}$	35.81$_{\pm 0.60}$	16.36	15.86	15.90
BART	52.53$_{\pm 0.51}$	21.83$_{\pm 0.52}$	49.75$_{\pm 0.52}$	13.59	14.16	14.45
BART+CNN/DM	52.46$_{\pm 0.48}$	21.84$_{\pm 0.50}$	49.70$_{\pm 0.50}$	13.73	14.33	14.60
BART+PubMed	52.66$_{\pm 0.48}$	21.73$_{\pm 0.48}$	49.97$_{\pm 0.51}$	**13.30**	**13.80**	**14.28**
BART+CNN/DM+PubMed	**53.02**$_{\pm 0.48}$	**22.06**$_{\pm 0.49}$	**50.24**$_{\pm 0.49}$	13.60	14.11	14.41

Table 2.3: Test set performance evaluated by ROUGE and readability score. BART model pretrained on CNN/DM and PubMed is the best-performing model based on ROUGE, while BART model pretrained on PubMed is the best one based on readability score (Best model performance is in bold). $x_{\pm}$ indicates 95% interval: $[x-, x+]$

2.5.1 Methods

Summarization methods can be broadly categorized into extractive and abstractive approaches. The extractive approach creates summaries by selecting the most important sentences in a document, while the abstractive approach usually employs sequence-to-sequence models to generate summaries that may contain new phrases not included in the source document. We experimented with several state-of-the-art extractive and abstractive methods to check the feasibility and difficulty of the plain language summarization task.

Extractive methods

We applied two extractive methods – *Oracle extractive* and *BERT extractive* (Liu and Lapata, 2019) – to the CDSR dataset. *Oracle extractive* can be viewed as an upper bound for extractive models. It creates an oracle summary by selecting the set of sentences in the document that generates the highest ROUGE-2 score with respect to the gold standard summary. Since oracle extractive summarization takes the gold standard summary into consideration, it can't be applied summarization tasks in practice. *BERT extractive* is the state-of-the-art extractive method for text summarization. *BERT* (Devlin et al., 2018) is a bidirectional unsupervised language representation derived by pre-training a Transformer architecture on a unlabeled text corpus for reconstruction. Several inter-sentence Transformer layers are then stacked on top of BERT outputs, to capture document-level features for extracting summaries. A sigmoid classifier is added as the output layer for extractive summarization. The oracle summary in the *Oracle Extractive* method are used as supervision for training the *BERT Extractive* model.

Abstractive methods

Pointer-generator was a commonly used abstractive model before pretraining dominated the field. It enhances the standard sequence-to-sequence model with a pointer network that allows both copying words from the source and generating words from a fixed vocabulary. *BART* is a state-of-the-art summarization model based on a large transformer sequence-to-sequence architecture. It is pre-trained on large corpora by corrupting text with an arbitrary

noising function, and learning a model to reconstruct the original text. As a sequence-to-sequence model, BART can be directly fine tuned for abstractive summarization task.

Intermediate pre-training

To compensate for the limited training data, we added intermediate pre-training steps for the BART model before finetuning. We first experimented with adding labeled data for summarization task in other domains. We adopted the CNN/DM dataset (Nallapati et al., 2016), which contains about 287K document-summary pairs, and BART is among the best-performing systems for this task. Secondly, we tried to pre-train BART with an unlabeled biomedical corpus to expose the model to medical domain-specific language. We used the PMC articles dataset [8] which contains 300K PubMed abstracts. Following the BART paper, we corrupted these documents using several transformations, including text substitution and sentence shuffling. BART was then trained on the corrupted abstracts to reconstruct the original PubMed abstracts. Lastly, we combined these two strategies to train BART on CNN/DM and PubMed sequentially before finetuning it on our dataset.

Training details

All experiments were run using a single NVIDIA Tesla V-100 GPU. All models were developed using `PyTorch`. We used `neural-summ-cnndm-pytorch`[9] to implement the pointer-generator model. The batch size was set to 4. Other hyper-parameters were set to default values. We built the BERT extractive model using code released by the authors.[10] The learning rate was set to 2×10^{-3} and the batch size 140. Other hyper-parameters were set to default values. We used the Fairseq [11] BART implementation. All BART models were trained using the Adam optimizer. The learning rate was set to 3×10^{-5}, and learning decay was applied. The minimum length of the generated summaries was set to 100, and the maximum length was set to 700.

[8] https://www.kaggle.com/cvltmao/pmc-articles

[9] https://github.com/lipiji/neural-summ-cnndm-pytorch/

[10] https://github.com/nlpyang/presumm

[11] https://github.com/pytorch/fairseq

	Abstract A		Abstract B	
Perspectives	**Target**	**Generated**	**Target**	**Generated**
Grammaticality	4.25	4.50	3.50	4.00
Meaning Preservation	3.75	4.75	3.50	4.50
Understandability	3.75	3.50	2.75	2.50
Correctness of Key Information	3.50	4.50	4.00	4.00

Table 2.4: Human evaluation scores of the expert-generated summaries (*Target*) and the model-generated summaries (*Generated*) for two abstracts from the test set. Generated abstracts from BART+CNN/DM+PubMed model have better scores in grammaticality, meaning preservation, and correctness of key information.

2.5.2 Results

Automated evaluation

ROUGE and readability results on the CDSR test set are shown in Table 2.3. We compare the seven methods described above: Oracle extractive, BERT extractive, pointer-generator, BART, BART pre-trained on CNN/DM, BART pre-trained on Pubmed abstracts, and BART pre-trained on both CNN/DM and PubMed abstracts.

The oracle extractive method, as an upper bound for the extractive approach, produces the best ROUGE-1 and ROUGE-2 scores. However, it obtains approximately the same level of readability as the source text in our test set (Table 2.2) , which indicates that selecting the reference sentences will only result in a summary that is as difficult to read as the original abstract. In contrast, the BERT-based extractive model achieves better readability scores while performing worst in terms of ROUGE scores. This demonstrates that, in practice, training the model to extract the correct content from the original abstract might be difficult, even though the model learns to extract shorter and easier sentences.

Among the 5 abstractive models, the pointer-generator model performs significantly worse in both ROUGE and readability, emphasizing the importance of pre-training for our

task. BART-based models achieve surprisingly good performance in terms of both summarization and readability, suggesting contemporary NLP models have the potential to perform the task, and to help the general public access professional medical information. Additionally, BART pre-trained on CNN/DM and PubMed abstracts achieves the best performance in ROUGE, and BART pre-trained only on PubMed abstracts obtained the lowest readability. This demonstrates the usefulness of either adding task-relevant labeled data or domain-specific unlabeled data. However, our strategies for adding such data are quite straightforward, and we lacked resources to do hyperparameter search for the relatively expensive pre-training procedure. Therefore, we only see marginal improvement compared with the BART model. We will aggregate more relevant data, and develop better pre-training strategies to improve the performance in future work.

Human evaluation

Table 2.4 shows the human evaluation results. Intriguingly, human evaluators rated the model-generated summaries with comparable or even higher scores for all the four aspects, and for both abstract A and B. The average Kendall's coefficient (Sen, 1968) for the two biomedical abstracts among all evaluators' inter-rater aggrement is 0.62. Kendall's coefficient ranges from -1 to 1, indicating low to high association. Considering the subjectivity of the rating task, this number indicated high human agreement for the tasks. While larger scale study is required, this work provides preliminary evidence that automatically-generated plain language summaries are readable and interpretable to non-expert human readers.

2.5.3 Qualitative analysis

We present the output of our best two models in the last two columns of Table 2.1. This provides evidence that the best-performing models can address some transformations, and generate grammatical and meaningful outputs. Specifically, out of the five listed phenomena, we observed that model-generated summaries could achieve three transformation types to some extent, including removing unnecessary details, jargon explanation and sentence

structure simplification. Some capabilities the model demonstrated are encouraging for future research. For example, it learned to explain the term RCT from similar examples in the training data.

On the downside, the models are still struggle with some difficult transformations, such as relevant background explanation. This ability is harder to learn, and our dataset might not contain the required background knowledge. Therefore, external knowledge might be also useful. Furthermore, we also see risks in using the current abstractive models to generate reliable information for the public. For example, in the example of sentence structure simplification, *BART+PubMed* changed the meaning of the original sentence: the source sentence claims an association between the pattern of blood flow with poor prognosis, while the generated sentence focuses on the Doppler ultrasonography. *BART+CNN/DM+PubMed* performs better in this case.

2.6 Discussion

Automated lay language summarization of biomedical scientific reviews requires both summarization and the acquisition of domain knowledge. Previously, available datasets were constructed at sentence level. However, sentence-level simplification or transformation does not require the complex strategies used by experts when rendering biomedical literature understandable to a lay audience. Therefore, we consider the document-level dataset as an important outcome of our work, which can be useful for future research on this topic. Abstractive models are more practical than extractive ones, since extractive summaries are written in the same professional language as their source documents. The best performing model is BART pre-trained on both CNN/DM and PubMed abstracts, which preserves key information (based on ROUGE) while dropping the reading requirements a year or two (based on readability scores).

Human evaluation is necessary for our task. There is a considerable gap between the automatic evaluation merics and human judgement. Despite being widely used to evaluate summarization systems, ROUGE is not practical for our task because it can neither capture the required transformation phenomena nor assess difficulty in understanding. Similarly,

lower readability scores do not imply understandability. Readability scores consider only the surface forms, without considering the complexity introduced by medical abbreviations and domain-specific concepts. Human evaluation is the most robust method to evaluate the performance. However, aside from the small number of participants, the survey questions need a formal validity. Further studies are required to find that BART-derived summaries were more appealing to human raters on several fronts hold when more abstracts and human raters are involved.

2.7 Conclusion

We propose a novel plain language summarization task at the document level and construct a dataset to support training and evaluation. The dataset is of high quality, and the task is challenging due to typical transformation phenomena in this domain. We tried both extractive and abstractive summarization models, and obtained best performance with a BART model pre-trained further on CNN/DM and PubMed, as evaluated by automated metrics. Human evaluation suggests the automatically generated summaries may be at least as acceptable as their professionally authored counterparts.

Chapter 3

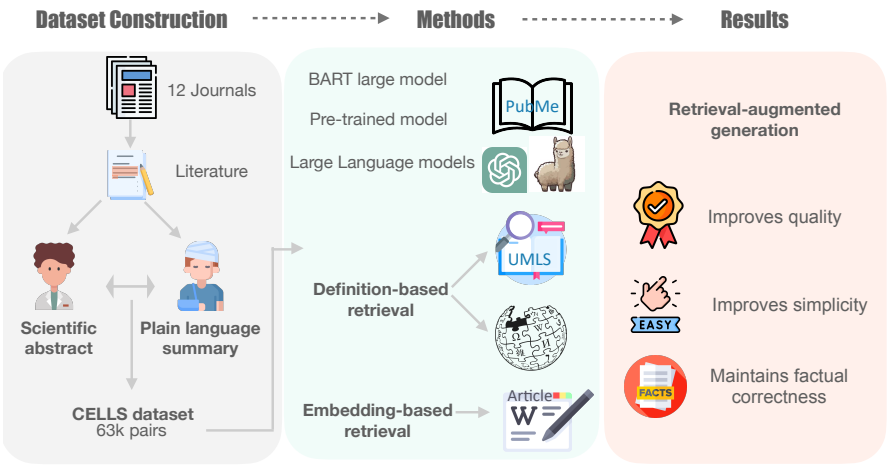
CHALLENGE 2. GENERATING BACKGROUND EXPLANATIONS

To tackle the specific challenge of generating background explanations, in the work described in this chapter, I first define what a background explanation is and emphasize its importance in this chapter. I then propose Retrieval-Augmented Lay Language (RALL) generation as a solution that intuitively addresses the need for external knowledge beyond what is provided in expert-authored source documents. This approach involves augmenting state-of-the-art text generation models with information retrieval, using either term definitions from the UMLS and Wikipedia or embeddings of explanations from Wikipedia documents. I also assess the performance of both open-source (Llama 2) and closed-source (GPT-4) Large Language Models in generating background explanations, with and without retrieval augmentation. Furthermore, to help mitigate the limited data challenge, I introduce CELLS, which is the largest (63k pairs) and most diverse (12 journals) parallel corpus for plain language summaries to date. Our code and data are publicly available at: https://github.com/LinguisticAnomalies/pls_retrieval.

A version of this chapter was previously published by Journal of Biomedical Informatics (JBI). ©JBI.

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

3.1 Introduction



The COVID-19 pandemic underscored the difficulties the general public faces when attempting to use scientific information to guide their health-related decisions (Soroya et al., 2021; Bin Naeem and Kamel Boulos, 2021). Though widely *available* in scientific papers and preprints, the information required to guide health-related decision making is often not *accessible*: medical jargon (Korsch et al., 1968), scientific writing styles (Kurtzman and Greene, 2016), and insufficient scientific background (Crossley et al., 2014) make this information opaque to non-experts. Consequently, there is a pressing need to deliver scientific knowledge in lay language, which has motivated research on automated generation of lay language summaries.

Prior work has framed lay language generation as a summarization or simplification task (Guo et al., 2021; Devaraj et al., 2021). However, adapting biomedical text for the lay public includes the distinct task of *background explanation*: adding external content in the form of definitions, history, or examples to enhance comprehensibility. Cognitive studies of text comprehension suggest that providing missing background information effectively improves reader comprehension, especially when readers lack the prerequisite domain knowledge to fill this in themselves (McNamara et al., 1996).

Text simplification, which modifies content to improve readability while retaining its key points, has been widely studied (Jonnalagadda et al., 2009a; Qenam et al., 2017). However, generating background information is especially challenging, because the source docu-

ment may not include the required background knowledge. Furthermore, background explanation capabilities have yet to be formally evaluated, and little is known about how best to enhance them. Retrieval augmentation methods, which use information retrieval to identify additional content to inform text generation, present an intuitive fit for the need to acquire external knowledge. In the current work, we explore methods for Retrieval-Augmented Lay Language (RALL) generation, augmenting state-of-the-art text generation models with information retrieval of either term definitions from the UMLS (Bodenreider, 2004) and Wikipedia, or embeddings of explanations from Wikipedia documents (Lewis et al., 2020c). Our findings indicate that RALL models improve summary quality and simplicity while maintaining factual correctness, suggesting that general knowledge from Wikipedia in particular is a good source for background explanation. With Large Language Models (LLMs) becoming increasingly accessible, we also tested the ability of two LLMs for background explanation: we prompted both open-source Llama 2 (Touvron et al., 2023) and closed-source GPT-4 (OpenAI, 2023a) with and without external knowledge from Wikipedia. Results indicate that these LLMs improve simplicity but do not preserve the summary quality.

Abstractive summarization methods require source/summary pairs, with the summaries written in plain language. The limited size and topical breadth of publicly-available paired corpora constrain the scope of applicability of models trained for this task and limit the generalizability of published evaluations. Therefore, a further contribution of this work is the Corpus for Enhancement of Lay Language Synthesis (CELLS): 62,886 pairs of scientific abstracts with corresponding lay language summaries (Table 3.1). Summaries are written by abstract authors or other domain experts, assuring the quality of our dataset. CELLS is larger and more diverse than prior datasets (Guo et al., 2021; Devaraj et al., 2021), aggregating papers from 12 journals (Table 3.2) spanning various biomedical domains. From CELLS, we derived a set of specialized paired corpora: 233,916 algorithm-aligned sentence pairs for *simplification* and 47,157 scientific/lay-language pairs emphasizing novel content that is absent from scientific abstracts for *background explanation* to support our research on background augmentation.

Abstract: Clinical reports of Zika Virus (ZIKV) RNA detection in breast milk have been described, but evidence conflicts as to whether this RNA represents infectious virus...
Summary: *Only 4 years have passed since the Zika virus outbreak in Brazil, and much remains to be understood about the transmission and health consequences of Zika infection.* To date, some case reports have detected Zika virus RNA in the breast milk of infected mothers, but the presence of a virus' RNA does not mean that intact virus is present...

Table 3.1: Example abstract/summary pair from CELLS. The lay language summary is written by the abstract's authors. There are two challenges in lay language summary generation: generating background explanations (*italicized*) and simplifying the original abstract (underlined).

3.2 Related Work

3.2.1 Lay language summary generation

Text summarization and simplification are two important aspects of lay language generation. Text summarization is a widely-studied research topic (Cohan et al., 2018b; Cachola et al., 2020a; Devaraj et al., 2022a). It has been a focus of research attention in the biomedical domain (Mishra et al., 2014; Bui et al., 2016; Givchi et al., 2022), with applications including summarization of radiology reports (Cai et al., 2021; Zhang et al., 2018, 2020), biomedical literature (Wang et al., 2021b; Plaza, 2014; Cai et al., 2022) and medical dialogue (Chintagunta et al., 2021; Joshi et al., 2020). Several of the biomedical text simplification datasets and methods have also been reported in the literature (Jonnalagadda et al., 2009b; Li et al., 2020b; Cao et al., 2020; Lu et al., 2023a). However, these were designed for sentence-level text simplification, rather than translation of paragraphs and longer documents into interpretable lay language. Compared to other paragraph-level lay language generation efforts (Guo et al., 2021; Devaraj et al., 2021), the current work is the first to focus on background explanation generation.

3.2.2 Lay language summarization datasets

Previous endeavors towards developing datasets for automated conversion of scientific text into lay language have been limited in scale and scope. The CL-SciSumm 2020 shared task series (Chandrasekaran et al., 2020) provided a training dataset encompassing 572 articles and corresponding author-constructed lay summaries, collated from a diverse array of scientific journals published by Elsevier. Guo et al. (Guo et al., 2021) and Devaraj et al. (Devaraj et al., 2021) introduce datasets of ∼5k scientific abstract and lay language summary pairs drawn from systematic reviews in the Cochrane Library. Goldsack et al. (Goldsack et al., 2022) present ∼30k biomedical literature abstract pairs from PLOS and eLife. Luo et al. (Luo et al., 2022c) developed a dataset from ∼28k biomedical abstract pairs from PLOS. Attal et al. (Attal et al., 2023) describe the PLABA dataset, encompassing 750 pairs of abstracts, each set featuring a sentence-aligned adaptation generated by human authors. The dataset developed for our study differs from these prior efforts in that: 1) CELLS is a large (∼63K) abstract-level dataset which includes different article types besides systematic reviews; and 2) we address the need for background explanation in lay language generation, deriving a specialized subset emphasizing content that is absent from the abstracts.

3.2.3 Lay language summary generation methodologies

Present text summarization methodologies predominantly fall into two categories: extractive and abstractive (Das and Martins, 2007). Extractive summarization involves ranking and selecting critical elements of the original text and combining them to form a condensed version (Erkan and Radev, 2004; Cheng and Lapata, 2016b). In contrast, abstractive summarization introduces novel words and phrases absent from the original text (Gupta and Gupta, 2019). The necessity to provide pertinent background, explain terminology, and apply straightforward sentence structures makes lay language summarization intrinsically an abstractive task (Guo et al., 2021). The emergence of Transformer-based approaches such as BART, T5, and PEGASUS has significantly advanced this field (Zhang et al., 2021; Yadav et al., 2022). BART, especially when pre-trained on domain-specific data, has demonstrated strong performance in the simplification of biomedical review articles (Guo et al., 2021;

Goldsack et al., 2022) and the summarization of randomized controlled trials (Wallace et al., 2021). We employ BART as the benchmark model in the current work, including a variant with additional PubMed-specific pre-training. In addition, newer work has indicated that auto-regressive LLMs can outperform other Transformer models in lay language generation tasks (Goldsack et al., 2023a). Therefore, we also evaluated the performance of two such LLMs: Llama 2 (Touvron et al., 2023) and GPT-4 (OpenAI, 2023a), on our dataset.

3.2.4 Retrieval-augmented text generation

Background explanation helps laypeople understand biomedical concepts (Srikanth and Li, 2020a). Furthermore, experiments in cognitive psychology have shown that providing explanatory content improves the recall of readers with limited domain knowledge (Britton and Gülgöz, 1991; McNamara et al., 1996). Information retrieval methods present an intuitive approach to identify content to inform background explanations, with established utility for clinical question answering (Simpson et al., 2014; Roberts et al., 2015; Luo et al., 2022b), biomedical text summarization (Alambo et al., 2022; Mishra et al., 2014; Plaza, 2014) and clinical outcome prediction (Naik et al., 2021). There are two main categories of information retrieval methods that have been used to augment the generation of natural language text. *Definition-based* retrieval methods identify terms that exist in predefined lexicons, and use their definitions to inform text generation (Alambo et al., 2022; Moradi and Ghadiri, 2018). *Embedding-based* retrieval methods retrieve documents with similar low-dimensional representations, instead of depending upon lexical overlap between terms (Deerwester et al., 1990; Cao and Xiong, 2018; Guu et al., 2020; Karpukhin et al., 2020; Lewis et al., 2020c). Retrieval augmentation has been shown to improve the performance of question answering systems (Lewis et al., 2020c), and reduce the frequency of so-called "hallucinations" (statements without grounding in training data) in text generated by language models (Shuster et al., 2021). However, these approaches have not been explored for lay language generation, despite their intuitive fit to the subtask of background explanation in particular. In the current work, we explore both definition- and embedding-based retrieval approaches and evaluate the utility of external information from the UMLS and Wikipedia

		Length	
Journal	**Num.**	**Src**	**Tgt**
PNAS	25,647	227	124
PLOS Genetics	8,030	256	192
PLOS Pathogens	7,345	260	193
PLOS Neglected Tropical Diseases	7,185	315	198
PLOS Computational Biology	7,072	253	188
Cochrane	5,377	624	334
PLOS Biology	2,149	243	212
Health Technology Assessment	557	645	318
Health Services and Delivery Response	510	623	316
Public Health Research	93	624	331
Programme Grants for Applied Research	78	722	311
Efficacy and Mechanism Evaluation	70	659	341

Table 3.2: Journals included in CELLS. The average length (token level) of lay language summaries (Tgt) is shorter than that of scientific abstracts (Src).

for this important subtask.

3.3 Materials and methods

3.3.1 The CELLS Dataset

We present CELLS, the largest dataset of parallel scientific abstracts and expert-authored lay language summaries (LLSs) developed to date (Section 3.3.1), offering unique opportunities to study the performance of lay language generation models. To facilitate research on key LLS generation subtasks, we have also derived subsets for simplification and background explanation (Section 3.3.1).

Data compilation

To develop CELLS, we manually reviewed biomedical journals and identified 19 with a LLS section (see Appendix Table 3.5). We collected scientific abstracts (*source*) and their aligned LLSs (*target*) from these journals. We excluded abstracts where LLSs are not associated with a full-length paper (i.e., LLS in a separate section for the journal's website or social media feed) that required extensive human inspection. After further excluding non-biomedical topics, we obtained 75,205 pairs of abstracts and LLSs. To ensure data quality, we identified outliers using source-target lexical similarity and length. As a result, we excluded pairs from eLife, Annals of the Rheumatic Diseases, and Reproductive Health. This left a set of 62,886 source-target pairs from 12 journals.

Dataset applications

Using CELLS, we developed three evaluation tasks:

Lay language generation For this task, we used the full-length scientific abstract and LLS pairs in CELLS for abstract-level lay language generation. As mentioned in Section 3.1, this task requires paragraph-level simplification, summarization, and background explanation to produce understandable summaries for laypeople. The following tasks focus on two of these challenges: abstract simplification and background explanation generation.

Simplification Paragraph-level simplification fits the lay language generation task, but simplification is difficult to isolate because of the frequent insertion of background explanations. To focus on sentence-level simplification as a subtask, we developed a Greedy Paired Sentence Search (GPSS) algorithm (Algorithm 1) to align sentences from the abstracts and LLSs. The underlying idea is to identify matched source and target sentences based on lexical overlap and sentence sequence. An example is provided in Figure 3.1. After applying GPSS, each source and target sentence was labeled as "matched" or "unmatched", resulting in a large set of 233,916 matching abstract- and LLS-derived sentence pairs for simplification.

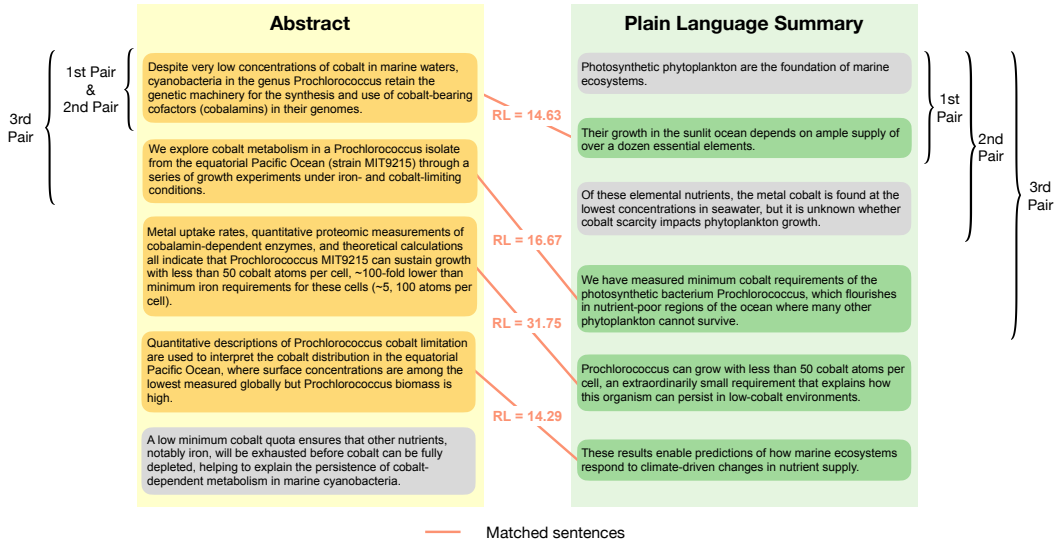


Figure 3.1: An example application of the GPSS algorithm. RL indicates the F1 score from ROUGE-L between the sentences in the abstract and plain language summary. For the background explanation subset, we combined unaligned target sentences (grey blocks) with proximal aligned sentences (green blocks). The example presented illustrates the generation of three paired examples ("pair") for the background explanation subset. All three pairs include the initial explanatory content that precedes the first matched sentence (RL = 14.63), as well as the sentence in the lay language summary that matches it. The second pair also includes the explanatory content after this matched sentences, and the third pair adds the following matched sentence also (i.e. the second sentence in the source abstract, and the lay language summary sentence that aligns with it). These combinations allow for the possibility that added content may relate to the preceding, or the subsequent sentence.

Algorithm 1 Greedy paired sentences search (GPSS) algorithm

Input: SRC – the list of sentences in the source abstract, TGT – the list of sentences in the target abstract **Output:** P – the set including the indices of the paired source and target sentences

function GPSS(src_start, src_end, tgt_start, tgt_end, score)
    if src_start $>$ src_end **or** tgt_start $>$ tgt_end **then**
        return $\emptyset$
    src_max,tgt_max $\leftarrow$ argmax$_{i,j}$(score[i,j], src_start $\leq$ i $<$ src_end, tgt_start $\leq$ j $<$ tgt_end)
    pairs $\leftarrow$ {(src_max, tgt_max)}
    pairs $\leftarrow$ pairs $\cup$ GPSS(src_start, src_max-1, tgt_start, tgt_max-1, score)
    pairs $\leftarrow$ pairs $\cup$ GPSS(src_max+1, src_end, tgt_max+1, tgt_end, score)
    return pairs
$N_{src} \leftarrow$ the number of sentences in SRC
$N_{tgt} \leftarrow$ the number of sentences in TGT
for i $\leftarrow$ 0 to N_{src}-1 **do**
    for j $\leftarrow$ 0 to N_{tgt}-1 **do**
        S[i,j] $\leftarrow$ ROUGE-L(TGT[j], SRC[i])
P $\leftarrow$ GPSS(0, N_{src}, 0, N_{tgt}, S)

Background explanation As mentioned in Section 3.1, adding explanations is a common strategy to enhance comprehension in 'Background' section. To support this research, we derived a large-scale paragraph-level dataset that emphasizes the insertion of novel content, i.e., background explanations. Focusing on explanation requires a reliable approach to extract sentences containing additional content in the LLS 'Background' section. Human annotation is reliable but costly, and exhaustive annotation of the 62,886 pairs in CELLS is infeasible. Therefore, we obtained paired source/target sub-paragraphs with the aforementioned GPSS algorithm. After applying GPSS, we considered the unaligned ("unmatched") sentences as putative *explanations*. We targeted the Background section, but section headers are unavailable for most abstracts. We therefore conducted human annotation to examine the utility of different empirically-defined boundaries, and the presence of external information. Fifty randomly selected abstracts from CELLS were annotated by two annotators: one medical student and one graduate student without medical training but with good familiarity

with the dataset. Cohen's Kappa among the annotators is 0.74, indicating substantial agreement (Artstein and Poesio, 2008). Informed by the results of the annotation process, we selected the 2nd pair (refer to Figure 3.1) to demarcate the 'background' section, given its superior integration of background and external information. Further details can be found in Section 3.4.5. Overall, we extracted 47,157 source/target pairs for background explanation [1].

3.3.2 Human Validation of Dataset

To ensure the robustness of the dataset used for background explanation and simplification, two expert annotators (same as above) assessed 250 paragraph pairs from the background explanation subset and 500 pairs from the simplification subset. The annotators were tasked with evaluating the pairs from the background explanation by: 1) confirming their presence within the background section; 2) identifying any external data not originally in the source; 3) classifying any external information as either definition, motivation, or example (detailed definitions of the categories can be found in Section 3.4.4); and 4) discerning whether the target and source information are aligned, where alignment is defined by the presence of common entities, i.e., at least one shared "triple" (A triple consists of three components: A subject, a predicate, and an object). Our annotators determined that 92.8% of pairs were situated in the background section and 62.8% included external information. Among the identified external information, 76.4% were motivations, 38.2% were definitions, and 10.8% were examples (these labels are not mutually exclusive). In addition, 87.2% of background explanation pairs were found to be in alignment. For the simplification subset, we focused on the alignment of paired sentences and found a 68.6% alignment. The challenge of sentence-level alignment in scientific summaries remains an active area of investigation (Krishna et al., 2023a), emphasizing the ongoing need for the advancement of alignment algorithms. Taking into account the inherent complexities of sentence alignment and external information detection, we consider the observed alignment and external rates

[1]The term "background explanation" refers to specific sentences found within a Background section, but not every sentence in this section qualifies as a "background explanation." Instead, a background explanation is specifically defined as unmatched sentences serving explanatory purposes.

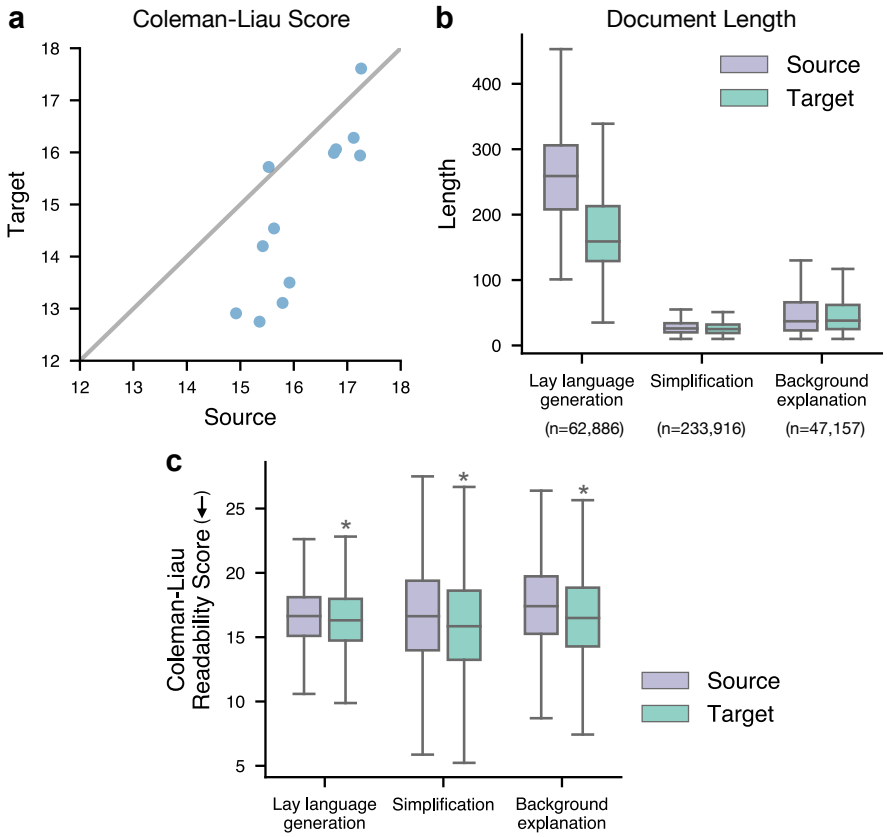


Figure 3.2: Dataset analysis. a, source and target Coleman-Liau readability scores for the 12 journals included in CELLS. Each dot represents one journal. Lower score indicates text is easier to read. b,c, Average length and Coleman-Liau readability score for source and target text for three tasks (i.e., lay language generation, simplification and background explanation). On average, target text is shorter and easier to read for all three tasks. "*" indicates that the score of the target significantly lower than that of the source with p-value < 0.05 (paired t-test).

to be within acceptable bounds for the purpose of the current work. However, these results also reinforce the need for further research on sentence alignment.

Dataset analysis

Dataset statistics are shown in Table 3.2. CELLS covers various topics including genetics, pathogens, neglected tropical diseases, computational biology, health services, and biomedical research. This diversity of topics and journals provides opportunities to study model generalizability. Background explanations notwithstanding, the average length of the source (professional language) is longer than the target (plain language summary) for each journal. The readability scores for each journal are shown in Figure 3.2a. The Coleman-Liau readability score indicates the estimated years of education required to understand a piece of text. Most of the average readability scores for the target are lower than those for the source, indicating that the target LLSs are generally easier to understand.

Figures 3.2b and 3.2c show lexical features of CELLS components for three tasks. On average, LLSs are shorter than corresponding scientific abstracts. Although the readability of both source and target texts is at the college level (Karačić et al., 2019), the difference in readability between them is statistically significant (paired t-test), indicating LLSs are easier to understand than source text.

We randomly split the dataset into 45,280; 11,295; and 6,311 abstract/LLS pairs as the train, validation, and test sets respectively.

3.3.3 Methods

We investigated the performance of language models with intermediate pre-training (i.e. further pre-training on in-domain text) and retrieval-augmented lay language generation (RALL).

In-domain pre-training for simplification

Abstractive models are more applicable than extractive ones for our tasks since extractive summaries are written in the same professional language as their source documents. There-

fore, we applied a state-of-the-art abstractive summarization model – *BART* (Lewis et al., 2020a) – to our tasks. BART uses hidden state representations of text sequences that are encoded bi-directionally (as is the case with BERT (Devlin et al., 2018)) to inform a decoder model that predicts the next word in the sequence (as is the case with GPT-series and related models e.g. (Brown et al., 2020b)). During semi-supervised pre-training of the model, input sequences are perturbed with a range of transformations (for example, some tokens may be masked), and the model attempts to reconstruct the original sequence. As such, BART has both the ability to encode hidden state representations that take an entire sequence into account, and a convenient mechanism to generate text in response to an input sequence. Once the model has been pre-trained on unlabeled text, it can be fine-tuned for particular tasks, such as summarization, by training it to generate target sequences in response to source sequences. We adopted a $\text{BART}_{\text{Large}}$ model that has been fine-tuned for general-domain text summarization on the widely-used CNN/DM summarization dataset (Cable News Network / Daily Mail (Nallapati et al., 2016)) as our baseline. To be concise, we use *Vanilla* to denote the BART-Large-CNN model in the following text. Due to the complexity of our task and the relatively small size of our data, we employed intermediate pre-training - further semi-supervised pre-training on additional unlabeled in-domain text - to attempt to improve the performance of BART. Previous work shows that adaptive pre-training with domain-relevant unlabeled data can improve model performance (Gururangan et al., 2020c). Therefore, we further pre-trained the BART model on a corpus from the biomedical domain. We first perturbed 300K PubMed abstracts[2] by text substitution and sentence shuffling, and trained the BART model to reconstruct the original text. The pre-trained model was then further fine-tuned for the tasks of summarization, sentence simplification and background explanation, using our datasets.

Definition-based explanation retrieval

As the source documents in our dataset may omit required background knowledge, models should be able to retrieve relevant background knowledge from external sources. We

[2]https://www.kaggle.com/cvltmao/pmc-articles

evaluated two approaches to retrieving this knowledge.

The definition-based retrieval model uses a straightforward method to add explanations of terms to the text, by identifying definitions of terms that exist in a predefined lexicon. In our experiments, we used datasets derived from the UMLS (Bodenreider, 2004) and Wikipedia to augment the context of the source document. The UMLS includes medical term (entity e) and definition (d) pairs $\mathcal{D}_u = \{(e_i, d_i)\}$. For each source document s, we used Scispacy (Neumann et al., 2019) to identify the expression of normalized UMLS terms $e_{u_1}, e_{u_2}, ..., e_{u_m}$ in s. Then we added corresponding UMLS term definitions $d_{u_1}, d_{u_2}, ..., d_{u_m}$ to s to obtain $\hat{s}$. The Wikipedia dataset includes keyword (e) and definition (d) pairs $\mathcal{D}_w = \{(e_i, d_i)\}$. For each source document s, we applied KeyBERT[3], which uses BERT embeddings and cosine similarity to find the sub-phrases in a document most similar to the document itself, to identify three keywords $(e_{w_1}, e_{w_2}, e_{w_3}) = \text{KeyBERT}(s)$. We obtained the definitions of those keywords, $d_{w_1} d_{w_2}, d_{w_3}$, from the Wikipedia dataset and added them to the end of the source document, s, to obtain $\hat{s}$. Lastly, we fine-tuned the BART model using $\hat{s}$.

Embedding-based explanation retrieval

For the embedding-based retrieval method, we adopted a state-of-the-art dense retrieval model to augment the source with related documents from an external set. Specifically, we applied the retrieval-augmented generation (RAG) model (Lewis et al., 2020c) using another Wikipedia-derived dataset Z_w of 21M 100-word documents z. The retrieval component $p_\phi(z|s) \propto \exp(\mathbf{d}(\mathrm{z})^{\mathrm{T}}\, \mathbf{q}(\mathrm{s}))$ is based on the Dense Passage Retriever (DPR) (Karpukhin et al., 2020), where $\mathbf{d}(z) = \text{BERT}_d(z)$ and $\mathbf{q}(s) = \text{BERT}_q(s)$ are the representations of the documents (in our case, the source document to be summarized and the Wikipedia documents that provide candidates for retrieval) produced by two $\text{BERT}_{\text{Base}}$ encoders. The DPR model retrieves the top k documents z with the highest prior probability $p_\phi(z|s)$ using the Maximum Inner Product Search method (Johnson et al., 2019). After applying DPR, we concatenated the source s and the retrieved content z as the input. We used the RAG-

[3]https://github.com/MaartenGr/KeyBERT

Sequence model whose generator produces the output sequence probabilities for each concatenated document:

$$p(t|s) \approx \sum_{z \in top\text{-}k(p_\phi(\cdot|s))} p_\phi(z|s) p_\theta(t|s,z).$$

$p_\theta(t|s,z)$ is the generator, and we used BART for this purpose. As can be seen from the formula, both the content of the retrieved documents (z) and their probabilities of retrieval ($p_\phi(z|s)$) inform the generated text. During training, we fixed BERT$_d(\cdot)$, and only fine-tuned BERT$_q(\cdot)$ and the BART generator.

3.3.4 LLMs

To evaluate the performance of LLMs in generating background explanations or plain language summaries, we utilized Llama 2 (Touvron et al., 2023) and GPT-4 (OpenAI, 2023a). We explored two prompts: 1) `"Summarize in plain language: input"`, and 2) `"Summarize in plain language, providing necessary explanations: input"`. To further assess the impact of the retrieval-augmented approach on LLMs, we established two settings for input: the source alone and the source combined with Wikipedia definitions as identified using KeyBERT.

3.3.5 Experiments

Experimental setup

All experiments except LLMs were run using a single NVIDIA Tesla V-100 GPU. Models were developed using PyTorch (Paszke et al., 2019). We used the Fairseq[4] BART implementation, and the HuggingFace Transformers Library (Wolf et al., 2019) to implement the RAG model. For RAG, we retrieved the top 5 documents for each input. The maximum length of generated texts was set to 700 for paragraph-level lay language generation and 150 for background explanation and sentence-level simplification. Other hyper-parameters were set to their default values.

[4]https://github.com/pytorch/fairseq

We used the `Llama-2-70B-chat`[5] model for Llama 2 [6] and GPT-4 [7] for the GPT model. The generation process was configured with a maximum length of 150 tokens. All other parameters were set to their default values.

Evaluation

Automated evaluation We first evaluated generation quality using ROUGE-L (Lin, 2004), BERTScore (Zhang* et al., 2020), BLEU (Papineni et al., 2002), and METEOR (Banerjee and Lavie, 2005)[8] to compare generated text to professionally-authored plain language target text. ROUGE-L depends on n-gram overlap, while BERTScore uses the similarity between embeddings and as such may be less sensitive to differences in word choice between human-authored and automatically-generated LLS. BLEU computes n-gram precision of generated text against target texts, including a brevity penalty. METEOR employs a relaxed matching criterion based on the F-measure, and addresses the exact match restrictions and recall consideration of BLEU.[9] The Coleman-Liau readability score (Coleman and Liau, 1975) assesses the ease with which a reader can understand a passage, and word familiarity (Leroy and Kauchak, 2014) measures the inverse document frequency of unigrams in text using frequency counts from Wikipedia. *Lower* Coleman-Liau score and word familiarity indicate that text is *easier* to read.[10]

To directly evaluate how representative of an LLS the generated text is, we trained a RoBERTa (Liu et al., 2019) model to classify the source of sentences from the original abstracts and LLSs. Specifically, we used the paired source-target sentences from the GPSS algorithm with a sentence length between 10 and 150 words. The input to the RoBERTa

[5]Model: https://ai.meta.com/llama/

[6]Implementation: The model was quantized to 4 bits using OPTQ as implemented in https://github.com/PanQiWei/AutoGPTQ, and hosted on a local server using https://github.com/turboderp/exllama for inference

[7]Implementation: https://openai.com

[8]Implementation: (Fabbri et al., 2020) BERTScore hash code: `bert-base-uncased_L8_no-idf_version = 0.3.12(hug_trans=4.27.3)`

[9]Please see BLEU and METEOR scores in Appendix.

[10]The familiarity measure is derived from *inverse* document frequency which is higher for rare terms, so *lower* familiarity scores indicate the use of *more familiar* words.

model is a sentence and the label is 0 for a source sentence (from a scientific abstract) and 1 for a target sentence (from a LLS). The RoBERTa model achieved an AUROC of 0.83 and an F1 score of 0.74 on the held-out test set, demonstrating that there are detectable and generalizable differences in language use by the intended audience. As the model is trained to output a higher prediction for a target sentence (i.e., a LLS sentence), we used the prediction of the model to evaluate how "plain" the input text is. We refer to the predicted probability of this model as the "Plainness Score". A *higher* Plainness Score indicates that the text is more representative of an LLS.

Human evaluation We set up our human evaluation similarly to (Guo et al., 2021), providing pairs of source and summary text to the human evaluators, where the summary is either expert-written or generated by one of our two best-performing BART models and two GPT-4 models (evaluators were not informed which summaries were human-authored). We asked human evaluators to rate the summary for grammatical correctness, meaning preservation, understandability, factual correctness, and the relevance of external information, each on a 1-4 point Likert scale (1-very poor, 4-very good). Questions can be found in Section 3.4.6. The study was considered exempt upon institutional IRB review. Twelve evaluators were recruited using an institutional NLP interest group channel. Each of them annotated four examples from the test set for background explanation. Three evaluators reviewed each example. All the evaluators have at least an undergraduate degree, lack specialized biomedical training, and half speak English at home. The Krippendorff's alpha coefficient (Krippendorff, 1970) was 0.40 for the four background explanation texts among all evaluators. Krippendorff's alpha coefficient measures multiple inter-rater agreements in ordinal data, and values range from 0 to 1. Considering the subjectivity of the task and the multiple choices per question, we considered this level of agreement to be acceptable.

3.4 Results

We experimented with five models using BART: the base (*Vanilla*) model, *Vanilla* further pre-trained on PubMed abstracts (*PubMed pre-trained*), *Vanilla* with UMLS (*UMLS definition-based retrieval*) and Wikipedia (*Wiki definition-based retrieval*) definition-based

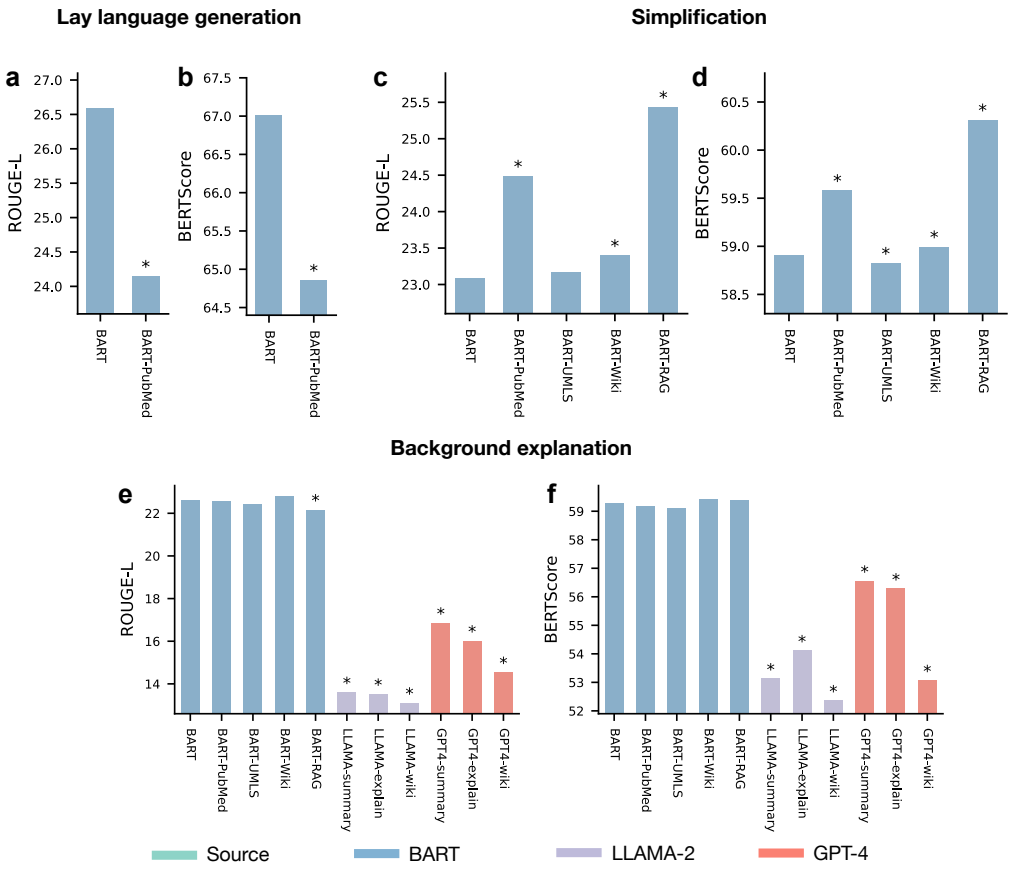


Figure 3.3: Models' performance in text generation. We used the F1 score of ROUGE-L and BERTScore to evaluate the generation quality of models on lay language generation, simplification, and background explanation tasks. P-values obtained through the t-test are employed to evaluate the performance of various models compared to the Vanilla model (BART). * indicates statistical significance with Bonferroni-Holm correction for multiple hypothesis testing Holm (1979).

retrieval, and *Vanilla* with embedding-based retrieval using Wikipedia (*Wiki embedding-based retrieval*). Additionally, we experimented with three prompts using LLMs for background explanation: ”Summarize in plain language: input” (*summary*), ”Summarize in plain language, providing necessary explanations: input (*explain*), and ”Summarize in plain language: input with Wiki definition-based retrieval (*wiki*).

3.4.1 RALL improves generation performance

We first evaluated the text generation performance of our models (Figure 3.3), using ROUGE-L and BERTScore to compare generated LLS text to the corresponding human-authored lay language text (the target) for a given abstract or sentence (the source). Due to the input length limitation of the BERT (512 tokens) and BART (1024 tokens) models, we did not perform retrieval-augmented generation for the abstract-level lay language generation task (Figure 3.3, 1st panel). However, for this task, pre-training on unlabeled data was not helpful.

We next compared text generation performance on the sentence-level simplification task (Figure 3.3, 2nd panel). Results indicate that the PubMed pre-trained model achieved better performance than the Vanilla model, suggesting that pre-training on domain-specific unlabeled data is helpful for sentence-level simplification, which aligns with the results from our prior work Guo et al. (2021). It can also be observed that the models with information retrieval from Wikipedia (Wiki definition- and embedding-based retrieval) achieve higher ROUGE-L scores than the Vanilla model, suggesting that retrieving this external information may also be helpful for text simplification tasks. One reason for this may be that Wikipedia articles target a broader audience than the intended audience of specialized biomedical literature, and are therefore written to be more accessible.

For background explanation (Figure 3.3, 3rd panel), the PubMed pre-trained model only shows marginal improvements, suggesting that pre-training on domain-specific unlabeled data is helpful but insufficient for this challenging task, perhaps because the authors of biomedical literature assume expert knowledge on the part of their readers and therefore seldom include the explanatory content that a non-expert reader might require.

Furthermore, the BART models with retrieval from the Wikipedia dataset (Wiki definition- and embedding-based retrieval) achieved higher BERTScore, establishing the benefits of information retrieval techniques for background explanation with BART. ROUGE-L results show a smaller advantage for Wiki definition-based RALL generation, and unlike with BERTScore this advantage is not apparent for embedding-based RALL methods. With auto-regressive LLMs ROUGE and BERTscore results are remarkably consistent: GPT-4 outperforms Llama 2 with all three prompts; however, both models are notably outperformed by BART-based architectures. Our results suggest that prompting exclusively for summarization yields superior outcomes compared to combining summarization with explanation. The weakest performance is observed when using the Wiki definition-based retrieval source combined with summarization prompting. This disparity could be attributed to our reliance on zero-shot LLMs, whereas BART benefits from fine-tuning. This suggests there may be avenues for improvement, such as exploring few-shot learning approaches within LLMs for background explanation, or using low-rank adaptation techniques to further improve auto-regressive LLM performance. To offer a comprehensive view of performance, BLEU and METEOR scores from the test set are also presented in Figure 3.6. The observed BLEU patterns are consistent with those for ROUGE and BERTScore, while METEOR highlights advantages for 'explain' LLM queries and BART-RAG.

In acknowledgement of the limitations of our GPSS algorithm, where the background explanation doesn't always pair paragraphs with external content and the simplification subset doesn't always produce aligned pairs, we provide results from the 'gold' annotated dataset. These are available in Figure 3.7 and Figure 3.8. This 'gold' background explanation subset features pairs in alignment with external content, while the 'gold' simplification subset showcases aligned pairs. The results derived from these 'gold' validated subsets are consistent with those from the test set, indicating that the observed patterns are not attributable to errors in algorithmic alignment.

Overall, our results suggest that both biomedical domain pre-training and information retrieval are helpful for background explanation and sentence-level simplification. Furthermore, the information retrieval models based on BART using Wikipedia produced text that was most similar to human-authored lay language. This indicates that general domain infor-

mation written for a broader audience (e.g., Wikipedia) is a good resource for background explanation generation using BART.

3.4.2 RALL improves text interpretability

We next evaluated the interpretability of the generated text using the data from the background explanation subset (Figure 3.4). Existing text interpretability metrics consistently return better scores for the models' outputs than for the source text. Of note, the Coleman-Liau readability scores of the models' outputs are even lower than those of the target text (Figure 3.4a.) This indicates that our datasets help the BART model to generate more straightforward and readily interpretable text. We also found that the retrieval-augmented BART models performed well in this interpretability evaluation, suggesting that the UMLS and Wikipedia datasets may be easier to understand than professional-language abstracts. Overall, the embedding-based RALL model, which used Wikipedia as a source, had the best readability, familiarity, and plainness scores. These results further support the utility of retrieval augmentation for lay language generation, suggesting it can benefit the style of generated text, as well as its content. For LLMs, the model outputs consistently score well across all metrics. Outputs generated with the Wiki-based definition retrieval source show improved results in the readability score and relative word familiarity compared to those without it. However, this advantage doesn't extend to the plainness score.

3.4.3 Human evaluation

Figure 3.5 shows the human rating scores across four pairs of target and generated texts. Evaluators generally rated generated background explanations higher than those from the expert-generated LLS. It is interesting to note that the Wiki definition-based retrieval BART model was judged to have the least relevant external information but the best understandability, according to raters. Exploring the tradeoff between the amount of external information and understandability, and jointly optimizing them presents a challenging direction for future work. These results confirm the effectiveness of our dataset for improving automated models' ability to generate LLS with relevant external information added. For GPT-4, when

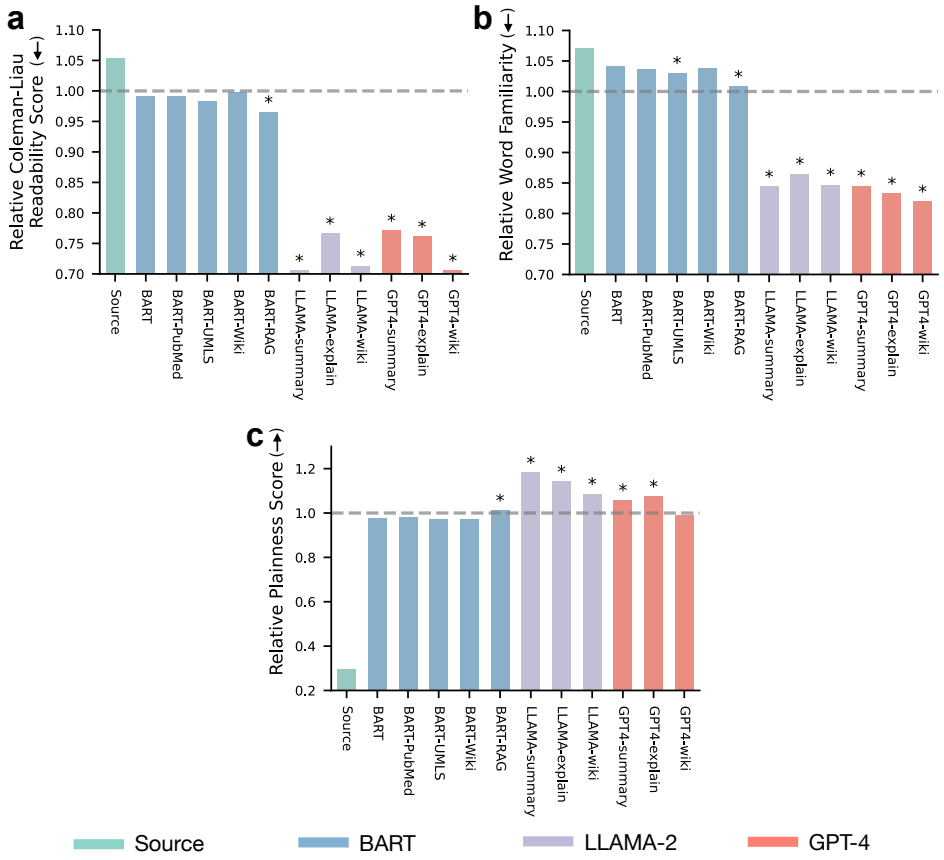


Figure 3.4: Readibility, familiarity and plainness of the background explanation subset, relative to professionally-authored lay language text. (a) Relative Coleman-Liau readability score, (b) word familiarity, and (c) plainness score of the source and models' generated text. The relative score is calculated by dividing by the score of the target text. A lower readability score and word familiarity indicate that the text is easier to read (values below the dashed line are lower than those from professionally-authored plain language). A higher Plainness Score indicates that the text is more representative of an LLS. P-values obtained through the t-test are employed to evaluate the performance of various models compared to the Vanilla model (BART). * indicates statistical significance with Bonferroni-Holm correction for multiple hypothesis testing Holm (1979).

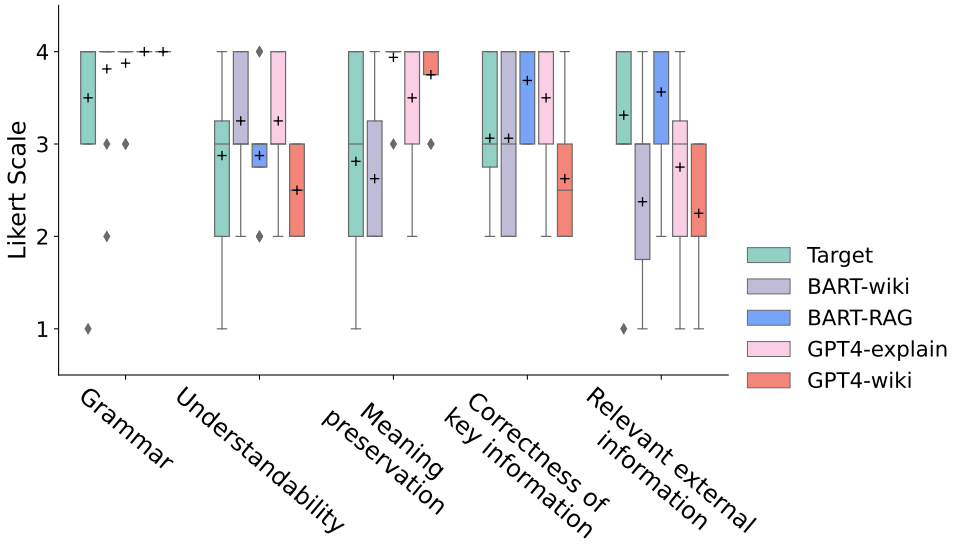


Figure 3.5: Human evaluation results for four generated texts from the background explanation task. Each text was assigned to four raters. For the Likert scale, 1 is very poor, and 4 is very good. "+" indicates the mean.

prompted with summarization and explanation, yields the highest scores in understandability, meaning preservation, and information correctness. However, it falls short in incorporating relevant external information. This suggests that GPT-4's explanatory outputs should be meticulously vetted to prevent potential misalignment with the topic. In contrast, the embedding-based Wiki approach (BART-RAG) excels in meaning preservation, maintaining the accuracy of key information, and integrating relevant external information. However, its outputs can be challenging to comprehend. This raises the potential of synergizing the embedding-based method with LLMs to achieve the ideal lay language summary.

3.4.4 Selected examples

To define the scope of background explanation, we identified three types of explanation in the dataset, as shown in Table 3.3. We did not aim to enumerate all possible categories. Rather, our goal was to provide some initial insights into explanation phenomena.

The most common explanation type we observed is a *definition*, including "common"

medical words, technical terms, and abbreviations, to avoid misunderstanding. *Motivation*, including prevalence, risk factors, history, etc., sustains readers' interest and establishes whether the topic under discussion meets their information needs. Providing a *concrete example* allows readers to link an otherwise obscure concept with a more familiar one. For example, connecting the increasing temperature in the ocean with coral reefs makes it easy for the reader to understand the importance of the study.

We present background explanations for the two best BART models and two GPT-4 models in Table 3.3. This provides evidence that the retrieval-augmented model can generate both term definitions and motivations for the main topic. However, the generated external content may not be aligned with the target (e.g. Marburg virus does not cause Ebola virus disease), highlighting the importance of improving the relevance and correctness of generated abstractive summaries as an area for future research. In addition, the models appear unable to produce illustrative examples. This ability goes beyond retrieving evidence, and appears difficult to learn.

3.4.5 Background explanation annotation

We provided annotators with the original abstract/LLS pair and the content (both matched and unmatched) before the 1st, 2nd, and 3rd matched sentence pairs. To make sure we have matching content from the corresponding LLS, the 1st matched sentence was included to capture situations in which explanation is provided before the first matching sentences (e.g. an introductory sentence defining terms). Examples of matching strategies can be found in Figure 3.1. We asked annotators to identify whether the filtered content 1) is in the background section (to confirm our heuristics indeed identify these sections); 2) truly contains external content (to confirm that unaligned GPSS sentences represent content that is absent from the source document); and 3) is paired (to confirm aligned GPSS sentences represent the same content). The results are shown in Table 3.4.

3.4.6 Human evaluation questions

The questions we included in the human evaluation questionnaire are as follows:

	Definition	Motivation	Example
Source	Marburg virus (MARV) disease is lethal, with fatality rates up to 90%…	Clinical reports of Zika Virus (ZIKV) RNA detection in breast milk have been described, but evidence conflicts as to whether this RNA represents infectious virus…	…Very little is known about how environmental changes such as increasing temperatu affect disease dynamics in the ocean, especially at large spatial scales…
Target	Marburg virus (MARV) is a member of the Filoviridae family that causes severe hemorrhagic fevers with high fatality rates in humans…	Only 4 years have passed since the Zika virus outbreak in Brazil, and much remains to be understood about the transmission and health consequences of Zika infection. To date, some case reports have detected Zika virus RNA in the breast milk…	…This hypothesis is supporte by local observations--for example, that some coral diseases become more prevalent in the summertime- but it has never been tested a large spatial scales or over relatively long periods….
BART-Wiki	Marburg virus (MARV) is a highly lethal pathogen that is transmitted by the bite of infected mosquitoes…	Zika Virus (ZIKV) is a mosquito-borne virus that has been circulating in tropical and subtropical regions of the world since 2015, and it has been linked to an outbreak of microcephaly and Guillain-Barre syndrome (GBS) in Brazil in 2015. ZIKV RNA detection in breast milk…	…Disease dynamics in the ocean are influenced by environmental changes such increasing temperature, but little is known about how thes changes affect disease dynamics over large spatial scales…
BART-RAG	Marburg virus (MARV) is the causative agent of Ebola virus disease, which is fatal in up to 90% of cases…	Zika virus (ZIKV) is a mosquito-borne flavivirus that has caused a worldwide health crisis since its discovery in the Americas in 2015. Clinical reports of Zika virus RNA detection in human breast milk…	…Disease dynamics in the ocean are affected by environmental changes such temperature, but very little is known about how these changes affect disease dynamics at large scales…
GPT4-Wiki	The Marburg virus is a deadly disease with a high fatality rate.	The Zika virus, which gained worldwide attention due to an outbreak in the Americas from 2015 to 2016, affects a protein-coding gene in a species called Drosophila melanogaster…	We don't know much about how changes in the environment, like rising temperatures, impact the spread and development of diseases in the ocean, particularly on a large scale.
GPT4-explain	The virus called Marburg Virus Disease is extremely dangerous and can kill up to 90% of the people it infects.	The Zika virus has gained worldwide attention over the last five years mainly because of its comeback in the Americas from 2015 to 2016…	We don't know a lot about hov changes in the environment, like rising temperatures, impa the spread and behavior of diseases in the ocean. This is especially true when we're looking at larger areas.

Table 3.3: Typical types of background explanation from scientific abstracts to lay language summaries, and the corresponding generated text using two best-performing models. Our retrieval-augmented models can generate both term definitions and related epidemiological data for the main topic, but fail to link examples to related concepts.

- Is the grammar of the plain text correct?
- Is the plain text easier to understand than the original text?
- Does the plain text provide all the important information from the original text?
- Is the information in the plain text correct compared to the original text?
- Does the plain text provide relevant additional information compared to the original text?

	Annotator 1	Annotator 2
1st Pair		
Background	48	48
External	20	18
Pair	43	40
2nd Pair		
Background	47	47
External	36	28
Pair	47	46
3rd Pair		
Background	21	23
External	35	28
Pair	50	47

Table 3.4: Background extraction annotation results. The columns show the number of the 50 annotated examples that each annotator labeled as containing content from the background section, including content external to the source, and being aligned with content from the source.

3.5 Discussion

We have two key observations from the model development and evaluation. Regarding BART models, RALL variants outperform the vanilla model, though these improvements are larger with the sentence simplification subset than when background explanation is emphasized. Background explanation generation is challenging and requires considering both the knowledge sources from which information is drawn, and the understandability of generated text. Examples suggest that the models can add term definitions and relevant epidemiological data but fail to provide illustrative examples for related concepts. These abilities may be beyond current models' capabilities and fall outside the scope of the resources used in our study for information retrieval. Therefore, generating high-quality explanations may

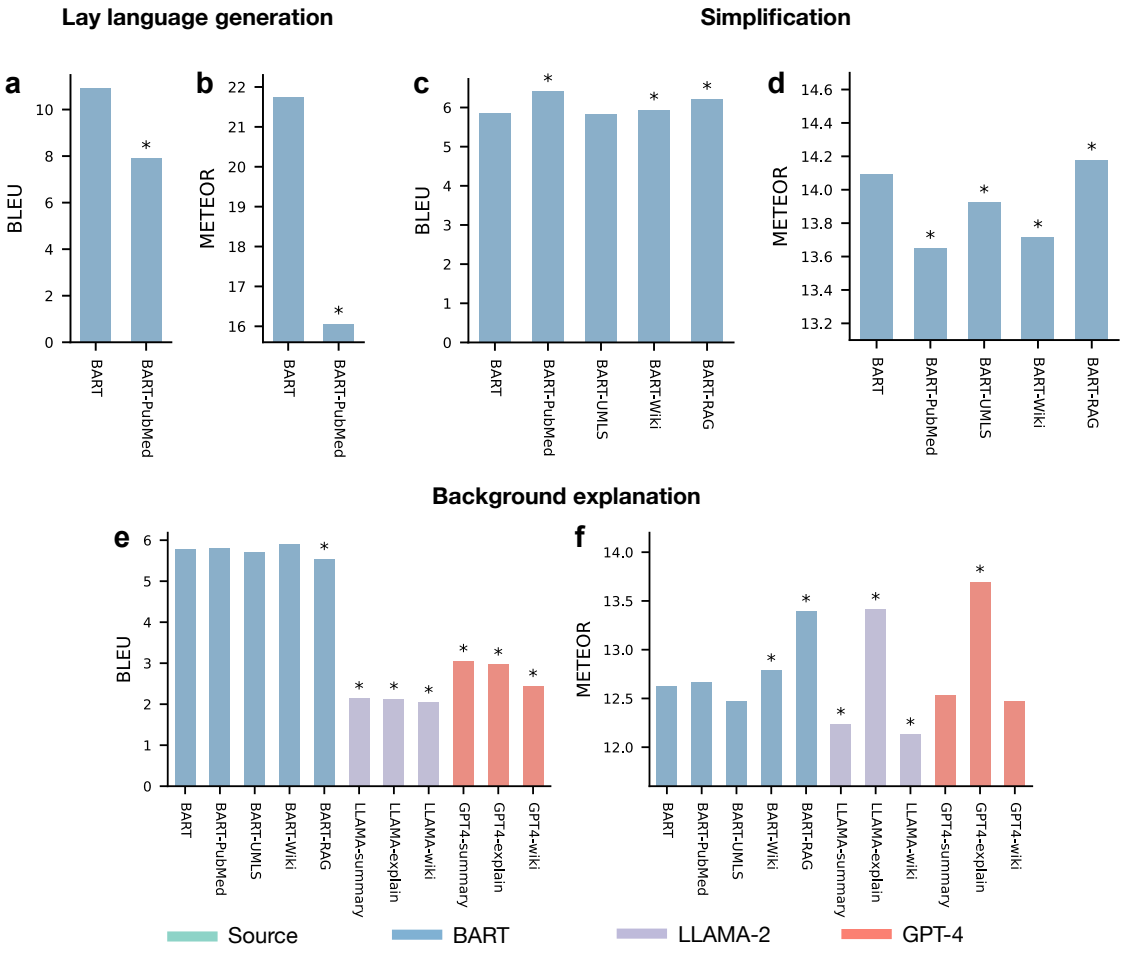


Figure 3.6: Models' performance in text generation. We used the F1 score of BLEU and METEOR to evaluate the generation quality of models on lay language generation, simplification, and background explanation tasks. P-values obtained through the t-test are employed to evaluate the performance of various models compared to the Vanilla model (BART). A p-value less than 0.05 is indicated by (*).

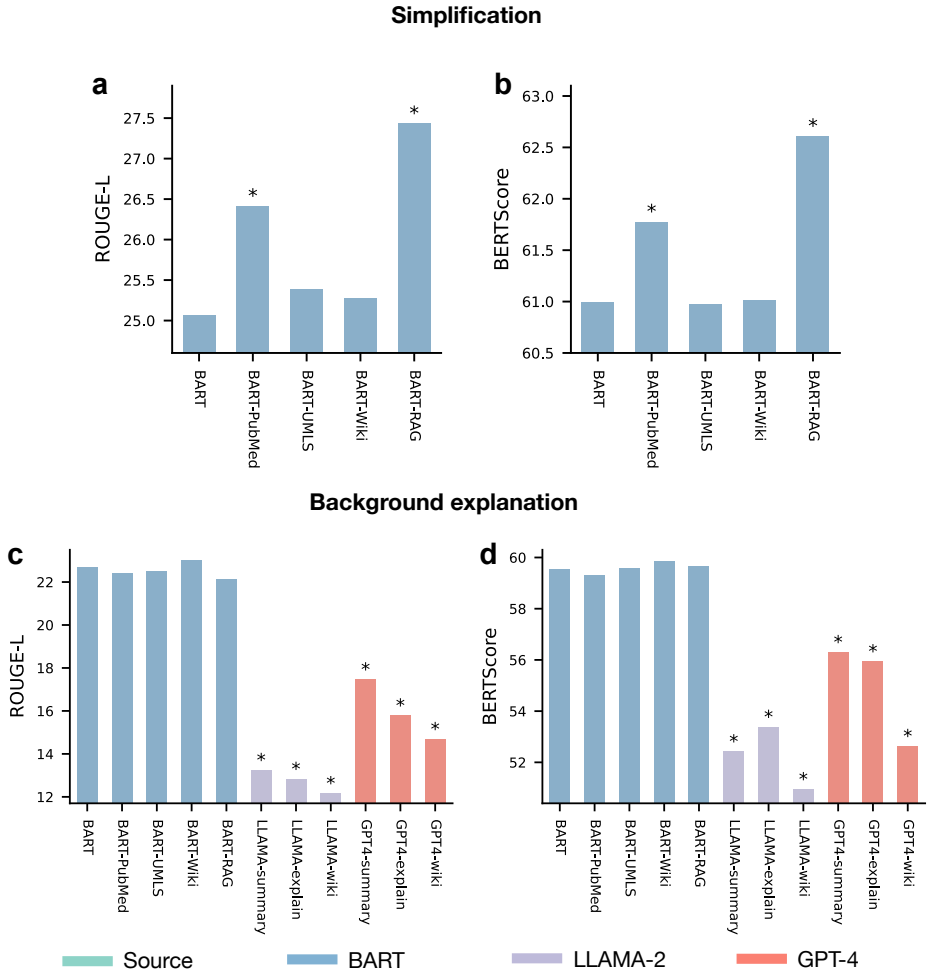


Figure 3.7: Models' performance in text generation on the validated dataset. We used the F1 score of ROUGE-L and BERTScore to evaluate the generation quality of models on lay language generation, simplification, and background explanation tasks. P-values obtained through the t-test are employed to evaluate the performance of various models compared to the Vanilla model (BART). A p-value less than 0.05 is indicated by (*).

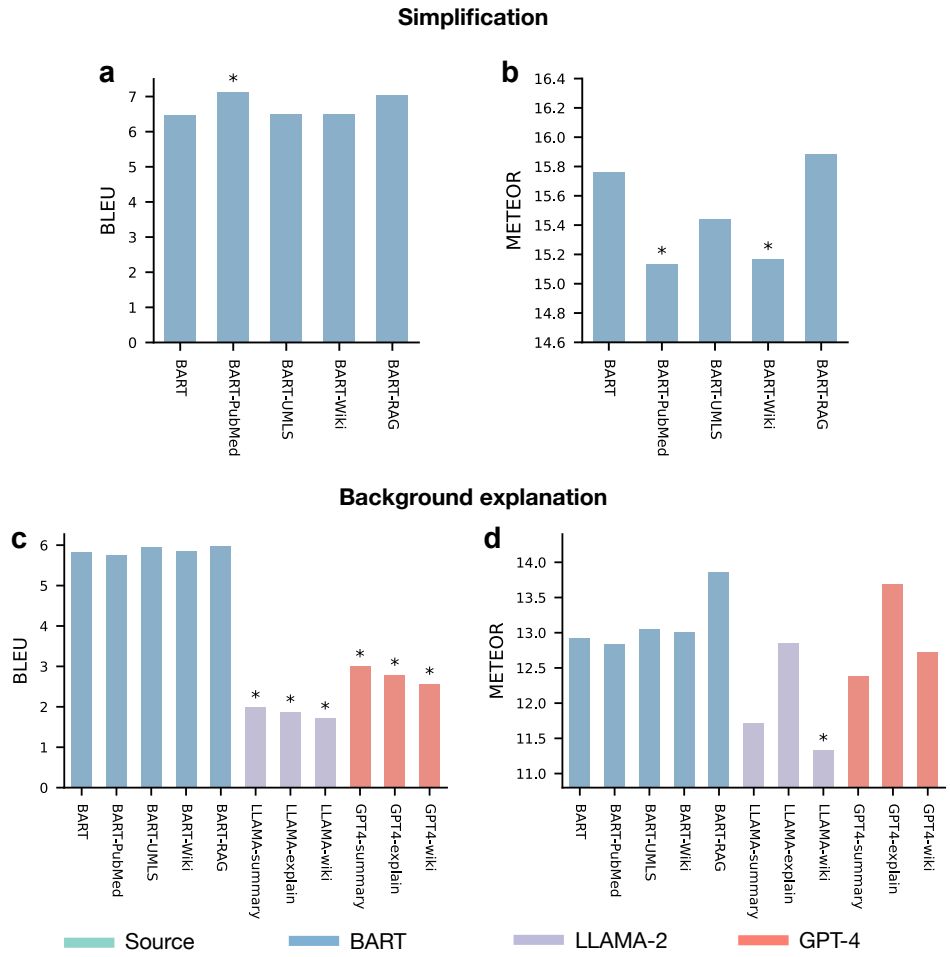


Figure 3.8: Models' performance in text generation on the validated dataset. We used the F1 score of BLEU and METEOR to evaluate the generation quality of models on lay language generation, simplification, and background explanation tasks. P-values obtained through the t-test are employed to evaluate the performance of various models compared to the Vanilla model (BART). A p-value less than 0.05 is indicated by (*).

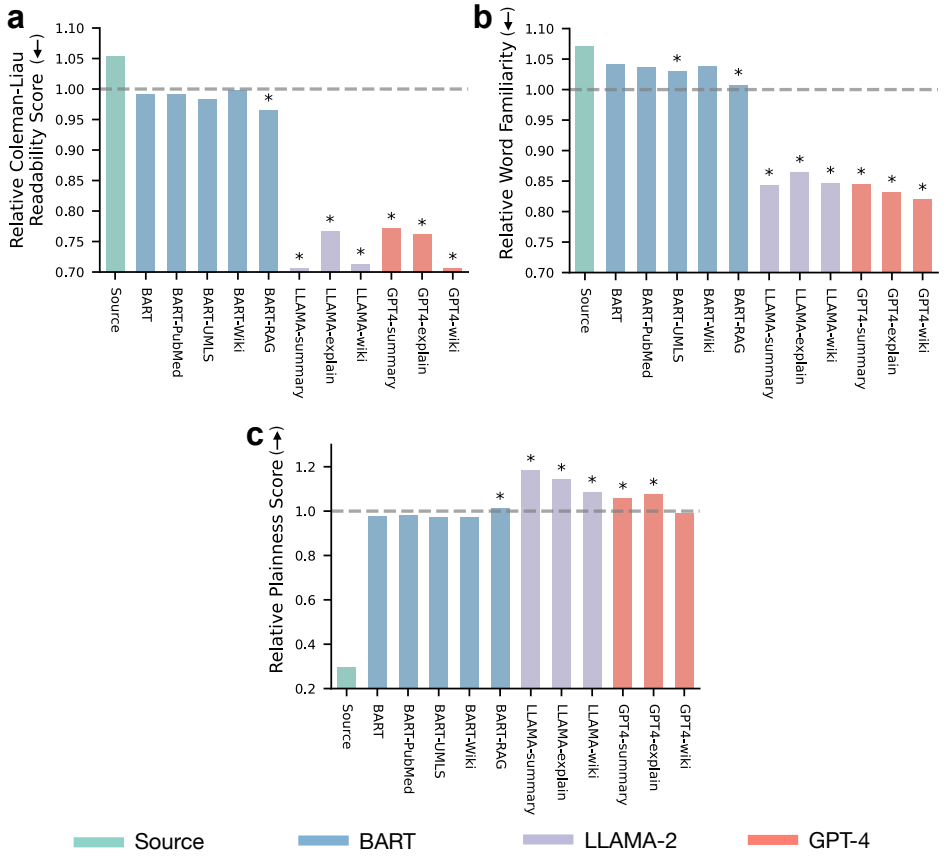


Figure 3.9: Readibility, familiarity, and plainness of the validated background explanation subset. (a) Relative Coleman-Liau readability score, (b) word familiarity, and (c) plainness score of the source and models' generated text. The relative score is calculated by dividing by the score of the target text. A lower readability score and word familiarity indicate that the text is easier to read (values below the dashed line are lower than those from professionally-authored plain language). A higher Plainness Score indicates that the text is more representative of an LLS. P-values obtained through the t-test are employed to evaluate the performance of various models compared to the Vanilla model (BART). A p-value less than 0.05 is indicated by (*).

Publisher	Type	Name	Where are they displayed	Written by	Start from
Annals of the Rheumatic Diseases	Journal	Patient summary	Dedicated section of website	Authors	2013
Autism	Journal	Lay abstract	Dedicated section of website	Authors	2011
Autism Research	Journal	Lay abstract	Dedicated section of website	Authors	2008
British Journal of Dermatology	Journal	QAs	Issues searching page	Authors/editorial team	2013
Cochrane	Journal	Plain language summary	Within article	Authors/editorial team	1997
eLife	Blog	eLife digest	Section on blog	Editorial team	2012
European Urology	Journal	Patient summary	Within article	Authors	2014
NIHR Efficaacy and Mechanism Evaluation	Journal	Plain English summary	Within article	Authors	2014
NIHR Health Services and Delivery Response	Journal	Plain English summary	Within article	Authors	2014
NIHR Health Technology Assessment	Journal	Plain English summary	Within article	Authors	2014
NIHR Programme Grants for Applied Research	Journal	Plain English summary	Within article	Authors	2015
PLOS Biology	Journal	Author summary	Within article	Authors	2007
PLOS Computational Biology	Journal	Author summary	Within article	Authors	2005
PLOS Genetics	Journal	Author summary	Within article	Authors	2005
PLOS Medicine	Journal	Author summary	Within article	Authors	2006
PLOS Neglected Tropical Diseases	Journal	Author summary	Within article	Authors	2007
PLOS Pathogens	Journal	Author summary	Within article	Authors	2005
Proceedings of the National Academy of Sciences	Journal	Significance	Within article	Authors	2013
Reproductive Health	Journal	Plain English summary	Within article	Authors	2016

Table 3.5: Summary of journals with Plain Language Summary

require acquiring other knowledge resources, or decomposing the background explanation task into more granular subtasks. Another key observation is that human ratings are es-

sential to assessment of background explanation task performance. Although we included automated evaluation metrics for generation quality and text simplicity, they cannot capture explanation quality. Rater evaluations of the external content for existence, relevance, and correctness of background explanations suggest additional advantages for RALL models that are opaque to automated evaluation methods.

To the best of our knowledge, CELLS is the largest lay language generation dataset developed to date, and the derived dataset for background explanation serves as the first explanation generation benchmark. We envision these data broadly applying to biomedically-related applications, and other NLP methods. On the biomedical applications side, we provide a benchmark to develop new NLP tools to generate LLSs for scientific literature. With the assistance of such NLP tools, researchers can write more understandable text, allowing healthcare consumers to interpret and apply the information it contains to guide their health-related decision-making. From an NLP methodological perspective, these datasets offer an excellent opportunity to develop and evaluate novel LLS generation techniques. CELLS can also support sentence-level and paragraph-level simplification research, and with additional annotation could provide a basis for open-question answering and informational retrieval tasks.

While we evaluated the correctness of key information in human evaluation, it remains difficult for non-experts to identify the factuality or external information relevance. An improved model with factuality enforcement could promote sequences with higher accuracy. Medical experts (i.e. medical students) could be recruited for evaluation of factual correctness. More abstracts and human raters are required to confirm the apparent appeal of LLS from retrieval-augmented text generation models. Furthermore, to improve the quality of the dataset for background explanation, larger-scale verification is needed. We also note that our strategies for adding entity-driven explanations are straightforward, and that we did not perform a hyperparameter search to optimize the relatively expensive dense retrieval procedure, on account of resource constraints.

Evaluating lay language generation inherently poses challenges due to the multifaceted nature of the task, including aspects such as incorporating background explanations and omitting technical terms. While the ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002)

metrics are conventionally applied to evaluate lay language generation, their applicability is constrained due to inherent limitations associated with their reliance on lexical overlap, and the need for high-quality reference summaries. Moreover, these metrics are not adept at detecting hallucinations, a critical consideration, especially in the healthcare domain where the accuracy of lay language plays a pivotal role in informing health decisions (Wallace et al., 2021; Pagnoni et al., 2021). A recent investigation indicated that ROUGE, BLEU, METEOR, and BERTScore face challenges in capturing text simplification precisely (Guo et al., 2023a). Additionally, Mac et al. found that automated readability scores frequently display inconsistency and lack accuracy (Mac et al., 2022). While human evaluations provide comprehensive feedback, they are resource-intensive, making them challenging to scale to extensive datasets. Therefore, a metric tailored specifically for lay language generation is much needed, and one should exercise caution when interpreting results using existing measures.

The GPSS algorithm searches for external content by calculating the lexical similarity between the source and target sentences using the ROUGE score. However, this may fail to recognize alignment at the semantic level when meanings but not terms overlap. To address this challenge, end-to-end approaches that can learn embedding-derived similarities from the source and target and classify the external content accordingly may be worth exploring. Since language models pre-trained in the medical domain have achieved state-of-the-art performance on several biomedical NLP tasks, exploring the benefits of these models is an important direction for future work. Regarding lay language generation, one remaining challenge that is a possible direction for future work involves directly applying retrieval-augmented methods to full-length abstracts instead of the background sections. Also, it would be intriguing yet challenging to generate LLS for different education levels. One potential solution may be incorporating a reward function that responds to readability, interpretability, or plainness metrics.

LLMs show promise in the realm of lay language generation. While the outputs from LLMs may not align closely with the target, the produced text is notably easy to comprehend. This ability to simplify addresses a key challenge in the existing lay language generation datasets, which typically offer only a single target. This suggests the potential

to develop a pipeline that first broadens the source and then tailors the content for varying levels of readability. Moreover, the less-than-optimal performance of LLMs underscores the potential of exploring few-shot learning further. Finally, incorporating Wiki-definitions could be problematic for LLMs operating in a zero-shot learning mode. For those wishing to employ retrieval-augmented methods with LLMs, a more judicious selection of external resources or a thorough vetting of the incorporated resources is imperative.

3.6 Conclusion

To improve the interpretability of lay language text generated by neural language models, we applied state-of-the-art text generation models augmented with retrieval components and achieved promising quality and readability scores as compared with reference lay language summaries generated by human experts. Results from human evaluation support the benefits of retrieval-augmented lay language generation for the generation of background explanations in particular. The new dataset and results provide a foundation for advancement in the challenging area of automated background explanation generation and lay language generation, with the potential to mediate clearer communication of biomedical science for better informed health-related decision making.

Chapter 4

CHALLENGE 3. LACK OF TAILORED EVALUATION METRICS

To address the concern of the efficiency of existing metrics in evaluation PLS, in the work described in this chapter, I first define four criteria that a PLS metric should capture. I then design a granular meta-evaluation testbed, APPLS, designed to evaluate metrics for PLS based on those criteria. An analysis of metrics using our testbed reveals that current metrics fail to capture simplification consistently. In response, I introduce POMME, a new metric designed to assess text simplification in PLS. The APPLS testbed and POMME is available at https://github.com/LinguisticAnomalies/APPLS.

A version of this chapter was previously published in arXiv. ©arXiv.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. "APPLS: A Meta-evaluation Testbed for Plain Language Summarization." arXiv preprint arXiv:2305.14341 (2023).

4.1 Introduction

Plain language summaries of scientific information are important to make science more accessible (Kuehne and Olden, 2015; Stoll et al., 2022) and inform public decision-making (Holmes-Rovner et al., 2005; Pattisapu et al., 2020). Recently, generative models have made gains in translating scientific information into plain language approachable to lay audiences (August et al., 2022c; Goldsack et al., 2023b; Devaraj et al., 2021). Despite these gains, the field has not reached consensus on effective automated evaluation metrics for plain language summarization (PLS) (Luo et al., 2022a; Ondov et al., 2022) due to the multifaceted nature of the PLS task. Removal of unnecessary details (Pitcher et al., 2022), adding relevant background explanations (Guo et al., 2021), jargon interpretation (Pitcher et al., 2022), and text simplification (Devaraj et al., 2021) are all involved in PLS, posing challenges for comprehensive evaluation.

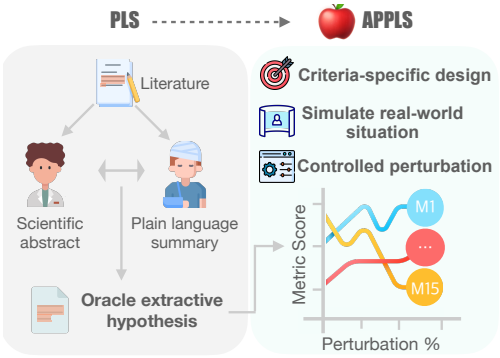


Figure 4.1: We present APPLS, the first granular testbed for analyzing evaluation metric performance for plain language summarization (PLS). We assess performance of 15 existing metrics and our new metric POMME.

We aim to assess how well existing metrics capture the multiple criteria of PLS. We define four criteria, informed by prior work (Pitcher et al., 2022; Ondov et al., 2022; Stoll et al., 2022; Jain et al., 2022), that a PLS metric should be sensitive to: *informativeness*, *simplification*, *coherence*, and *faithfulness*. We introduce a set of perturbations to probe metric sensitivity to these criteria, where each perturbation is designed to affect a single criterion with ideally minimal impact to others.[1] We produce the APPLS meta-evaluation testbed by incrementally introducing perturbations to the texts of two scientific PLS datasets, CELLS (Guo et al., 2022a) and PLABA (Attal et al., 2023).

Using APPLS, we analyze 15 metrics, including the most widely used metrics in text simplification and summmarization, and recently-proposed prompt-based evaluation (Gao et al., 2023; Luo et al., 2023). We find that while established metrics like ROUGE (Lin, 2004), BERTScore (Zhang et al., 2019), and QAEval (Deutsch et al., 2021) demonstrate mixed sensitivities to perturbations associated with informativeness, coherence, and faithfulness; all tested metrics, including those explicitly crafted for text simplification (Xu et al., 2016; Maddela et al., 2022), display a lack of sensitivity towards simplification perturbations.

[1]We acknowledge that introducing any change in language likely affects multiple criteria, though we seek to minimize these impacts through our suite of controlled perturbations.

In response to the lack of effective metrics for simplification, we introduce POMME, a new metric that evaluates text simplicity by calculating normalized perplexity differences between language models (LMs) trained on in-domain (i.e., scientific) and out-of-domain (i.e., web) text. We show POMME's effectiveness at capturing differences in text simplicity through extensive experiments on APPLS and other text simplification datasets. Because POMME is normalized to a reference dataset, it also allows text simplicity to be compared across different text corpora.

Our main contributions are as follows:

- We present APPLS, the first granular testbed for analyzing evaluation metric performance for plain language summarization(§4.3, 4.4);
- We assess the performance of existing evaluation metrics, demonstrating mixed effectiveness in evaluating informativeness, coherence, faithfulness, and simplification (§4.5, 4.7);
- We introduce a new metric, POMME, which employs language model perplexity to assess text simplicity, and validate its performance in our testbed and two additional datasets (§4.6, 4.7).

4.2 Related Work

Limitations of Existing Metrics The primary approach for evaluating plain language summaries incorporates evaluation metrics for summarization and simplification, and human evaluation (Jain et al., 2021; Ondov et al., 2022). While ROUGE (Lin, 2004) and BLEU (Sulem et al., 2018) are frequently employed in PLS assessment, their efficacy is limited due to the reliance on high-quality reference summaries, which are often challenging to obtain for PLS. Further, these metrics struggle to accurately identify hallucinations, especially crucial for PLS in the health domain to accurately inform health decisions (Wallace et al., 2021; Pagnoni et al., 2021). Though human evaluation offers thorough assessment, the high costs and time needed impede scalability for larger datasets. While recent progress in prompt-based evaluation shows potential for assessing factuality (Luo et al., 2023) and summarization quality (Gao et al., 2023), their efficacy for PLS is yet to be validated. Our work aims to fill these gaps through a systematic examination of these metrics within the

PLS context.

Robust Analysis with Synthetic Data Synthetic data has been widely used in NLP tasks to evaluate metrics, including text generation (He et al., 2022; Sai et al., 2021), natural language inference (Chen and Eger, 2022; McCoy et al., 2019), question answering (Ribeiro et al., 2019), and reading comprehension (Sugawara et al., 2020). Yet, no prior work has focused on the PLS task or incorporated simplification into their benchmarks. Additionally, previous studies lack granular analyses to capture the nuanced relationship between text changes and score changes. Our research endeavors to bridge these gaps by crafting perturbations that mirror real-world errors within the PLS context.

4.3 Criteria-Specific Perturbation Design

Notations: ~~removals~~/ additions/ modifications

Original text	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them. The first step is an accurate assessment of the population prevalence of past infections… (Kline et al., 2021)		
Criterion	**Perturbation**	**Simulated real-world situation**	**Perturbed text**
Informativeness	Delete sentences	Salient information missing	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them. ~~The first step is an accurate assessment of the population prevalence of past infections~~…
	Add out-of-domain sentences	Out-of-domain hallucination	In this paper we address the problem of aggregating the outputs of classi ers solving different nlp tasks. Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them…
	Add in-domain sentences	In-domain hallucination	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them. This review synthesised the latest evidence on the reduction of antipsychotic doses for stable individuals with schizophrenia…
	Add definitions	Background explanation	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them. Coronaviruses are species in the genera of virus belonging to the subfamily Coronavirinae in the family Coronaviridae. Coronaviruses are enveloped viruses with a positive-sense RNA genome and with a nucleocapsid of helical symmetry. The genomic size of coronaviruses ranges from approximately 26 to 32 kilobases, extraordinarily large for an RNA virus. …
Simplification	Replace sentences	Paraphrasing with simple terms	SARS-CoV-2 is a virus that has infected over 59 million people globally and killed more than 1.39 million. Scientists are trying to learn more about the virus in order to design interventions to slow and stop its spread. One of the first steps is understanding how many people have been infected in the past, which requires accurate population prevalence studies…
Coherence	Reorder sentences	Poor writing flow	The first step is an accurate assessment of the population prevalence of past infections. Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them…
Faithfulness	Number swap	Human errors	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 64 million people and killed more than one of them…
	Entity swap	Human errors	Worldwide, canine adenovirus (CaV-2), a severe acute respiratory syndrome, has infected more than 59 million people and killed more than one of them…
	Synonym verb swap	Human errors	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, has infected more than 59 million people and stamped out more than one of them…
	Antonym verb swap	Human errors	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, infected more than 59 million people and saved more than one of them…
	Negate	Human errors	Worldwide, coronavirus 2 (SARS-CoV-2), a severe acute respiratory syndrome, hasn't infected more than 59 million people and killed more than one of them.

Table 4.1: Example perturbations for criteria in APPLS. Original text comes from the CELLS (Guo et al., 2022a).

We define four criteria that an effective PLS evaluation metric should be sensitive to based on both abstractive summarization (Sai et al., 2022) and plain language summarization paradigms (Pitcher et al., 2022; Ondov et al., 2022; Stoll et al., 2022; Jain et al., 2022). To assess metric sensitivity, we develop perturbations for each criteria (illustrative examples in Table 4.1. We define *sensitivity* similar to prior work (Gabriel et al., 2020) as being correlated in the expected direction with the amount of perturbation. These criteria and our designed perturbations are:

4.3.1 Informativeness

Delete sentences We simulate the omission of information by ranking sentences based on similarity to others (assuming greater similarity indicates more important content) (Zhong et al., 2020) and removing sentences starting from the most to least similar.

Add sentences We simulate the inclusion of two forms of unrelated information by adding sentences from *out-of-domain* (i.e., unrelated dataset) and *in-domain* (i.e., within the same domain but on a different topic).

Add definitions Background explanation is fundamental to PLS and involves adding external content like definitions or examples (Guo et al., 2022a; Srikanth and Li, 2020a). To simulate background explanations, we add definitions[2] of keywords identified by KeyBERT (Grootendorst, 2020).

4.3.2 Simplification

Replace sentences Taking advantage of the LLMs' ability to simplify text (Lu et al., 2023b), we replace sentences in the original text with LLM-simplified versions. We use a language model to generate simplified summaries using the prompt ``explain the text in layman's terms to a primary school student.'' We use GPT-3 (*text-davinci-003*), Llama 2 (*llama-2-13b-chat*), and Claude (*claude-instant-v1.0*)[3] for text simplification. The maximum length of generation is set to 200.

[2]Taken from http://wikidata.dbpedia.org/develop/datasets

[3]https://www.anthropic.com

4.3.3 Coherence

Reorder sentences We simulate changes in text coherence by randomly shuffling the order of sentences, as suggested by (Sai et al., 2021).

4.3.4 Faithfulness

Number swap We randomly add a number from 1 to 5 to the original numerical value in the text.

Verb swap An appropriate metric should ignore verb synonyms but be sensitive to antonyms. We introduce two perturbations by substituting verbs with either synonyms or antonyms.

Entity swap We replace entities using the KBIN method (Wright et al., 2022), which links entity spans to concepts in the Unified Medical Language System (UMLS) and replaces them with different entities while maximizing NLI contradiction and minimizing LM perplexity. This results in a fluent sentence that contradicts the original one.

Negate sentences We negate sentences by identifying verbs and adding negation terms (e.g., not) preceding them. The goal of this perturbation is to create sentences similar to the original but communicating the opposite information.

Designed perturbations allow us to control each criteria. We ensure perturbed text quality by manually validating a subsample of perturbations (§4.4.3).

4.4 Constructing the APPLS *Testbed*

We implement our perturbations in two existing large-scale PLS datasets (§4.4.1). We describe how perturbations are incorporated into the dataset and our approach for managing perturbation magnitude (§4.4.2) and validating perturbation quality (§4.4.3). We employ this testbed in an analysis of existing (§4.5) and novel (§4.6) metrics for PLS (§4.7).

Dataset	Version	Word	Sentence
CELLS	Abstract (*src.*)	$283_{\pm 132}$	$11_{\pm 6}$
(n=6,311)	PLS (*tgt.*)	$178_{\pm 74}$	$7_{\pm 3}$
	Oracle Hypothesis	$134_{\pm 58}$	$5_{\pm 2}$
	GPT-simplified	$98_{\pm 57}$	$4_{\pm 3}$
PLABA	Abstract (*src.*)	$240_{\pm 95}$	$10_{\pm 4}$
(n=750)	Adaptation (*tgt.*)	$244_{\pm 95}$	$12_{\pm 5}$

Table 4.2: Diagnostic datasets statistics (mean±std).

4.4.1 Diagnostic datasets

For our experiments, we use the CELLS (Guo et al., 2022a) and PLABA (Attal et al., 2023) datasets. CELLS (Guo et al., 2022a) is a parallel corpus of scientific abstracts (*source* texts) and their corresponding plain language summaries (*target* texts), which are written by the abstract authors or by other domain experts. CELLS aggregates papers from 12 biomedical journals, representing a diverse set of topics and summaries, and serves as the primary dataset in our testbed. The PLABA (Attal et al., 2023) dataset includes expert-modified biomedical abstracts, simplified to improve understanding of health-related content. The modification involves rule-based adjustments, such as lexical simplification, shifting from passive to active voice, and segmenting long sentences, leading to greater word overlap compared to the CELLS dataset. PLABA includes sentence-level alignment, which is useful for controlled perturbations, but is nonetheless not used as our primary dataset in APPLS due to the simplifications being relatively contrived. Moreover, PLABA's pronounced n-gram overlap between sources and summaries tend to skew results for evaluation metrics that prioritize n-gram overlap, potentially compromising the generalizability of the assessment. Therefore, PLABA serves as an auxiliary dataset to CELLS, helping to address its limitations, discussed in Sections §4.4.2 and §4.4.3. We report simplification perturbation results for PLABA in the main paper and remaining perturbation results in Sup. A.5.

4.4.2 *Applying perturbations to datasets*

While the majority of metrics we assess only require targets and model-generated text (*hypothesis*), SARI and LENS additionally make use of the source text. For the APPLS testbed, we propose an *oracle hypothesis*, a reasonable extractive hypothesis that summarizes the source text with lexical variations while minimizing factual inaccuracies (Guo et al., 2021). For CELLS, the oracle hypothesis is created by selecting the set of source sentences yielding the highest ROUGE-L score when compared to the target summary and then introducing lexical variability through round-trip translation (Ormazabal et al., 2022).[4] Because PLABA is sentence-aligned, no extraction is needed, and the oracle hypothesis is created simply by round-trip translating the target. Details are in Sup. A.1.

We apply all perturbations to the oracle hypotheses, where each perturbation introduces a *change* (e.g., add/swap sentences) at some *magnitude* (e.g., replace 50% of sentences). Given costs associated with some of our perturbations (e.g., LLM-based simplification), we restrict perturbation experiments to dataset test splits (stats in Table 4.2).

For *informativeness*, we add sentences to the oracle hypothesis from ACL papers (Bird et al., 2008) to simulate out-of-domain hallucinations and Cochrane abstracts[5] for in-domain hallucinations. For sentence addition, we add up to the same number of sentences as in the oracle hypothesis. For sentence deletion, we delete sentences until a single sentence remains. For keyword definitions, we add up to three definitions, the average number of nouns explained in CELLS abstracts (Guo et al., 2022a), i.e., 100% perturbed adds three definitions.

For *simplification*, for CELLS, we first generate an LLM-simplified summary from the oracle extractive hypothesis. We then align sentences between the oracle hypothesis and LLM-simplified summary using the sentence alignment algorithm from Guo et al. (2022a). We perturb the text by replacing hypothesis sentences with their corresponding LLM-simplified sentences randomly until full replacement. We use GPT-3 (Brown et al., 2020a) to generate LLM-simplified text due to its accessibility and demonstrated proficiency in

[4] Why not use the extractive summary directly? Metrics like SARI expect simplified hypotheses and exhibit degenerate behavior when used to evaluate extractive summaries.

[5] https://community.cochrane.org

text simplification (Lu et al., 2023b). To ensure that our findings are not specific to the chosen model, we conduct additional experiments using Llama2 (Touvron et al., 2023) and Claude[6] (details in Sup.§A.6). For PLABA, we perturb text by replacing source sentences with round-trip translated versions of their aligned simplified targets (no LLM is used). Source and target lengths in PLABA are roughly equivalent, allowing us to evaluate metric response when there are minimal changes in text length.

For *coherence*, we shuffle sentences in the hypothesis and quantify perturbation percentage as the distance between the original and shuffled hypotheses in terms of absolute difference in sentence order. A document with reversed sentence order would be 100% perturbed.

For *faithfulness*, perturbation percentage of number, entity, and verb swaps is determined by comparing the count of altered spans to the total number of eligible spans in the hypothesis. Full perturbation means all eligible spans are swapped. For sentence negation, we constrain the maximum number of negations to the sentence count in the hypothesis, allowing for a max of one negation per sentence. Therefore, full perturbation is achieved when each sentence contains a negation.

To mitigate the effects of randomness, we use two random seeds to produce perturbations.

4.4.3 Human validation of oracle extractive hypotheses and GPT-simplified summaries

We assess the quality of oracle extractive hypotheses and GPT-simplified summaries through human evaluation. We sample 100 pairs each of (i) pre- and post-round-trip translation (RTT) oracle hypotheses and (ii) GPT-simplified summaries paired with oracle hypotheses. Annotators were asked to assess content alignment (defined as having identical relation triples) and rate informativeness, simplification, faithfulness, and coherence on 5-point Likert scales. Annotations were performed by two independent annotators, both with doctorates in the biological sciences, who were hired on UpWork and compensated at 21 USD/hr. Each annotator reviewed all sampled pairs for both evaluation tasks. Inter-rater

[6] https://www.anthropic.com

agreement measured by Cohen's Kappa was 0.29, implying fair agreement (Artstein and Poesio, 2008). For task details, refer to Sup. A.2.

Human annotators affirmed that RTT text retained its informativeness (98%), faithfulness (83%), coherence (100%), and simplicity (96%) compared to the original. Evaluators considered GPT-simplified sentences more simplified (98%), informative (63%), faithful (61%), and coherent (99%). In this context, neutral implies the same level of informativeness/simplicity between the two texts, so we report annotations equal to or better than neutral as positive. We observe that the alignment algorithm employed for simplification can lead to decreased informativeness and faithfulness; to mitigate the impact of such misalignment, we utilize the PLABA dataset for auxiliary diagnostics because it contains sentence-level alignments.

4.5 Existing Metrics

Our analysis spans 8 established evaluation metrics, including the 5 most commonly reported in ACL'22 summarization/generation papers (empirical results in Sup. A.3). We also assess 5 lexical features associated with text simplification (§4.5.2) and LLM-based evaluations (§4.5.3).

4.5.1 Existing automated evaluation metrics

Overlap-based metrics measure n-gram overlaps, and are popular due to their ease of use.

- **ROUGE**[7] (Lin, 2004) measures n-gram overlap between generated and reference summaries, focusing on recall. We report the average of ROUGE-1, ROUGE-2, and ROUGE-L.
- **BLEU**[7] (Papineni et al., 2002) computes n-gram precision of generated text against reference texts, including a brevity penalty.
- **METEOR**[7] (Banerjee and Lavie, 2005) employs a relaxed matching criterion based on the F-measure, and addresses the exact match restrictions and recall consideration of BLEU.

[7]Implementation: Fabbri et al. (2021) BERTScore hash code: bert-base-uncased_L8_no-idf_version = 0.3.12(hug_trans=4.27.3).

- **SARI**[8] (Xu et al., 2016) is specifically designed to evaluate text simplification tasks. The score weights deleted, added, and kept n-grams between the source and target texts.

Model-based metrics use pretrained models to evaluate text quality.

- **GPT-PPL**,[9] usually computed with GPT-2, measures fluency and coherence by calculating the average log probability assigned to each token by the GPT model, with lower scores indicating higher fluency and coherence.
- **BERTScore**[7] (Zhang et al., 2019) quantifies the similarity between hypothesis and targets using contextualized embeddings from the BERT model, computing the F1-score between embeddings to capture semantic similarity beyond n-gram matching.
- **LENS** (Maddela et al., 2022) employs an adaptive ranking loss to focus on targets closer to the system output in edit operations (e.g., splitting, paraphrasing, deletion).

QA-based metrics capture content quality using a question-answering approach.

- **QAEval** (Deutsch et al., 2021) generates question-answer pairs from the target text, then uses a learned QA model to answer these questions using the generated text. The score is computed as the proportion of questions answered correctly. We report QAEval LERC scores.

4.5.2 Lexical features

We also assess lexical features that have been shown to be associated with text simplicity:

- **Length**: Shorter sentences are easier to understand (Kauchak et al., 2017). We report both sentence length and paragraph length.
- **Familiarity**: Simple text contains more common words (Leroy et al., 2018). We compute the percentage of text that is made up of the 1,000 most common English words.[10]
- **Specificity**: Specificity quantifies the level of detail in the text. We use Speciteller (Ko et al., 2019) to compute the domain agnostic specificity of terms in the paragraph.
- **Phrase Transitions**: Conjunctions (e.g., therefore) are important for flow and can assist

[8] Implementation: Alva-Manchego et al. (2019)

[9] https://huggingface.co/transformers/v3.2.0/perplexity.html

[10] https://gist.github.com/deekayen/4148741

with comprehension (Kauchak et al., 2017). We report the number of conjunctions.

- **Function Words**: Simple text contains more verbs and fewer nouns (Mukherjee et al., 2017). We report the number of verbs, nouns, adjectives, adverbs, and numbers.

4.5.3 LLM prompt-based evaluations

Prompting LLMs for text generation evaluation has been explored in recent work (Gao et al., 2023; Luo et al., 2023). We adopt the prompt template from Gao et al. (2023) to have GPT-3 (*text-davinci-003*) evaluate each hypothesis on four criteria—informativeness, simplification, coherence, and faithfulness—and provide an overall quality score. All scores range from 0 (worst) to 100 (best). We supply definitions for each criterion in the prompt. We evaluate under two settings: providing only the source abstract (reference-free) and providing both source and target (reference-provided). Model configurations and prompts are available in Sup. A.4.

4.6 Novel Metric: POMME

We introduce a novel, lightweight metric (POMME) to assess text simplification by leveraging pretrained LMs. LMs like GPT-2 have been used to assess readability through perplexity scores (Zhao et al., 2022; Kanthara et al., 2022), but these measures exhibit considerable sensitivity to text length (Wang et al., 2022), which is undesirable for PLS evaluation. Our own investigation corroborates this, showing divergent raw perplexity scores for different simplification datasets (Tables 4.3; 4.4).

POMME addresses this issue by employing the *difference* in perplexity scores from an in-domain and out-of-domain LM, leveraging the inherent domain shift from scientific to plain language in PLS and minimizing sensitivity to text length. The perplexity scores from these LMs are normalized relative to a reference dataset of complex and plain language texts, which addresses differences in magnitude when comparing perplexity scores from models with distinct vocabulary sizes. Specifically, POMME is computed by taking the difference of perplexity Z-scores rather than using raw values. POMME is computed as:

$$Z(x) = \frac{\log(x) - \mu_{\text{ref}}}{\sigma_{\text{ref}}}$$

$$\text{POMME} = Z\left(\text{PPL}_{\text{id}}\right) - Z\left(\text{PPL}_{\text{ood}}\right)$$

where μ is the mean and σ the standard deviation of the perplexity of texts in the reference dataset (we use CELLS). We use BioMedLM (Bolton et al., 2023) as our in-domain ("scientific") LM and T5 (Raffel et al., 2020) as our out-of-domain ("plain"). BioMedLM was trained exclusively on PubMed abstracts (matching the domain of our source texts) while T5 was trained primarily on general-domain data like web text and Wikipedia (more closely matching our target texts).

The core idea is that scientific LMs should assign lower perplexity scores to scientific texts than general English LMs, with the opposite holding true for plain language (Harris et al., 2012). Similar logic has been used successfully for controlled text generation (Liu et al., 2021; August et al., 2022a). POMME, by quantifying a text's perplexity within a perplexity score distribution from a reference dataset, guarantees the compatibility of POMME scores across varied datasets. An advantage of POMME is its model-agnosticism, enabling any two models to serve as in- and out-of-domain LMs. Thus, POMME could be adapted to evaluate text simplification in other fields such as law (Jain et al., 2021) or finance (Salo et al., 2016). In this work, we limit POMME evaluation to biomedical text, given the available pretrained models and paired PLS datasets in this domain.

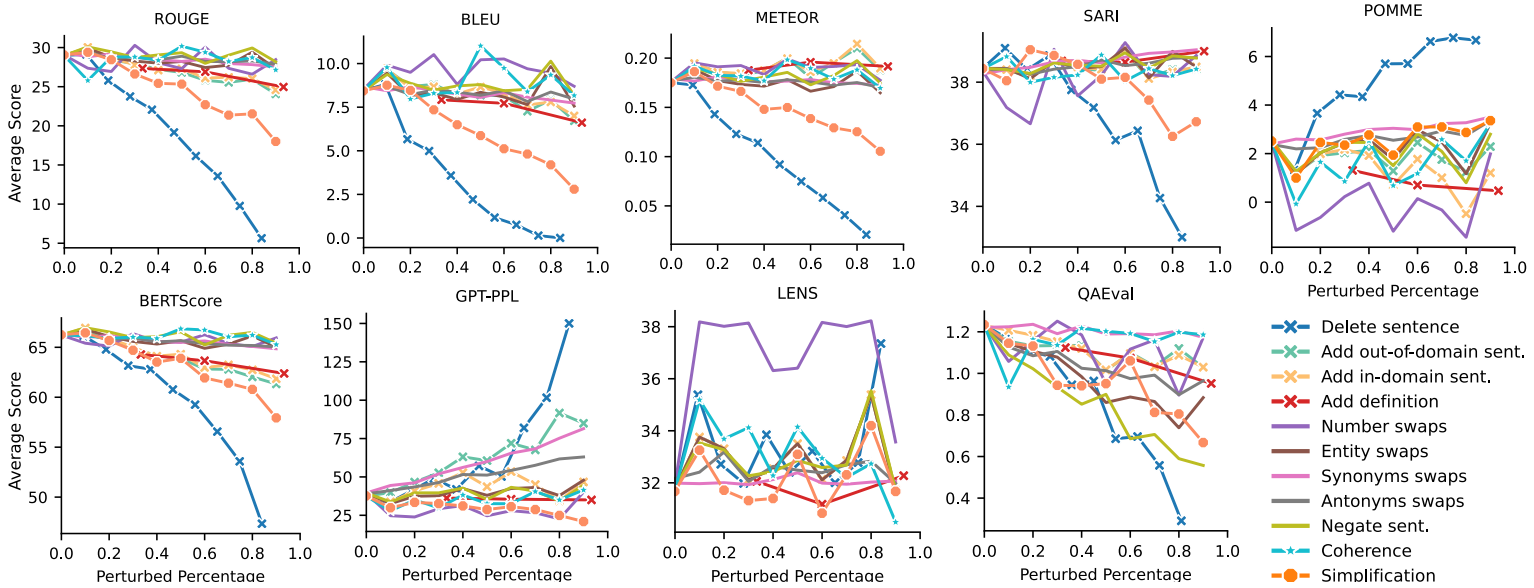


Figure 4.2: Average scores of existing metrics and our POMME score for perturbed texts. Scores are averaged in 10 bins by perturbation percentage. Markers denote the defined criteria associated with that perturbation. GPT-PPL is the only metric exhibiting sensitivity to the simplification perturbation (i.e., PPL decreases when simplification perturbation % increases, signifying simpler text). Median reported improvements in ACL'22 summarization and generation papers are ROUGE (+0.89), BLEU (+0.69), METEOR (+0.50), SARI (+1.71), BERTScore (+0.55), and PPL (-2.06).

4.7 Analysis Results

Metric responses to perturbations are presented in Figure 4.2. All score trends are consistent across two random seeds. For contextualizing metric performance in APPLS, we survey reported metric changes in ACL'22 papers on text generation and summarization (full results in Sup. A.3). The median reported improvements are: ROUGE (+0.89), BLEU (+0.69), METEOR (+0.50), SARI (+1.71), BERTScore (+0.55), and PPL (-2.06). We summarize our main findings below.

Current metrics exhibit shortcomings in evaluating simplicity. Metrics that are sensitive to simplification should consistently distinguish between more and less simplified text. As shown in Figure 4.2, the only metric that exhibits appropriate sensitivity to simplification perturbations is GPT-PPL (decreasing as more perturbations are introduced; lower PPL is better). However, in follow-up evaluations with other datasets (discussed below and shown in Table 4.4), we see that GPT-PPL has undesirable sensitivity to text length, as found in prior work (Zhao et al., 2022). ROUGE, BLEU, METEOR, SARI, BERTScore, and QAEval decrease in response to the simplification perturbation. While their response is consistent, we posit this is due to sensitivity to n-gram overlap rather than text simplicity. To confirm, we report metric changes when reversing sources and targets (perturbing simplified texts to increase complexity). Metrics also decrease in this case (Sup. Figure A.11), suggesting that they are not sensitive to text simplicity. LENS and LLM prompt-based evaluations (Sup. Table A.2) are erratic or insensitive to simplification perturbations.

POMME is sensitive to simplification perturbations. In Figure 4.2, we observe that POMME increases with simplification perturbations. We further validate POMME in PLABA (perturbation results in Sup. Figure A.8) and two other text simplification datasets: MSD (Cao et al., 2020) and WikiSimple (Woodsend and Lapata, 2011). Using POMME to compare source and target texts from these datasets (Table 4.3), we observe consistently higher POMME for the source texts compared to the target texts (Δ is positive). We also present single-model PPL scores as computed by the in- and out-of-domain LMs used to compute POMME, and find that inconsistency is evident. For instance, BioMedLM-PPLΔ for MSD is

	BioMedLM-PPL			T5-PPL			POMME		
Datasets	**Source**	**Target**	**Δ (↑)**	**Source**	**Target**	**Δ (↓)**	**Source**	**Target**	**Δ (↑)**
CELLS	-0.36	0.36	**0.72**	0.52	-0.52	**-1.04**	-0.88	0.88	**1.76**
PLABA	-0.79	-0.14	**0.65**	0.29	0.31	0.02	-1.08	-0.45	**0.63**
MSD	3.30	3.30	0.0	-1.89	-1.94	**-0.05**	5.19	5.24	**0.05**
WikiSimple	1.28	2.47	**1.19**	-1.12	-3.23	**-2.11**	2.40	5.70	**3.30**

Table 4.3: BioMedLM-PPL, T5-PPL and POMME scores for four simplification datasets, comparing source (complex text) and target (simple text). A higher POMME score indicates a higher degree of text simplification. The difference, denoted Δ, is calculated by subtracting the source score from the target score. **Bold** indicates statistical significance in the correct direction (i.e., target is simpler than source) with Bonferroni-Holm correction for multiple hypothesis testing (Holm, 1979). The CELLS dataset functions as the reference in all POMME computations.

	CELLS		PLABA	
Perturb%	**PPL**	**POMME**	**PPL**	**POMME**
20%	**-44.81**	-5.32	3.06	-0.38
40%	**-17.52**	-0.88	5.01	**0.69**
60%	**-14.14**	-0.29	4.59	**0.83**
80%	**-11.55**	0.27	3.38	**0.76**
100%	**-15.45**	**0.42**	1.80	**0.68**

Table 4.4: Delta in PPL and POMME scores at various levels of perturbation. **Bold** indicates statistical significance in the correct direction with Bonferroni-Holm correction for multiple hypothesis testing (Holm, 1979).

0.0 and T5-PPLΔ for PLABA is 0.02, suggesting incorrectly that the source texts (scientific abstracts) are simpler or as simple as the targets (plain language summaries).

To further validate the sensitivity of POMME to simplification, we show results over CELLS and PLABA by % perturbed in Table 4.4. Especially for PLABA, GPT2-PPL is insensitive to perturbations, potentially due to similar text lengths between the scientific abstracts and plain text in PLABA. Conversely, POMME consistently reacts to perturbations, producing higher scores for more extensively altered text. To ascertain that POMME responds to text simplification itself and not merely to the characteristics of GPT-simplified text, we conduct further tests by generating simplified text using Llama2 and Claude. These results, illustrated in Sup. Figure A.10, reveal that POMME trends across these three models exhibit similar patterns.

Using a reference dataset to normalize perplexity enables POMME to be used for cross-dataset comparisons of text simplicity. In Table 4.3, we observe that based on POMME, the source and target texts of both MSD and WikiSimple are much simpler than those of CELLS. This aligns with their content: MSD contains consumer-health information, usually simpler than plain language summaries of research papers, and WikiSimple is sourced from English and Simple Wikipedias, both of which feature language suited for the general public. This supports the use of POMME to compare text simplicity across corpora. Domain adaptation can be further enabled by selecting a domain-specific reference dataset and domain-adapted LMs (Gururangan et al., 2020b) to compute POMME.

Metrics effectively capture informativeness, coherence, and faithfulness, with room for improvement. For informativeness, ROUGE, BLEU, BERTScore, GPT-PPL, and QAEval are sensitive to information deletion and irrelevant additions, but decrease with the addition of background explanations through keyword definitions. For coherence, BERTScore and LENS excel in detecting perturbations, largely due to their ability to assess structural and contextual sentence relationships. BERTScore, GPT-PPL, and QAEval generally perform well for faithfulness-related perturbations, although GPT-PPL and BERTScore are somewhat sensitive to synonym verb swaps, an undesirable trait. QAEval is best at being unresponsive to synonym verb swaps. Number swaps, however, remain undetected by all metrics. Results in Figure 4.2.

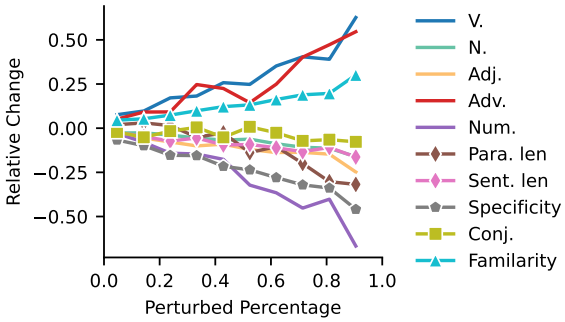


Figure 4.3: Relative change of each lexical feature with respect to the unperturbed state (0%). Different markers represent lexical feature categories.

Lexical features are useful measures of text simplicity. Figure 4.3 illustrates the response of lexical features to degrees of text simplification in CELLS, confirming trends observed in previous studies (Kauchak et al., 2014; Leroy et al., 2018; Kauchak et al., 2017; Mukherjee et al., 2017). As simplification increases, sentence length decreases; common words and verbs increase; and nouns, adjectives, and term specificity decrease. Although prior work emphasizes the importance of conjunctions for comprehension (Kauchak et al., 2017), our study reveals a reduction rather than increase in conjunctions as texts become simpler. Overall, these trends demonstrate that lexical features are valuable indicators for text simplification. Results on PLABA are similar, with an inverse trend for paragraph length (Sup. Figure A.9).

LLM prompt-based evaluations do not distinguish between PLS criteria. Prompt-based evaluations are insensitive to simplification perturbations, and in most cases, do not distinguish between the four criteria when scoring summaries (Sup. Figure A.7). Despite findings from Luo et al. (2023) showing agreement between ChatGPT scores and human ratings, our results suggest that the capacity of LLMs for generative text evaluation warrants further examination. We also note that the reference-free and reference-provided settings yield very different scores along all four criteria, indicating that scores produced with this method are difficult to compare across settings and datasets. Detailed results are provided in Sup. A.4.

4.8 Discussion & Conclusion

Recent advances point to the possibility of automated plain language summarization (PLS); however, the multifaceted nature of PLS makes evaluation challenging. We introduce the first—to our knowledge—meta-evaluation testbed, APPLS, for evaluating PLS metrics. In APPLS, we apply controlled text perturbations to existing PLS datasets based on several criteria (informativeness, simplification, coherence, and faithfulness). Using APPLS, we find that while some metrics reasonably capture informativeness, faithfulness, and coherence, they face challenges assessing simplification. Most metrics decrease, contrary to expectation, when computed for more simplified text. GPT-2 perplexity, the sole metric sensitive to simplification, exhibits inconsistencies across datasets.

In response to these shortcomings, we propose POMME. By using normalized perplexity differences between in and out-of-domain language models, POMME maintains the desirable qualities of language model perplexity while being robust and comparable across datasets. It is worth noting, though, that while POMME is sensitive to simplification, it is less sensitive to other PLS criteria. In other words, no single metric is capable of capturing all desired PLS criteria and a holistic evaluation will necessitate a combination of metrics.

The primary advantage of our testbed and metric is their extensibility. Using the perturbation pipeline, APPLS can transform any PLS dataset into a granular meta-evaluation testbed. Similarly, POMME can be easily adapted to other domains, requiring only a domain-specific dataset and two language models representing the source and target domains. Our testbed and metric lay the groundwork for further advancements in automated PLS, aiming to foster more impactful, accessible, and equitable scientific communication.

Limitations

Our perturbations use synthetic data to simulate real-world textual phenomena seen in PLS. Although our approach is informed by theory and provides valuable insights into metric behavior, further exploration of more sophisticated methods to simulate changes in these criteria is warranted. This is especially true for aligning sentences between scientific abstracts and plain language summaries, as sentence-level alignment for scientific summaries

is still an open problem (Krishna et al., 2023b).

We also acknowledge that text quality may deteriorate with synthetic perturbations in a way that affects multiple PLS criteria. However, by using synthetic data, we are benefiting from the ability to control our perturbations and extend our testbed creation framework to any dataset. It is infeasible to find naturally occurring text with the same controlled levels of each perturbation, with minimal changes to other aspects. Our aim is not to produce perfect outputs, but rather to establish a robust baseline that enables controlled text perturbations, assisting in evaluating shifts in metric scores. The results of our analysis complement qualitative examinations of model output conducted in other work, which further suggests that automated text generation evaluation metrics may be limited in their ability to assess generation performance of post-GPT-3 LLMs (Goyal et al., 2022).

We have also focused our analysis on commonly used metrics reported in prior work on simplification, summarization, and generation. Investigating the performance of metrics not included in this work, as well as the generalizability of our methods to meta-evaluation for other generative NLP tasks, is a future goal.

Chapter 5

CHALLENGE 4. PERSONALIZATION

In the work described in this chapter, I address the need for personalized plain language summaries (PLS) by defining two key tasks: 1) identifying jargon on a personalized basis, and 2) determining individual users' additional information needs. To support this work, I collect a dataset containing over 10,000 annotations related to term familiarity. I explore various features that represent background knowledge and compare the performance of supervised learning methods and prompt-based approaches in tackling these personalization tasks. The code and anonymized version of our dataset are available at: https://github.com/talaugust/PersonalizedJargon.

A version of this chapter was previously published in the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

Yue Guo, Joseph Chee Chang, Maria Antoniak, Erin Bransom, Trevor Cohen, Lucy Lu Wang, and Tal August. "Personalized Jargon Identification for Enhanced Interdisciplinary Communication." In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2024.

5.1 Introduction

An important challenge to communicating knowledge across scientific domains is aligning on a shared vocabulary (Strober, 2006). Each scientific domain has unique terminology that optimizes communication within the field but can pose a barrier to researchers in other domains (Lucy et al., 2022; Choi and Pak, 2007). As science becomes more specialized, so too does its terminology (Barnett and Doubleday, 2020; Plaven-Sigray et al., 2017), raising the barrier of learning and collaborating across disciplines. We envision systems that can identify whether specialized terminology will be unfamiliar to an individual scholar, so that

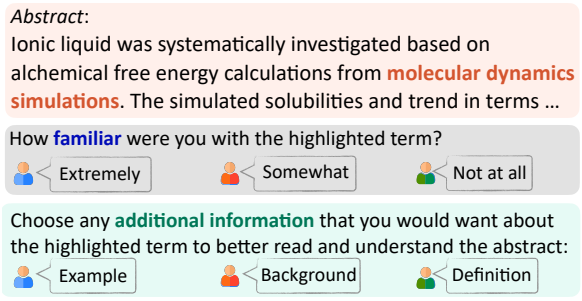


Figure 5.1: An annotated term from our dataset, with annotations by computer science researchers. Despite sharing a common domain, these researchers exhibit variation in their familiarity and additional information needs about the term within the abstract. Abstract from Liu et al. (2014).

other systems can then translate this terminology.

NLP techniques have been developed to identify and simplify scholarly jargon (Gardner and Davies, 2013; Tanaka-Ishii and Terada, 2011; Guo et al., 2022b, 2021), a first step in our envisioned setting. The majority of these techniques use a corpus of documents as a proxy for what a reader knows (e.g., Wikipedia is assumed to contain words known to a general audience). However, an individual's specific background knowledge also plays a role in determining their familiarity with a word (Gooding and Tragut, 2022). For example, a theoretical computer science (CS) researcher might struggle more with jargon in a chemistry paper than in a mathematics paper, but the opposite may be true for a computational biologist. Information on a researcher's background should help determine what they already know and what they need explained.

In this paper, we introduce the task of *personalized scholarly jargon identification*. We ground our investigation in the real-world setting of interdisciplinary reading: researchers reading papers in less familiar domains. We first validate our setting with an initial study on interdisciplinary reading. The results reveal a clear preference for supplementary information beyond what is provided in a paper abstract, especially in less familiar domains. Building on initial findings, we propose the task of predicting the familiarity and any asso-

ciated information needs of a term for an individual researcher.

To study this problem, we collect a dataset of over 10K individual familiarity ratings and information needs from 11 CS researchers about terms drawn from 100 out-of-domain abstracts (example in Figure 5.1). We enumerate features representing an individual's background knowledge based on papers they have written and read. Using these features and our dataset, we investigate baselines for estimating term familiarity, including regression models and prompt-based approaches using large language models (LLMs). Our analysis reveals that incorporating individual-level information improves the accuracy of predicting term familiarity, though the task is difficult and no one model performs best for all annotators. Our project contributes the following:

- We define the novel task of predicting personalized jargon familiarity. We motivate our task based on initial experiments with interdisciplinary computer science researchers.
- We collect a dataset of over 10K term familiarity ratings and individual information needs.
- We enumerate features representing an individual researcher's knowledge and investigate integrating these features into supervised and prompt-based methods.

5.2 Task Description

We conduct an initial study with 10 computer science researchers to (i) validate our intuition that term familiarity is important for interdisciplinary reading, and (ii) identify what information needs researchers have for unfamiliar terms when reading across domains.

We recruited participants from two subdomains of computer science: from Natural Language Processing (NLP) and from Human-Computer Interaction (HCI). Participants were asked to read two paper abstracts, one from a closer domain (Linguistics or Psychology) and one from a more distant domain (Medicine). For each paper, we provided two abstract variants: the original author-written abstract and a generated abstract personalized to the participant's background. To personalize the abstracts, we prompted GPT-3 (*text-davinci-003*) to rewrite the provided abstract as an abstract personalized to an author. The model was given the author's paper abstracts and a sampled list of citation sentences from the author's papers (details can be found in Sup. Table B.2). After reading each abstract pair, participants

identified what modifications they liked/disliked in the personalized abstract, and provided a free-text response on whether they preferred the generated abstract and why. The study was considered exempt by University of Washington's IRB.

5.2.1 Initial Findings

We collected a total of 20 responses from 10 researchers (7 from NLP and 3 from HCI). Participants generally preferred the personalized abstract over the original, with 9 preferring the personalized abstract for the medical paper (90%) and 5 preferring the personalized abstract for the linguistics or psychology paper (50%). There was a general preference for modifying the abstract by *adding* information (82% of additions preferred) over *removing* it (9% of removals preferred). Full results in Sup. Figure B.1.

We categorize participants' preferred modifications as satisfying the following information needs:

- **Definition**: key information about the term independent of any context. A definition answers the question, "What is/are [term]?"
- **Background**: information that is important for understanding the term in the context of the abstract, e.g., how the term relates to the overall problem, significance, and motivation.
- **Example**: specific instances that help illustrate the usage of the term within the abstract.
- **Method/Result Details**: details on the methodology and results of the paper.
- **Relevant Downstream Connections**: insights about how the current paper's findings relate to the reader's own research.

The first three information needs pertain to additional information for specific terms, while the latter two require further contextualization of the information in the abstract. Participants in our study generally requested additional term-specific information when they were less familiar with the term and domain, and requested contextualizing information when they were more familiar with the domain. Given the more clear association between the need for term-specific information and term/domain unfamiliarity, we focus on the first three needs—definitions, background, and examples—in the remainder of this work. We

also note that while participants generally reacted positively to relevant downstream connections in all cases, these texts were usually hallucinated by the model, so we avoid targeting these as well. Examples of modifications that the models made to the abstracts when personalizing are provided in Sup. Table B.1.

5.2.2 Task Definition

Based on these initial findings, we identify the tasks of individual term familiarity prediction and information need prediction as important steps for assisting interdisciplinary reading of scientific abstracts. We formalize the first task as: given an individual researcher defined by their authored publications $R = \{r_1, r_2, ...r_m\}$ and an abstract to personalize A, which includes a set of terms $T = \{t_1, t_2, ...t_n\}$, our goal is to predict the subset of terms unfamiliar to R. In addition, we aim to predict R's indicated information need from among {definition, background, example} for each term, as defined above.

5.3 Dataset

As no pre-existing datasets exist for personalized scientific jargon identification, we construct a new dataset of terms from abstracts with human annotations of familiarity and additional information needs. We direct our focus to abstracts that are outside the individual's domain, with CS researchers as the annotators.

5.3.1 Data Source

To ensure that the out-of-domain abstracts could realistically be read by our annotators, we compile a corpus of non-CS papers often viewed by CS researchers, published after 2010, using the Semantic Scholar API (Kinney et al., 2023). We define CS researchers as anyone who has (co-)authored a paper categorized as 'Computer Science' as classified in the API. From the top 500 viewed papers not categorized as CS, we take a stratified sample of 100 abstracts covering the 22 domains found in the top 500 papers (paper counts are in Sup. Figure B.2).

For each abstract, the top-10 significant terms are identified using the OpenAI model

text-davinci-003 (details and prompt in Sup. Table B.2). We manually review 10 abstracts and confirm that their top 10 terms align with our notion of salient terms in each abstract—we define salient terms as terms that could be provided as keywords of the paper.

5.3.2 Annotation

Annotators were asked to annotate each term with the following information:

- **Familiarity** on a scale of 1 (not at all familiar) to 5 (extremely familiar). Not at all familiar was defined as "you have never heard of this term." Extremely familiar was defined as "you have a deep, comprehensive understanding of this term."

- **Additional information needs** that could help annotators understand the abstract. These included definitions, background, and examples for each term (defined in §5.2.1). Annotators could select more than one information need per term.

We recruited 11 annotators from UpWork with a Master's (N=4) or Doctorate (N=7) degree in CS who had published at least one paper (see Table 5.1). We paid each annotator $20-30 hourly based on their degree. Each annotator reviewed all sampled abstracts and answered all questions, providing a total of 10,571 familiarity ratings for 956 terms.[1]

Familiarity data check To ensure that annotator familiarity ratings were consistent, we conducted a data check for all annotators. We selected 10 entities from each annotator, 5 rated as familiar and 5 rated as unfamiliar. For each entity, we asked annotators to provide a definition of the entity without looking up any information. If they could not define the term, we instructed them to write 'N/A'. Annotators were generally consistent with their initial scores, with 81% of responses matching initial ratings (i.e., if they were familiar, they wrote a correct definition). When initial scores did not align with the data check, annotators generally wrote definitions based on the term's context in the abstract.

[1]GPT 3.5 identified ¡10 terms from some abstracts. Upon inspecting these abstracts, the authors agreed that there were fewer than 10 salient terms to list.

ID	Degree	# of Papapers	Self-Defined Subfield
1	Master	20	Computer Vision
2	PhD	10	Networking
3	Master	1	NLP
4	PhD	20	NLP
5	PhD	30	Cyber Security
6	PhD	4	General CS theory
7	PhD	3	Neural Networks
8	PhD	60	NLP
9	PhD	15	Complex Networks
10	Master	2	Computer Vision
11	Master	2	Computer Vision

Table 5.1: Annotators' characteristics

5.3.3 Outcomes

We define a binary term familiarity outcome measure by grouping the collected 5-point familiarity ratings into the binary classes of "familiar" (ratings ≥ 3) and "unfamiliar" (ratings ≤ 2). We treat the need for additional definitions, background, and examples as separate binary classification tasks (covered by RQ5 in §5.5) .

5.3.4 Analysis

Below we describe characteristics of our dataset, focusing on how familiarity ratings and information needs exhibit variation across abstract domain and annotator background.

Domain-specific variation Figure 5.2 illustrates the differences in familiarity and additional information needs across abstract domains. Annotators were most often familiar with terms from Art, while Chemistry received the lowest familiarity ratings. This is in line with prior work, which has suggested that the technical sciences often develop more specialized vocabulary within a domain, while the social sciences and humanities share more terminol-

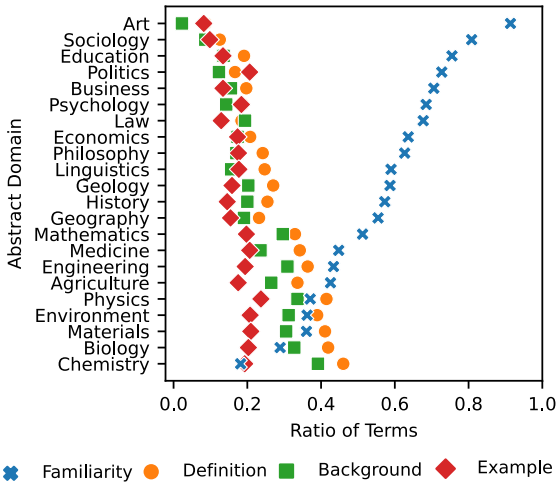


Figure 5.2: Mean familiarity and additional information needs (definition, background, and example) across abstract domains. The ratio of terms shows how many terms in the abstract domain are familiar, and require definitions, background, and examples.

ogy between domains (Lucy et al., 2022; Vilhena et al., 2014). Mathematics, which in prior work has been found to contain a large amount of discipline-specific terminology (Lucy et al., 2022), was one of the most familiar outside the social sciences. One possible reason was that annotators were generally from CS sub-domains that share overlap with Mathematics (e.g., ML, Computer Vision). Examples of terms and annotations are in Sup. Table B.7.

The same trend we observed for familiarity ratings held for definition and background information requests. However, annotators preferred a roughly constant rate of examples regardless of domain. Looking at common terms that annotators requested examples for, we see that generally these terms refer to a category rather than a single concept. For example, annotators requested examples for terms in the humanities like “mental operations” (5/11 annotators) and “mass communication technologies” (6/11 annotators) even when most annotators rated these terms as familiar (10/11 and 9/11, respectively).

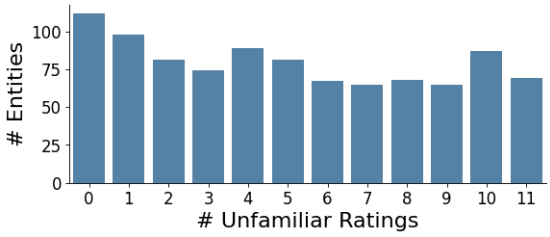


Figure 5.3: The number of terms rated as unfamiliar by annotators, broken down by the number of annotators. Generally there is a uniform number of ratings for each level of agreement.

Individual-specific variation There is substantial variation in term familiarity across annotators. As Figure 5.3 shows, terms vary widely in how many annotators rated them as familiar, with a slightly greater number of terms being rated as familiar by all annotators. For 15% of the terms, half the annotators disagreed (i.e., 5 or 6 deemed it familiar while the rest did not). The field most commonly found in this split was Mathematics (12% of terms).

Annotator backgrounds were associated with what terms they found familiar. Taking Mathematics again as an example, two annotators self-identifying as working in CS Theory and Neural Networks rated Mathematics terms as more familiar (mean=72%, std=4%) compared to other annotators (mean=44%, std=22%). There is similar variation for Linguistics. The three annotators identifying as NLP researchers rated an average of 64% of terms from Linguistics abstracts as familiar with little variation between them (std=6%). The remaining annotators rate an average of 57% Linguistics terms as familiar, with much greater variation (std=20%).

5.4 Prediction

Given our dataset and annotator backgrounds, we investigate the effectiveness of a set of features and methods for predicting individual term familiarity and additional information needs.

5.4.1 Features

Past work has explored using readability measures, frequency statistics, and embeddings to predict term familiarity at the population-level (e.g., for all lay readers) (Rakedzon et al., 2017; August et al., 2022d; Lucy et al., 2022). We adapt the following features for predicting individual-level familiarity:

- **Frequency**: The number of times a term appears in a researcher's publications $R = \{r_1, r_2, ...\}$.
- **Specificity**: The term's uniqueness to a corpus (Zhang et al., 2017), computed as the log probability ratio:
$$S_c(t) = log \frac{P_c(t)}{P_C(t)}$$
In our case, c corresponds to the target abstract A and C to the researcher's publications R.
- **Embedding similarity**: The minimum Euclidean distance between the target abstract A's embedding and any of the author's publications $R = \{r_1, r_2, ...\}$'s embeddings. We use embeddings from SPECTER 2.0 (Singh et al., 2022), a citation-based transformer model encoding semantic document similarity.

For each feature, we start by defining different granularities of a researcher's publications R, representing domain, subdomain, and individual-level information. The following data is extracted from the Semantic Scholar API (Kinney et al., 2023):

- **Domain**: 10K randomly sampled CS papers from 2015-2022.
- **Subdomain**: 10K randomly sampled papers from each annotator's self-defined CS subdomain from 2015-2022. Subdomains are defined manually based on venues associated with a given subdomain.
- **Individual**: All an individual's publications. If the individual's number of publications is less than the necessary number for training in §5.4.2, the remaining quantity is supplemented by a random selection from the cited references within those publications.

In addition to these granular features, we include the following general-purpose measures of readability and metadata:

- **Readability**: We use the Flesch-Kincaid (F-K) (Flesch, 2007) readability score and the GPT-2 perplexity score (Martinc et al., 2021). F-K score is computed at the passage level; all terms from an abstract are assigned the same F-K score.
- **Metadata**: We include the target paper domain, the year of the annotator's first published paper, total number of published papers, and the annotators average citation count for published papers.

5.4.2 Models

We explore two modeling approaches: supervised and prompt-based.

Lasso regression We adopt a logistic regression model with L1-regularization to integrate features across various levels of granularity and determine feature importance. A binary label determination is made using a threshold value of 0.5. We train one model per annotator. There were two training settings:

- *Individual model*: the model is trained using data from one annotator to predict ratings from the same annotator.
- *Mixed model*: the model is trained on ratings from all other annotators (i.e., leave-one-annotator-out testing). To maintain the same sample size of training data as the individual Lasso models, we randomly select the same number of training data points as was used to train an annotator's individual model.

Prompt-based LLMs We design prompts to predict binary term familiarity using the GPT-4 model from OpenAI in September 2023. To explore different strategies, we use:

- *Baseline*: providing only the term and abstract containing the term. This is essentially a zero-shot setting with no personalization information.
- *Metadata*: providing annotator metadata along with the baseline prompt.
- *Context-enhanced learning*: providing publications at either the annotator's domain, sub-domain, or individual level.
- *Few-shot learning*: providing examples of term-abstract-rating tuples from our labeled dataset. Ratings are drawn from the three levels of granularity: ratings from other annota-

tors with no overlap in subdomain (domain), from other annotators in the same subdomain (subdomain), and from the annotator (individual).

For context-enhanced learning and few-shot learning, we experiment with 1, 5, and 10 examples.

Oracle settings We also include two oracle classifiers based on social recommendations and collaborative filtering (Kang et al., 2022a; Guy et al., 2009; Goldberg et al., 1992):

- *Majority oracle*: the majority rating from all other annotators.
- *Nearest-neighbor oracle*: the annotator's rating who has the most similar ratings to the current annotator (i.e., having the highest agreement in ratings). Similarity of ratings is based on the training set. Nearest-neighbor pairings are listed in Sup. Table B.6.

5.4.3 Evaluation

We split the entities randomly into an 80/20 train/test set split for each annotator, with the test set containing 2200 ratings across 200 entities. All models are evaluated on the same test set per annotator, reporting F1 score, recall, and precision to measure classification performance. To identify critical features in a Lasso model, we count the features with non-zero value coefficients. A higher frequency denotes greater and more consistent influence of the feature on prediction.

5.5 Results

RQ1. How do supervised and prompt-based methods perform? In Table 5.2, we present a comparison of precision, recall and F1 scores across the highest performing predictive models. Both oracles outperform all other methods, pointing to the benefit of using ratings from similar annotators for predicting term familiarity. The nearest-neighbor oracle achieves roughly the same performance as the majority oracle while using one tenth the data (i.e., one annotator rather than 10), suggesting that collaborative-filtering approaches could be effective for personalized familiarity prediction without collecting data from many annotators.

Model	F1	Recall	Precision
Majority Baseline	$62.9_{\pm 1.4}$	$100.0_{\pm 0.0}$	$45.9_{\pm 1.5}$
Oracle			
Majority	$71.5_{\pm 1.7}$	$69.6_{\pm 2.1}$	$73.4_{\pm 2.2}$
Nearest-neighbor	$71.9_{\pm 1.7}$	$76.0_{\pm 2.1}$	$68.2_{\pm 2.1}$
Lasso			
Mixed	$56.9_{\pm 1.9}$	$59.5_{\pm 2.3}$	$54.6_{\pm 2.3}$
Individual	$60.6_{\pm 2.0}$	$55.3_{\pm 2.3}$	$67.2_{\pm 2.3}$
GPT			
Baseline	$63.1_{\pm 1.5}$	$100.0_{\pm 0.0}$	$46.1_{\pm 1.6}$
Metadata	$64.0_{\pm 1.5}$	$94.4_{\pm 1.0}$	$48.4_{\pm 1.6}$
Context-enhanced	$64.2_{\pm 1.5}$	$98.7_{\pm 0.5}$	$47.6_{\pm 1.6}$
Few-shot	$62.8_{\pm 1.5}$	$99.6_{\pm 0.3}$	$45.8_{\pm 1.6}$

Table 5.2: Mean model performance ($\pm$std) on term familiarity prediction in the test set. Standard deviation is estimated by bootstrapping with 1,000 resamples and each size of 1,000. Context-enhanced and few-shot learning are prompted with individual-level data. Metadata model are prompted with all metadata.

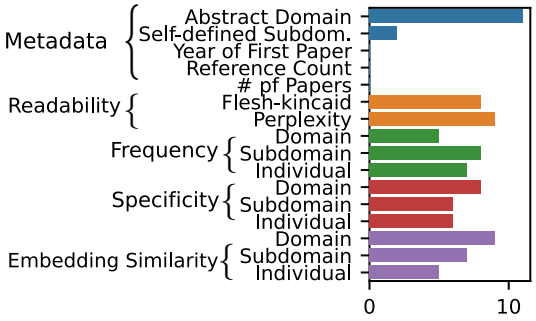


Figure 5.4: Frequency of non-zero coefficients in individual Lasso models across researchers. The Lasso penalty minimizes less critical coefficients to zero. Features with higher frequencies of non-zero values in individual models are consistently identified as important.

While other models slightly outperform the majority baseline, we generally see that the task of personalized jargon prediction is difficult. Among the models, the individual Lasso model shows superior performance compared to the mixed Lasso. Despite being provided less data, GPT-4 approaches slightly outperform the individual Lasso model.

Notably, the Lasso models achieve substantially higher precision (Table 5.2) than either GPT-4 or the majority baseline. Explaining terms a reader already knows (i.e., a false positive) distracted readers in our formative study (§5.2.1), making Lasso models that minimize false positives a promising alternative to prompt-based methods.

RQ2. What features and granularity level influence performance? Figure 5.4 reveals the non-zero coefficients of the individual Lasso models. The domain of the target abstract is consistently identified as a significant feature by all models. Word specificity and embedding similarity at the domain level also underscore the relevance of a researcher's domain in their familiarity with jargon, aligning with previous population-level familiarity research (Li et al., 2020a). However, the importance of individual-level features such as frequency, specificity, and embedding similarity also emerge, highlighting the dual influence of both domain-specific and individual-specific factors.

Model	F1	Recall	Precision
Metadata			
Self-defined Subdom.	$65.2_{\pm 1.5}$	$95.3_{\pm 1.0}$	$49.6_{\pm 1.7}$
# of Papers	$60.0_{\pm 1.7}$	$77.5_{\pm 1.9}$	$48.9_{\pm 1.9}$
Reference Count	$60.5_{\pm 1.6}$	$79.5_{\pm 1.8}$	$48.8_{\pm 1.7}$
Year of First Paper	$63.5_{\pm 1.5}$	$95.9_{\pm 0.9}$	$47.5_{\pm 1.6}$
Abstract Domain	$62.9_{\pm 1.5}$	$99.7_{\pm 0.3}$	$46.0_{\pm 1.6}$
All Metadata	$64.0_{\pm 1.5}$	$94.4_{\pm 1.0}$	$48.4_{\pm 1.6}$
Granularity			
Domain	$63.0_{\pm 1.5}$	$99.5_{\pm 0.3}$	$46.1_{\pm 1.6}$
Subdomain	$63.4_{\pm 1.5}$	$99.3_{\pm 0.4}$	$46.6_{\pm 1.6}$
Individual	$64.2_{\pm 1.5}$	$98.7_{\pm 0.5}$	$47.6_{\pm 1.6}$
Example Number			
n=1	$63.6_{\pm 1.5}$	$99.1_{\pm 0.4}$	$46.9_{\pm 1.6}$
n=5	$64.2_{\pm 1.5}$	$98.7_{\pm 0.5}$	$47.6_{\pm 1.6}$
n=10	$64.0_{\pm 1.5}$	$98.7_{\pm 0.5}$	$47.4_{\pm 1.6}$

Table 5.3: Mean GPT-4 model performance ($\pm$std) on term familiarity prediction in the test set. Context-enhanced learning with individual level data is used for granularity and example number. Underlined models are reported in Table 5.2.

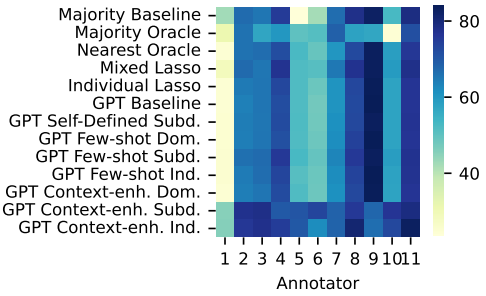


Figure 5.5: F1 score of models in familiarity prediction across annotators.

Table 5.3 details GPT-4 performance when including different metadata and publication granularities. Prompting with researcher self-defined subdomain seems to be the most effective strategy, suggesting that subdomain information (e.g., that a researcher is in NLP) is more useful than a researcher's broad field (e.g., CS).

RQ3. How do models perform across annotators? Model performance varies substantially for different annotators. Table 5.4 reports standard deviation of model performance calculated across annotators among top performing models and Figure 5.5 plots the F1 scores across annotators. We see ranges above 15 F1 points, well beyond the differences in performance we see across models. The high levels of variation measured across annotators indicate that models work much better for certain annotators over others. Looking at Figure 5.5, we also see that the best performing model is different for different annotators. For example, for Annotator 8, the *context-enhanced GPT-4* at the individual level performed best, while *individual lasso* performed better for Annotator 9.

RQ4. What errors do the models make? Figure 5.6 plots the rate of correct predictions for the two best performing model variants: GPT with subdomain metadata and the individual Lasso regression models. We see that the GPT method over-predicts unfamiliarity: incorrect predictions are usually for terms that most annotators were familiar with. Looking at these entities, most have common meanings that a reader could guess given context (e.g., "coffee production" or "food supply chains"). In contrast, the Lasso model generally

Model	F1	Recall	Precision
Oracle			
Majority	±9.4	±13.2	±18.2
Nearest-neighbor	±10.4	±9.4	±16.8
Lasso			
Individual	±18.7	±25.8	±10.4
GPT			
Metadata - all	±15.6	±4.1	±17.5
Metadata - subdomain	±15.8	±4.1	±17.5
Context-enhanced	±16.4	±1.8	±16.8

Table 5.4: Standard deviation of the highest performing models' scores across annotators.

performs better for highly familiar terms, but suffers for terms that are highly unfamiliar. Supporting this difference, we generally see that the individual lasso models perform better for annotators who were familiar with more entities, while the prompt-based methods performed better for annotators who rated more entities as unfamiliar.

RQ5. How does individual information contribute to additional information prediction? We investigate models that performed well on the familiarity prediction task to predict users' additional information needs. Results in Table 5.5 demonstrate that incorporating individual features into Lasso or prompt-based models do not substantially improve performance on this challenging prediction task. Predicting granular user information needs remains difficult without more tailored modeling and data to detect individual variations.

5.6 Related Work

Interdisciplinary communication Interdisciplinary research integrates knowledge from multiple disciplines to address a shared question (Daniel et al., 2022). Choi and Pak (2007) surveyed interdisciplinary researchers in the health sciences, finding that a mismatch in

Model	Def.	Bg.	Ex.
Majority Baseline	$44.4_{\pm 1.8}$	$37.2_{\pm 1.7}$	$31.5_{\pm 1.8}$
Oracle			
Majority	$50.7_{\pm 2.8}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Nearest-neighbor	$54.3_{\pm 2.4}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Lasso			
Mixed	$13.7_{\pm 3.6}$	$15.0_{\pm 5.1}$	$0.0_{\pm 0.0}$
Individual	$56.4_{\pm 2.8}$	$48.1_{\pm 3.3}$	$28.7_{\pm 3.9}$
GPT			
Baseline	$47.7_{\pm 1.8}$	$40.0_{\pm 1.9}$	$31.9_{\pm 1.8}$
Self-defined Subfield	$48.4_{\pm 1.9}$	$39.3_{\pm 1.9}$	$32.4_{\pm 1.8}$
Context-enhanced	$47.5_{\pm 1.8}$	$38.6_{\pm 1.9}$	$32.5_{\pm 1.8}$

Table 5.5: Mean F1 score (±std) on **additional information needs** (definition, background, and example) prediction in the test set. Recall and precision are in Sup. Table B.3, B.4, and B.5.

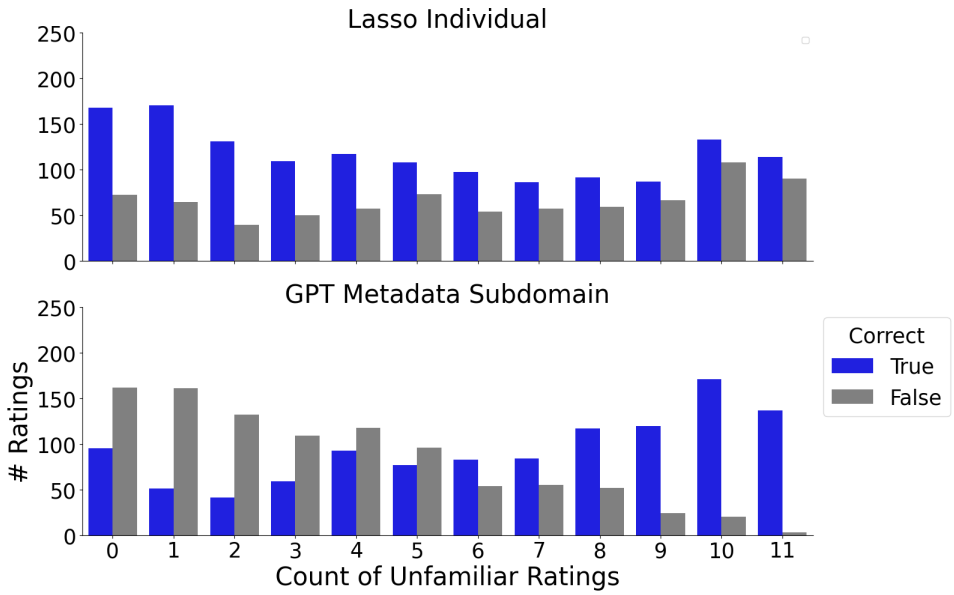


Figure 5.6: Error counts for two best performing model variants; entities are binned by their total count of unfamiliarity ratings summed over annotators. Generally, the GPT-based method over-predicts unfamiliarity.

terminology complicates efforts in communicating between disciplines. Lucy et al. (2022) found that papers that used more discipline-specific terminology (i.e., jargon) had fewer citations across disciplines, and Martínez and Mammola (2021) found that papers that use more jargon are generally cited less.

Scientific text simplification Detecting scholarly jargon is commonly done using corpus-based approaches (Tanaka-Ishii and Terada, 2011). For example, Gardner and Davies (2013) identified scholarly jargon in English by studying the frequency of words within scientific papers compared to a background corpus of general English writing. Similar methods have identified jargon in specific fields of science, including medical studies (August et al., 2022d), and computer science papers (Salatino et al., 2018). Gooding and Tragut (2022) found that training models at the individual level improves general English complex word identification, and Lin et al. (2012) found that using social media posts written by an indi-

vidual can help predict word familiarity. Murthy et al. (2021) generated alternate definitions of scientific terms to better align with a scientist's knowledge. Work has also explored interactive systems to augment scientific abstracts (Fok et al., 2023) and provide term definitions (Head et al., 2020; August et al., 2022d).

In contrast to prior work, we focus on predicting familiarity of scholarly jargon. Our focus on scientists provides unique opportunities to model individual knowledge. Scientists develop deep knowledge of their field by reading and writing scientific papers. Models that can achieve accurate, individualized predictions for scientists could greatly improve interactive systems by focusing aids (e.g., definitions, additional information) on only the unfamiliar words for an individual reader (Ridder, 2002; Head et al., 2020).

5.7 Discussion

This paper introduces the novel task of personalized scholarly jargon detection. We collect a dataset of over 10K term familiarity and information need annotations from 11 CS researchers. Our dataset reveals significant variation in term familiarity based on annotators' background (§5.3.4). When predicting familiarity, subdomain information and sampled paper abstracts written by the author can improve prediction. This is particularly evident in prompt-based methodologies, where the use of paper abstracts was more beneficial than term familiarity labels from that researcher. This might be because researcher abstracts/metadata provide more generalizable information about the annotator than individual-specific term familiarity labels.

The high levels of variation, of both familiarity labels (§5.3.4) and model performance (§5.5) across annotators, indicate that certain models work much better for certain annotators than others. Personalized scientific jargon identification is a difficult task, with no clear best modeling strategy, and performance is dependent on the individual researcher being modeled. This opens up new research questions and avenues for future work. For example, we see that collaborative oracles saw noticeable performance boosts, suggesting that taking ratings from similar researchers might improve over general-purpose jargon identification methods. Subdomain information was also helpful for prediction, suggesting that modeling

familiarity at the subdomain level can bring benefits beyond current general jargon identification techniques. Furthermore, for researchers with a small publication history (and therefore a small amount of unlabeled individual-level data), subdomain data can provide valuable information for predicting jargon familiarity.

Our methods provide an exciting first step to support researchers in reading and communicating outside their domain (Wudarczyk et al., 2021). Combined with text simplification techniques (Srikanth and Li, 2020b), models that can predict an individual researcher's familiarity could assist researchers by rewriting an abstract tuned to particular research audiences, a capability that our formative study findings highlighted as a need for researchers (§5.2).

5.8 Conclusion

This paper introduces the novel task of scientific jargon detection for the individual researcher. We collect and release a dataset of over 10K term familiarity annotations from computer science researchers and investigate supervised and prompt-based methods to predict term familiarity. We find that leveraging a researcher's publication history, self-reported subdomain, and general domain information can improve term familiarity prediction. Our results provide insight on integrating an individual's knowledge into scientific jargon detection.

5.9 Limitations

We focus on CS researchers for selecting annotators and abstracts in other fields. This might limit our ability to generalize to other domains or researchers. Publication venues (e.g., journals or conferences) and norms vary widely across fields, which might complicate how research publications can be used to model term familiarity. Because of the cost of collecting annotations, our dataset is also relatively small, which might further limit its generalizability. 11 CS researchers is not representative of all of CS nor does it adequately cover each subfield (e.g., only one annotator self-identified as a CS theory researcher). Our goal with the dataset and analysis is to show the potential of modeling individual term familiairty and

information needs. We are excited about future work expanding this goal into new domains. Therefore, we published the procedures and questionnaire used in our data collection to encourage future research on personalized jargon detection for scientists in additional domains. Details can be found at: https://github.com/talaugust/PersonalizedJargon.

5.10 Ethics Statement

Some of the methods in the paper include personal data (e.g., publication record, labeled terms), which might pose a privacy risk for some researchers. Systems identifying term familiarity and information needs must keep any personal data stored locally and allow researchers to remove or view their data at any time. Focusing on CS researchers as a first step for predicting term familiarity might also allow CS researchers to more effectively read outside their discipline, but not researchers in other domains reading within CS. While encouraging interdisciplinary reading can improve two-way communication, it is also important to consider the voices of researchers in other domains.

Chapter 6

CONCLUSION

6.1 Discussion

This thesis presents novel contributions to improving the accessibility of health information using LLMs, spanning dataset construction, algorithm development, and evaluation methods.

In this work, I proposed the task of plain summary generation and identified the unique transformations involved in translating biomedical scientific language into plain language. I also demonstrated the utility of encoder-decoder models in generating PLSs and highlighted the significance of domain-specific continual pre-training for enhancing the quality of the generated summaries for the first time.

The major material contribution is the CELLS dataset, which is, to the best of my knowledge, the largest lay language generation dataset developed to date. The derived dataset for background explanation also serves as the first explanation generation benchmark. These datasets have broad potential applications in biomedical NLP, providing a foundation for developing tools to generate PLSs of scientific literature, as well as supporting research on sentence-level and paragraph-level simplification, open-question answering, and information retrieval.

During the course of this work, I identified subtasks specific to PLS, that distinguish this task from the summarization tasks that language models had been applied to previously. The most challenging of these subtasks involves the generation of *background explanations*, as this requires drawing on information from beyond that provided in the scientific abstract to be simplified. To address this particular challenge of PLS, I explored methods for generating retrieval-augmented lay language, augmenting state-of-the-art text generation models with information retrieval from sources such as the UMLS, Wikipedia, and embeddings of explanations from Wikipedia documents. My findings indicate that RALL models improve

summary quality and simplicity while maintaining factual correctness, suggesting that general knowledge from Wikipedia is a valuable source for background explanation.

Meaningful and accurate quality measurement is crucial for ensuring proper system performance and efficiency, as well as attracting further research by enabling reliable assessment. While prior studies relied on readability scores or costly human evaluation, these approaches have clear limitations. To address this, I assessed how well existing metrics capture the requirements of PLS, defining four key dimensions of PLS quality: informativeness, simplification, coherence, and faithfulness. Using a set of perturbations to probe metric sensitivity, I developed the APPLS meta-evaluation testbed and used it to analyze 15 metrics. My findings show that while established metrics demonstrate mixed sensitivities to perturbations associated with informativeness, coherence, and faithfulness, all tested metrics display a lack of sensitivity towards simplification perturbations.

In response, I introduced POMME, a new metric that evaluates text simplicity by calculating normalized perplexity differences between language models trained on in-domain and out-of-domain text. My results demonstrate POMME's effectiveness at capturing differences in text simplicity through extensive experiments on APPLS and other text simplification datasets.

To address the necessity of personalization in PLS, I introduced the novel task of scientific jargon personalization for individual researchers. I collected and released a dataset of over 10,000 term familiarity annotations from computer science researchers and investigate supervised and prompt-based methods to predict term familiarity. My results show that leveraging a researcher's publication history, self-reported subdomain, and general domain information can improve term familiarity prediction, providing insight on integrating an individual's knowledge into scientific jargon detection.

In summary, the main contributions of this thesis include the introduction of the novel task of automated generation of plain language summaries of scientific literature, the construction of the CELLS dataset, the meta-evaluation of existing evaluation metrics on this task, the introduction of POMME for assessing text simplicity, and the exploration of personalization in text generation. Through this work, I aim to bridge the gaps in health literacy, break down barriers to understanding, and ultimately improve health communication for all.

6.2 Limitations and Future Work

While this work has made progress in dataset development, evaluation, and algorithms for improving health information accessibility, there are several limitations to address in future research.

First, the current plain language text is written by experts and assessed by a subset of the population, which may introduce bias. Future work should focus on expanding the reach of the system and evaluating it among underserved groups, such as youth, people with disabilities, the elderly, and individuals from diverse socioeconomic backgrounds.

Second, the current approach to capturing researchers' knowledge in the interdisciplinary communication project is limited to the papers they have published. To improve the accuracy and personalization of the system, future research should explore additional features, such as citations, papers read, and expressed preferences.

Finally, while the current research focuses on making biomedical literature more accessible, patients' information needs in real-life healthcare situations extend beyond scientific articles. Future work should address the challenge of helping patients understand their medical test results, imaging reports, and after-visit summaries by integrating patient-centered communication systems directly into Electronic Health Record platforms.

Moreover, several challenges persist in the field of PLS generation. Hallucinations, where the model generates information not present in the source text, remain a concern, and additional research is necessary to characterize and mitigate this issue. The evaluation of generated summaries, particularly in terms of simplicity or understandability, is still an open question, as these aspects are highly personalized and difficult to determine for specific subpopulations or individuals. Personalized PLS based on individual reader preferences and background knowledge is another area that warrants exploration to improve the accessibility and effectiveness of the generated summaries. However, obtaining such data is challenging, especially in the healthcare domain due to privacy concerns. Further research is needed to investigate deidentification techniques and identify features that reflect understandibility. Future work should also concentrate on exploring the application of more advanced architectures and the incorporation of domain-specific knowledge to enhance the quality and

reliability of PLS generation systems.

6.3 Applications

The application of personalized plain language generation systems would benefit healthcare consumers, researchers, and healthcare providers. Healthcare providers can be relieved of the burden of writing clinical notes and after-visit summaries, instead reviewing and signing automatically generated plain language summary notes. Researchers can use PLS systems to identify relevant research papers, communicate more effectively across disciplines, and facilitate novel research ideas and collaborations. Healthcare consumers, both patients and caregivers, can access the most up-to-date health information in an understandable format, informing their decision-making processes. The ultimate goal is to create a friendly, approachable, and intelligent virtual health assistant that empowers patients to take control of their health journey. I envision a future where every patient has access to a personalized ChatGPT-like interface containing all their relevant health information, capable of breaking down complex medical jargon, providing easy-to-understand explanations, and offering gentle guidance on next steps. This would greatly enhance the healthcare experience, providing patients with a knowledgeable, always-available resource to help navigate the often-confusing world of healthcare.

BIBLIOGRAPHY

Action, P.L., Network, I., 2011. Federal Plain Language Guidelines. CreateSpace Independent.

Aho, A.V., Ullman, J.D., 1972. The Theory of Parsing, Translation and Compiling. volume 1. Prentice-Hall, Englewood Cliffs, NJ.

Alambo, A., Banerjee, T., Thirunarayan, K., Raymer, M., 2022. Entity-driven fact-aware abstractive summarization of biomedical literature. arXiv preprint arXiv:2203.15959 .

Alfirevic, Z., Stampalija, T., Dowswell, T., 2017. Fetal and umbilical doppler ultrasound in high-risk pregnancies. Cochrane database of systematic reviews .

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E.D., Gutierrez, J.B., Kochut, K., 2017. Text summarization techniques: a brief survey. arXiv preprint arXiv:1707.02268 .

Altmami, N.I., Menai, M.E.B., 2020. Automatic summarization of scientific articles: A survey. Journal of King Saud University - Computer and Information Sciences .

Alva-Manchego, F., Martin, L., Scarton, C., Specia, L., 2019. EASSE: Easier automatic sentence simplification evaluation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations, Association for Computational Linguistics, Hong Kong, China. pp. 49–54. URL: https://aclanthology.org/D19-3009, doi:10.18653/v1/D19-3009.

American Psychological Association, 1983. Publications Manual. American Psychological Association, Washington, DC.

Amith, M.T., Cohen, T.A., Cunningham, R., Savas, L.S., Smith, N., Cuccaro, P.M., Gabay, E.K., Boom, J.A., Schvaneveldt, R.W., Tao, C., 2020. Mining hpv vaccine knowledge

structures of young adults from reddit using distributional semantics and pathfinder networks. Cancer Control : Journal of the Moffitt Cancer Center 27. URL: https://api.semanticscholar.org/CorpusID:210086429.

Amith, M.T., Cunningham, R., Savas, L.S., Boom, J.A., Schvaneveldt, R.W., Tao, C., Cohen, T.A., 2017. Using pathfinder networks to discover alignment between expert and consumer conceptual knowledge from online vaccine content. Journal of biomedical informatics 74, 33–45. URL: https://api.semanticscholar.org/CorpusID:7095540.

Ando, R.K., Zhang, T., 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research 6, 1817–1853. URL: https://www.jmlr.org/papers/volume6/ando05a/ando05a.pdf.

Andrew, G., Gao, J., 2007. Scalable training of L_1-regularized log-linear models, in: Proceedings of the 24th International Conference on Machine Learning, pp. 33–40. URL: https://dl.acm.org/doi/abs/10.1145/1273496.1273501.

Artstein, R., Poesio, M., 2008. Inter-coder agreement for computational linguistics. Computational linguistics 34, 555–596.

Attal, K., Ondov, B., Demner-Fushman, D., 2023. A dataset for plain language adaptation of biomedical abstracts. Scientific Data 10, 8.

August, T., Card, D., Hsieh, G., Smith, N.A., Reinecke, K., 2020. Explain like i am a scientist: The linguistic barriers of entry to r/science, in: Proceedings of the 2020 CHI conference on human factors in computing systems, pp. 1–12.

August, T., Reinecke, K., Smith, N.A., 2022a. Generating scientific definitions with controllable complexity, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland. pp. 8298–8317. URL: https://aclanthology.org/2022.acl-long.569, doi:10.18653/v1/2022.acl-long.569.

August, T., Reinecke, K., Smith, N.A., 2022b. Generating scientific definitions with controllable complexity, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8298–8317.

August, T., Wang, L.L., Bragg, J., Hearst, M.A., Head, A., Lo, K., 2022c. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. arXiv preprint arXiv:2203.00130 .

August, T., Wang, L.L., Bragg, J., Hearst, M.A., Head, A., Lo, K., 2022d. Paper plain: Making medical research papers approachable to healthcare consumers with natural language processing. ACM Transactions on Computer-Human Interaction 30, 1 – 38. URL: https://api.semanticscholar.org/CorpusID:247187606.

Banerjee, S., Lavie, A., 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, pp. 65–72.

Barnett, A.P.A., Doubleday, Z., 2020. The growth of acronyms in the scientific literature. eLife 9. URL: https://api.semanticscholar.org/CorpusID:222631123.

Barzilay, R., Elhadad, N., 2002. Inferring strategies for sentence ordering in multidocument news summarization. Journal of Artificial Intelligence Research 17, 35–55.

Barzilay, R., Lapata, M., 2008. Modeling local coherence: An entity-based approach. Computational Linguistics 34, 1–34.

Basu, C., Vasu, R., Yasunaga, M., Yang, Q., 2023. Med-easi: Finely annotated dataset and models for controllable simplification of medical texts. ArXiv abs/2302.09155. URL: https://api.semanticscholar.org/CorpusID:257038377.

Baumgartner, W.A., Lu, Z., Johnson, H.L., Caporaso, J.G., Paquette, J., Lindemann, A., White, E.K., Medvedeva, O., Cohen, K.B., Hunter, L., 2008. Concept recognition for extracting protein interaction relations from biomedical text. Genome biology 9, S9.

Beck, I.L., McKeown, M.G., Sinatra, G.M., Loxterman, J.A., 1991. Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. Reading research quarterly , 251–276.

Belz, A., Reiter, E., 2006. Comparing automatic and human evaluation of nlg systems, in: 11th conference of the european chapter of the association for computational linguistics, pp. 313–320.

Van den Bercken, L., Sips, R.J., Lofi, C., 2019. Evaluating neural text simplification in the medical domain, in: The World Wide Web Conference, pp. 3286–3292.

Berghella, V., Saccone, G., 2019. Cervical assessment by ultrasound for preventing preterm delivery. Cochrane database of systematic reviews .

Bin Naeem, S., Kamel Boulos, M.N., 2021. Covid-19 misinformation online and health literacy: a brief overview. International journal of environmental research and public health 18, 8091.

Bingel, J., Paetzold, G., Søgaard, A., 2018. Lexi: A tool for adaptive, personalized text simplification, in: Proceedings of the 27th International Conference on Computational Linguistics, pp. 245–258.

Bird, S., Dale, R., Dorr, B.J., Gibson, B.R., Joseph, M.T., Kan, M.Y., Lee, D., Powley, B., Radev, D.R., Tan, Y.F., et al., 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics., in: LREC, pp. 00–08.

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

Bolton, E., Hall, D., Yasunaga, M., Lee, T., Manning, C., Liang, P., 2023. Pubmedgpt 2.7b. URL: https://crfm.stanford.edu/2022/12/15/pubmedgpt.html.

Britton, B.K., Gülgöz, S., 1991. Using kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Journal of educational Psychology 83, 329.

Brok, J., Gluud, L.L., Gluud, C., 2010. Ribavirin plus interferon versus interferon for chronic hepatitis c. Cochrane database of systematic reviews .

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020a. Language models are few-shot learners. ArXiv abs/2005.14165. URL: https://api.semanticscholar.org/CorpusID:218971783.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020b. Language models are few-shot learners. ArXiv abs/2005.14165.

Bui, D.D.A., Del Fiol, G., Hurdle, J.F., Jonnalagadda, S., 2016. Extractive text summarization system to aid data extraction from full text in systematic review development. Journal of biomedical informatics 64, 265–272.

Cachola, I., Lo, K., Cohan, A., Weld, D., 2020a. Tldr: Extreme summarization of scientific documents. Findings of EMNLP .

Cachola, I., Lo, K., Cohan, A., Weld, D.S., 2020b. Tldr: Extreme summarization of scientific documents. ArXiv abs/2004.15011.

Cai, D., Wang, Y., Bi, W., Tu, Z., Liu, X., Lam, W., Shi, S., 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1219–1228.

Cai, X., Liu, S., Han, J., Yang, L., Liu, Z., Liu, T., 2021. Chestxraybert: A pretrained language model for chest radiology report summarization. IEEE Transactions on Multimedia .

Cai, X., Liu, S., Yang, L., Lu, Y., Zhao, J., Shen, D., Liu, T., 2022. Covidsum: A linguistically enriched scibert-based summarization model for covid-19 scientific papers. Journal of Biomedical Informatics 127, 103999.

Cao, Q., Xiong, D., 2018. Encoding gated translation memory into neural machine translation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3042–3047.

Cao, Y., Shui, R., Pan, L., Kan, M.Y., Liu, Z., Chua, T.S., 2020. Expertise style transfer: A new task towards better communication between experts and laymen. arXiv preprint arXiv:2005.00701 .

Carlsson, F., Öhman, J., Liu, F., Verlinden, S., Nivre, J., Sahlgren, M., 2022. Fine-grained controllable text generation using non-residual prompting, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6837–6857.

Chan, J., Chang, J.C., Hope, T., Shahaf, D., Kittur, A., 2018. Solvent: A mixed initiative system for finding analogies between research papers. Proc. ACM Hum. Comput. Interact. 2, 31:1–31:21. URL: https://api.semanticscholar.org/CorpusID:53236685.

Chandra, A.K., Kozen, D.C., Stockmeyer, L.J., 1981. Alternation. Journal of the Association for Computing Machinery 28, 114–133. doi:10.1145/322234.322243.

Chandrasekaran, M.K., Feigenblat, G., Hovy, E., Ravichander, A., Shmueli-Scheuer, M., de Waard, A., 2020. Overview and insights from the shared tasks at scholarly document processing 2020: Cl-scisumm, laysumm and longsumm, in: Proceedings of the First Workshop on Scholarly Document Processing, pp. 214–224.

Chen, Y., Eger, S., 2022. Menli: Robust evaluation metrics from natural language inference. arXiv preprint arXiv:2208.07316 .

Cheng, J., Lapata, M., 2016a. Neural summarization by extracting sentences and words. ArXiv abs/1603.07252.

Cheng, J., Lapata, M., 2016b. Neural summarization by extracting sentences and words. arXiv preprint arXiv:1603.07252 .

Chintagunta, B., Katariya, N., Amatriain, X., Kannan, A., 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization, in: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pp. 66–76.

Choi, B.C.K., Pak, A.W.P., 2007. Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. promotors, barriers, and strategies of enhancement. Clinical and investigative medicine. Medecine clinique et experimentale 30 6, E224–32. URL: https://api.semanticscholar.org/CorpusID:211765.

Cirocchi, R., Trastulli, S., Abraha, I., Vettoretto, N., Boselli, C., Montedori, A., Parisi, A., Noya, G., Platell, C., 2012. Non-resection versus resection for an asymptomatic primary tumour in patients with unresectable stage iv colorectal cancer. Cochrane Database of Systematic Reviews .

Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N., 2018a. A discourse-aware attention model for abstractive summarization of long documents, in: NAACL-HLT, pp. 615–621.

Cohan, A., Dernoncourt, F., Kim, D.S., Bui, T., Kim, S., Chang, W., Goharian, N., 2018b. A discourse-aware attention model for abstractive summarization of long documents, in: Proceedings of NAACL-HLT, pp. 615–621.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S., 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers, in: ACL, pp. 2270–2282.

Coleman, M., Liau, T.L., 1975. A computer readability formula designed for machine scoring. Journal of Applied Psychology 60, 283.

Cook, D.A., Beckman, T.J., Bordage, G., 2007. A systematic review of titles and abstracts of experimental studies in medical education: many informative elements missing. Medical education 41 11, 1074–81. URL: https://api.semanticscholar.org/CorpusID:3592666.

Cooley, J.W., Tukey, J.W., 1965. An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation 19, 297–301. URL: https://www.ams.org/journals/mcom/1965-19-090/S0025-5718-1965-0178586-1/S0025-5718-1965-0178586-1.pdf.

Crossley, S.A., Yang, H.S., McNamara, D.S., 2014. What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. Reading in a Foreign Language 26, 92–113.

Daniel, K.L., McConnell, M., Schuchardt, A., Peffer, M.E., 2022. Challenges facing interdisciplinary researchers: Findings from a professional development workshop. PLoS ONE 17. URL: https://api.semanticscholar.org/CorpusID:248262202.

Das, D., Martins, A., 2007. A Survey on Automatic Text Summarization. Technical Report. Carnegie Mellon University. https://www.cs.cmu.edu/~afm/Home_files/Das_Martins_survey_summarization.pdf.

De Belder, J., Moens, M.F., 2010. Text simplification for children, in: Prroceedings of the SIGIR workshop on accessible search systems, ACM; New York. pp. 19–26.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American society for information science 41, 391–407.

Dell, N., Vaidyanathan, V., Medhi, I., Cutrell, E., Thies, W., 2012. " yours is better!" participant response bias in hci, in: Proceedings of the sigchi conference on human factors in computing systems, pp. 1321–1330.

Desmarais, M.C., Baker, R., 2012. A review of recent advances in learner and skill modeling in intelligent learning environments. User Modeling and User-Adapted Interaction 22, 9–38. URL: https://api.semanticscholar.org/CorpusID:14826104.

Deutsch, D., Bedrax-Weiss, T., Roth, D., 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. Transactions of the Association for Computational Linguistics 9, 774–789.

Deutsch, D., Roth, D., 2022. Benchmarking answer verification methods for question answering-based summarization evaluation metrics. arXiv preprint arXiv:2204.10206 .

Devaraj, A., Marshall, I., Wallace, B.C., Li, J.J., 2021. Paragraph-level simplification of medical texts, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4972–4984.

Devaraj, A., Sheffield, W., Wallace, B.C., Li, J.J., 2022a. Evaluating factuality in text simplification, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, NIH Public Access. p. 7331.

Devaraj, A., Sheffield, W., Wallace, B.C., Li, J.J., 2022b. Evaluating factuality in text simplification. Proceedings of the conference. Association for Computational Linguistics. Meeting 2022, 7331–7345.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, pp. 4171–7186.

DS, M., 2001. Reading both high-coherence and low-coherence texts: effects of text sequence and prior knowledge. Can J Exp Psychol 55, 51–62.

Durmus, E., He, H., Diab, M., 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. arXiv preprint arXiv:2005.03754 .

Erkan, G., Radev, D.R., 2004. Lexrank: Graph-based centrality as salience in text summarization. Journal of Artificial Intelligence Research .

Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R., Radev, D., 2020. Summeval: Re-evaluating summarization evaluation. arXiv preprint arXiv:2007.12626 .

Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R., Radev, D., 2021. Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics 9, 391–409.

Feblowitz, J.C., Wright, A., Singh, H., Samal, L., Sittig, D.F., 2011. Summarization of clinical information: a conceptual model. Journal of biomedical informatics 44, 688–699.

Flesch, R., 1948. A new readability yardstick. Journal of applied psychology 32, 221.

Flesch, R., 2007. Flesch-kincaid readability test. Retrieved October 26, 2007.

Fok, R., Chang, J.C., August, T., Zhang, A.X., Weld, D.S., 2023. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries. ArXiv abs/2310.07581. URL: https://api.semanticscholar.org/CorpusID:263835343.

Friedman, C., Kra, P., Rzhetsky, A., 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. Journal of biomedical informatics 35, 222–235.

Gabriel, S., Celikyilmaz, A., Jha, R., Choi, Y., Gao, J., 2020. Go figure: A meta evaluation of factuality in summarization. arXiv preprint arXiv:2010.12834 .

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., Wan, X., 2023. Human-like summarization evaluation with chatgpt. arXiv preprint arXiv:2304.02554 .

Gardner, D., Davies, M., 2013. A new academic vocabulary list. Applied Linguistics 35, 305–327.

Gehlbach, H., Barge, S., 2012. Anchoring and adjusting in questionnaire responses. Basic and Applied Social Psychology 34, 417–433.

Givchi, A., Ramezani, R., Baraani, A., 2022. Graph-based abstractive biomedical text summarization. Journal of Biomedical Informatics , 104099.

Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Saez, G.G., Viviani, M., Xu, C., 2020. Overview of the clef ehealth evaluation lab

2020, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer. pp. 255–271.

Goldberg, D., Nichols, D.A., Oki, B.M., Terry, D.B., 1992. Using collaborative filtering to weave an information tapestry. Commun. ACM 35, 61–70. URL: https://api.semanticscholar.org/CorpusID:1591394.

Goldenberg, J.Z., Yap, C., Lytvyn, L., Lo, C.K.F., Beardsley, J., Mertz, D., Johnston, B.C., 2017. Probiotics for the prevention of clostridium difficile-associated diarrhea in adults and children. Cochrane Database of Systematic Reviews .

Goldsack, T., Luo, Z., Xie, Q., Scarton, C., Shardlow, M., Ananiadou, S., Lin, C., 2023a. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles, in: Proceedings of the 22st Workshop on Biomedical Language Processing, Toronto, Canada. Association for Computational Linguistics, pp. 468–477.

Goldsack, T., Zhang, Z., Lin, C., Scarton, C., 2022. Making science simple: corpora for the lay summarisation of scientific literature. arXiv preprint arXiv:2210.09932 .

Goldsack, T., Zhang, Z., Lin, C., Scarton, C., 2023b. Domain-driven and discourse-guided scientific summarisation, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I, Springer. pp. 361–376.

Gooding, S., Tragut, M., 2022. One size does not fit all: The case for personalised word complexity models. ArXiv abs/2205.02564. URL: https://api.semanticscholar.org/CorpusID:248524891.

Goyal, T., Durrett, G., 2021. Annotating and modeling fine-grained factuality in summarization. ArXiv abs/2104.04302.

Goyal, T., Li, J.J., Durrett, G., 2022. News summarization and evaluation in the era of gpt-3. ArXiv abs/2209.12356. URL: https://api.semanticscholar.org/CorpusID:252532176.

Grootendorst, M., 2020. Keybert: Minimal keyword extraction with bert. URL: https://doi.org/10.5281/zenodo.4461265, doi:10.5281/zenodo.4461265.

Gunning, R., et al., 1952. Technique of clear writing. McGraw-Hill.

Guo, Y., August, T., Leroy, G., Cohen, T., Wang, L.L., 2023a. Appls: A meta-evaluation testbed for plain language summarization. arXiv preprint arXiv:2305.14341 .

Guo, Y., August, T., Leroy, G., Cohen, T.A., Wang, L.L., 2023b. Appls: A meta-evaluation testbed for plain language summarization. ArXiv abs/2305.14341. URL: https://api.semanticscholar.org/CorpusID:258841161.

Guo, Y., Chang, J.C., Antoniak, M., Bransom, E., Cohen, T., Wang, L.L., August, T., 2024. Personalized jargon identification for enhanced interdisciplinary communication, in: Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 0–8. URL: https://api.semanticscholar.org/CorpusID:265220712.

Guo, Y., Qiu, W., Leroy, G., Wang, S., Cohen, T., 2022a. Cells: A parallel corpus for biomedical lay language generation. arXiv preprint arXiv:2211.03818 .

Guo, Y., Qiu, W., Leroy, G., Wang, S., Cohen, T.A., 2022b. Cells: A parallel corpus for biomedical lay language generation. ArXiv abs/2211.03818. URL: https://api.semanticscholar.org/CorpusID:253397732.

Guo, Y., Qiu, W., Wang, Y., Cohen, T., 2021. Automated lay language summarization of biomedical scientific reviews, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 160–168.

Gupta, S., Gupta, S.K., 2019. Abstractive summarization: An overview of the state of the art. Expert Systems with Applications 121, 49–65.

Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L., 2012. Development of a benchmark corpus to support the automatic extraction of drug-

related adverse effects from medical case reports. Journal of biomedical informatics 45, 885–892.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020a. Don't stop pretraining: Adapt language models to domains and tasks. ArXiv abs/2004.10964.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020b. Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 8342–8360. URL: https://aclanthology.org/2020.acl-main.740, doi:10.18653/v1/2020.acl-main.740.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.A., 2020c. Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8342–8360.

Gusfield, D., 1997. Algorithms on Strings, Trees and Sequences. Cambridge University Press, Cambridge, UK. URL: https://www.cambridge.org/core/books/algorithms-on-strings-trees-and-sequences/F0B095049C7E6EF5356F0A26686C20D3.

Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M., 2020. Retrieval augmented language model pre-training, in: International Conference on Machine Learning, PMLR. pp. 3929–3938.

Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., Ofek-Koifman, S., 2009. Personalized recommendation of social software items based on social relations, in: ACM Conference on Recommender Systems, p. 53–60. URL: https://api.semanticscholar.org/CorpusID:7108068.

Habibi, M., Weber, L., Neves, M., Wiegandt, D.L., Leser, U., 2017. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33, i37–i48.

Harris, Z., Gottfried, M., Ryckman, T., Daladier, A., Mattick, P., 2012. The form of information in science: analysis of an immunology sublanguage. volume 104. Springer Science & Business Media.

He, T., Zhang, J., Wang, T., Kumar, S., Cho, K., Glass, J., Tsvetkov, Y., 2022. On the blind spots of model-based evaluation metrics for text generation. arXiv preprint arXiv:2212.10020 .

Head, A., Lo, K., Kang, D., Fok, R., Skjonsberg, S., Weld, D.S., Hearst, M.A., 2020. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems URL: https://api.semanticscholar.org/CorpusID:222066998.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., Song, D., 2020. Pretrained transformers improve out-of-distribution robustness, in: ACL, p. 2744–2751.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics , 65–70.

Holmes-Rovner, M., Stableford, S., Fagerlin, A., Wei, J.T., Dunn, R.L., Ohene-Frempong, J., Kelly-Blake, K., Rovner, D.R., 2005. Evidence-based patient choice: a prostate cancer decision aid in plain language. BMC Medical Informatics and Decision Making 5, 1–11.

Holyoak, K.J., Thagard, P., 1996. The analogical scientist, in: Holyoak, K.J., Thagard, P. (Eds.), Mental Leaps: Analogy in Creative Thought. Cambridge, MA, pp. 185–209.

Howes, F., Doyle, J., Jackson, N., Waters, E., 2004. Evidence-based public health: the importance of finding 'difficult to locate'public health and health promotion intervention studies for systematic reviews. Journal of public health 26, 101–104.

Hughes, R.A., Mehndiratta, M.M., Rajabally, Y.A., 2017. Corticosteroids for chronic inflammatory demyelinating polyradiculoneuropathy. Cochrane Database of Systematic Reviews .

Jain, D., Borah, M.D., Biswas, A., 2021. Summarization of legal documents: Where are we now and the way forward. Computer Science Review 40, 100388.

Jain, R., Jangra, A., Saha, S., Jatowt, A., 2022. A survey on medical document summarization. arXiv preprint arXiv:2212.01669 .

Johnson, J., Douze, M., Jégou, H., 2019. Billion-scale similarity search with gpus. IEEE Transactions on Big Data 7, 535–547.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C., Gonzalez, G., 2009a. Towards effective sentence simplification for automatic processing of biomedical text, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, Association for Computational Linguistics, Boulder, Colorado. pp. 177–180. URL: https://aclanthology.org/N09-2045.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C., Gonzalez, G., 2009b. Towards effective sentence simplification for automatic processing of biomedical text, in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, pp. 177–180.

Jonnalagadda, S., Tari, L., Hakenberg, J., Baral, C., Gonzalez, G., 2010. Towards effective sentence simplification for automatic processing of biomedical text. arXiv preprint arXiv:1001.4277 .

Joshi, A., Katariya, N., Amatriain, X., Kannan, A., 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures., in: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3755–3763.

Kang, H.B., Kocielnik, R., Head, A., Yang, J., Latzke, M., Kittur, A., Weld, D.S., Downey, D., Bragg, J., 2022a. From who you know to what you read: Augmenting scientific recommendations with implicit social networks. Proceedings of the 2022 CHI Con-

ference on Human Factors in Computing Systems URL: https://api.semanticscholar.org/CorpusID:248299830.

Kang, H.B., Mysore, S., Huang, K., Chang, H.S., Prein, T., McCallum, A., Kittur, A., Olivetti, E.A., 2022b. Augmenting scientific creativity with retrieval across knowledge domains. ArXiv abs/2206.01328. URL: https://api.semanticscholar.org/CorpusID:249375678.

Kanthara, S., Leong, R.T.K., Lin, X., Masry, A., Thakkar, M., Hoque, E., Joty, S., 2022. Chart-to-text: A large-scale benchmark for chart summarization. arXiv preprint arXiv:2203.06486 .

Karačić, J., Dondio, P., Buljan, I., Hren, D., Marušić, A., 2019. Languages for different health information readers: multitrait-multimethod content analysis of cochrane systematic reviews textual summary formats. BMC medical research methodology 19, 1–9.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., Yih, W.t., 2020. Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781.

Kauchak, D., Leroy, G., Hogue, A., 2017. Measuring text difficulty using parse-tree frequency. Journal of the Association for Information Science and Technology 68, 2088–2100.

Kauchak, D., Mouradi, O., Pentoney, C., Leroy, G., 2014. Text simplification tools: Using machine learning to discover features that identify difficult text, in: 2014 47th Hawaii international conference on system sciences, IEEE. pp. 2616–2625.

Kim, H., Goryachev, S., Rosemblat, G., Browne, A., Keselman, A., Zeng-Treitler, Q., 2007. Beyond surface characteristics: a new health text-specific readability measurement, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association. p. 418.

Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S., 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report. Naval Technical Training Command Millington TN Research Branch.

Kinney, R.M., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., Crawford, M., Downey, D., Dunkelberger, J., Etzioni, O., Evans, R., Feldman, S., Gorney, J., Graham, D.W., Hu, F., Huff, R., King, D., Kohlmeier, S., Kuehl, B., Langan, M., Lin, D., Liu, H., Lo, K., Lochner, J., MacMillan, K., Murray, T., Newell, C., Rao, S.R., Rohatgi, S., Sayre, P.L., Shen, Z., Singh, A., Soldaini, L., Subramanian, S., Tanaka, A., Wade, A.D., Wagner, L.M., Wang, L.L., Wilhelm, C., Wu, C., Yang, J., Zamarron, A., van Zuylen, M., Weld, D.S., 2023. The semantic scholar open data platform. ArXiv abs/2301.10140. URL: https://api.semanticscholar.org/CorpusID:256194545.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., et al., 1983. Optimization by simulated annealing. science 220, 671–680.

Kirsch, I.S., et al., 1993. Adult literacy in America: A first look at the results of the National Adult Literacy Survey. ERIC.

Ko, W.J., Durrett, G., Li, J.J., 2019. Domain agnostic real-valued specificity prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6610–6617.

Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P., 2020. Concept bottleneck models. ArXiv abs/2007.04612. URL: https://api.semanticscholar.org/CorpusID:220424448.

Korsch, B.M., Gozzi, E.K., Francis, V., 1968. Gaps in doctor-patient communication: I. doctor-patient interaction and patient satisfaction. Pediatrics 42, 855–871.

Kotseruba, I., Tsotsos, J.K., 2018. 40 years of cognitive architectures: core cognitive abilities and practical applications. Artificial Intelligence Review 53, 17–94. URL: https://api.semanticscholar.org/CorpusID:51888132.

Krippendorff, K., 1970. Estimating the reliability, systematic error and random error of interval data. Educational and Psychological Measurement 30, 61–70.

Krishna, K., Bransom, E., Kuehl, B., Iyyer, M., Dasigi, P., Cohan, A., Lo, K., 2023a. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. arXiv preprint arXiv:2301.13298 .

Krishna, K., Bransom, E., Kuehl, B., Iyyer, M., Dasigi, P., Cohan, A., Lo, K., 2023b. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization, in: European Chapter of the Association for Computational Linguistics, pp. 1650–1669.

Kuehne, L.M., Olden, J.D., 2015. Lay summaries needed to enhance science communication. Proceedings of the National Academy of Sciences 112, 3585–3586.

Kurtzman, E.T., Greene, J., 2016. Effective presentation of health care performance information for consumer decision making: a systematic review. Patient education and counseling 99, 36–43.

Laban, P., Schnabel, T., Bennett, P., Hearst, M.A., 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. arXiv preprint arXiv:2107.03444 .

Laban, P., Vig, J., Kryscinski, W., Joty, S.R., Xiong, C., Wu, C.S., 2023. Swipe: A dataset for document-level simplification of wikipedia pages, in: Annual Meeting of the Association for Computational Linguistics, p. 10674–10695. URL: https://api.semanticscholar.org/CorpusID:258967312.

Lee, J.S.Y., Yeung, C.Y., 2018. Personalizing lexical simplification, in: International Conference on Computational Linguistics, p. 224–232. URL: https://api.semanticscholar.org/CorpusID:52012455.

Leroy, G., Carroll, E.L., Bruford, M.W., DeWoody, J.A., Strand, A., Waits, L., Wang, J., 2018. Next-generation metrics for monitoring genetic erosion within populations of conservation concern. Evolutionary Applications 11, 1066–1083.

Leroy, G., Helmreich, S., Cowie, J.R., Miller, T., Zheng, W., 2008. Evaluating online health information: Beyond readability formulas, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association. p. 394.

Leroy, G., Kauchak, D., 2014. The effect of word familiarity on actual and perceived text difficulty. Journal of the American Medical Informatics Association 21, e169–e172.

Levy, H., Janke, A., 2016. Health literacy and access to care. Journal of Health Communication 21, 43–50.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 .

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2020b. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. ArXiv abs/1910.13461.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al., 2020c. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems 33, 9459–9474.

Li, H., Su, Y., Cai, D., Wang, Y., Liu, L., 2022a. A survey on retrieval-augmented text generation. arXiv preprint arXiv:2202.01110 .

Li, H., Zhu, J., Zhang, J., Zong, C., He, X., 2020a. Keywords-guided abstractive sentence summarization, in: AAAI Conference on Artificial Intelligence, pp. 8196–8203. URL: https://api.semanticscholar.org/CorpusID:213639967.

Li, J., Lester, C., Zhao, X., Ding, Y., Jiang, Y., Vydiswaran, V., 2022b. Pharmmt: a neural machine translation approach to simplify prescription directions. arXiv preprint arXiv:2204.03830 .

Li, J., Lester, C., Zhao, X., Ding, Y., Jiang, Y., Vydiswaran, V.V., 2020b. Pharmmt: A neural machine translation approach to simplify prescription directions, in: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2785–2796.

Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, pp. 74–81.

Lin, H.N., Hsieh, S.K., Chan, S.H., 2012. Measuring individual differences in word recognition: The role of individual lexical behaviors, in: ROCLING/IJCLCLP, pp. 61–74. URL: https://api.semanticscholar.org/CorpusID:5616068.

Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N.A., Choi, Y., 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online. pp. 6691–6706. URL: https://aclanthology.org/2021.acl-long.522, doi:10.18653/v1/2021.acl-long.522.

Liu, H., Dai, S., Jiang, D.e., 2014. Solubility of gases in a common ionic liquid from molecular dynamics based free energy calculations. The Journal of Physical Chemistry B 118, 2719–2725.

Liu, Y., Lapata, M., 2019. Text summarization with pretrained encoders, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3721–3731.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer,

L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .

Lu, J., Li, J., Wallace, B.C., He, Y., Pergola, G., 2023a. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. arXiv preprint arXiv:2302.05574 .

Lu, J., Li, J., Wallace, B.C., He, Y., Pergola, G., 2023b. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization, in: Findings, pp. 1079–1091.

Luan, Y., Eisenstein, J., Toutanova, K., Collins, M., 2021. Sparse, dense, and attentional representations for text retrieval. Transactions of the Association for Computational Linguistics 9, 329–345.

Lucy, L., Dodge, J., Bamman, D., Keith, K.A., 2022. Words as gatekeepers: Measuring discipline-specific terms and meanings in scholarly publications, in: Annual Meeting of the Association for Computational Linguistics, pp. 6929–6947. URL: https://api.semanticscholar.org/CorpusID:254854025.

Luo, J., Lin, J., Lin, C., Xiao, C., Gui, X., Ma, F., 2022a. Benchmarking automated clinical language simplification: Dataset, algorithm, and evaluation, in: Proceedings of the 29th International Conference on Computational Linguistics, pp. 3550–3562.

Luo, M., Mitra, A., Gokhale, T., Baral, C., 2022b. Improving biomedical information retrieval with neural retrievers. arXiv preprint arXiv:2201.07745 .

Luo, Z., Xie, Q., Ananiadou, S., 2022c. Readability controllable biomedical document summarization. arXiv preprint arXiv:2210.04705 .

Luo, Z., Xie, Q., Ananiadou, S., 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. arXiv preprint arXiv:2303.15621 .

Luria, S.E., Delbrück, M., 1943. Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28, 491.

Luu, K., Koncel-Kedziorski, R., Lo, K., Cachola, I., Smith, N.A., 2020. Citation text generation. ArXiv abs/2002.00317.

Ma, Y., Liu, J., Yi, F., 2023. Is this abstract generated by ai? a research for the gap between ai-generated scientific text and human-written scientific text. ArXiv abs/2301.10416. URL: https://api.semanticscholar.org/CorpusID:256231054.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of machine learning research 9.

Mac, O., Ayre, J., Bell, K., McCaffery, K., Muscat, D.M., 2022. Comparison of readability scores for written health information across formulas using automated vs manual measures. JAMA Network Open 5, e2246051–e2246051.

Maddela, M., Dou, Y., Heineman, D., Xu, W., 2022. Lens: A learnable evaluation metric for text simplification. arXiv preprint arXiv:2212.09739 .

Manor, L., Li, J.J., 2019. Plain english summarization of contracts. arXiv preprint arXiv:1906.00424 .

Martinc, M., Pollak, S., Robnik-Šikonja, M., 2021. Supervised and unsupervised neural approaches to text readability. Computational Linguistics 47, 141–179.

Martínez, A., Mammola, S., 2021. Specialized terminology reduces the number of citations of scientific papers. Proceedings of the Royal Society B 288. URL: https://api.semanticscholar.org/CorpusID:233129354.

Maynez, J., Narayan, S., Bohnet, B., McDonald, R.T., 2020. On faithfulness and factuality in abstractive summarization. ArXiv abs/2005.00661.

McCoy, R.T., Pavlick, E., Linzen, T., 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv preprint arXiv:1902.01007 .

McIlwain, C., Santesso, N., Simi, S., Napoli, M., Lasserson, T., Welsh, E., Churchill, R., Rader, T., Chandler, J., Tovey, D., et al., 2014. Standards for the reporting of Plain

Language Summaries in new Cochrane Intervention Reviews (PLEACS). Cochrane database of systematic reviews. https://methods.cochrane.org/sites/default/files/public/uploads/pleacs_2019.pdf.

McNamara, D.S., Kintsch, E., Songer, N.B., Kintsch, W., 1996. Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and instruction 14, 1–43.

Miller, G.A., 1995. Wordnet: a lexical database for english. Communications of the ACM 38, 39–41.

Mishra, R., Bian, J., Fiszman, M., Weir, C.R., Jonnalagadda, S., Mostafa, J., Del Fiol, G., 2014. Text summarization in the biomedical domain: a systematic review of recent research. Journal of biomedical informatics 52, 457–467.

Mollá, D., Santiago-Martinez, M.E., et al., 2011. Development of a corpus for evidence based medicine summarisation, in: Proceedings of the Australasian Language Technology Association Workshop 2011, Canberra, Australia. pp. 86–94. URL: https://www.aclweb.org/anthology/U11-1012.

Moradi, M., Ghadiri, N., 2018. Different approaches for identifying important concepts in probabilistic biomedical text summarization. Artificial intelligence in medicine 84, 101–116.

Moradi, M., Ghadiri, N., 2019. Text summarization in the biomedical domain. ArXiv abs/1908.02285.

Mukherjee, P., Leroy, G., Kauchak, D., Navarrete, B.A., Diaz, D.Y., Colina, S., 2017. The role of surface, semantic and grammatical features on simplification of spanish medical texts: A user study, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association. p. 1322.

Murad, M.H., Asi, N., Alsawas, M., Alahdab, F., 2016. New evidence pyramid. BMJ Evidence-Based Medicine 21, 125–127.

Murthy, S.K., King, D., Hope, T., Weld, D.S., Downey, D., 2021. Towards personalized descriptions of scientific concepts, in: The Fifth Widening Natural Language Processing Workshop at EMNLP, pp. 0–8. URL: https://api.semanticscholar.org/CorpusID:247410502.

Mysore, S., Jasim, M., McCallum, A., Zamani, H., 2023. Editable user profiles for controllable text recommendations. Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval URL: https://api.semanticscholar.org/CorpusID:258049038.

Naik, A., Parasa, S., Feldman, S., Wang, L.L., Hope, T., 2021. Literature-augmented clinical outcome prediction. arXiv preprint arXiv:2111.08374 .

Nallapati, R., Zhou, B., Santos, C.D., Çaglar Gülçehre, Xiang, B., 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond, in: CoNLL, pp. 1–14.

Narechania, A., Karduni, A., Wesslen, R., Wall, E., 2021. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. IEEE Transactions on Visualization and Computer Graphics 28, 486–496. URL: https://api.semanticscholar.org/CorpusID:236957094.

Neumann, M., King, D., Beltagy, I., Ammar, W., 2019. Scispacy: Fast and robust models for biomedical natural language processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327.

Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P., 2014. The quaero french medical corpus: A ressource for medical entity recognition and normalization, in: In proc biotextm, reykjavik, Citeseer. pp. 1–14.

Nunn, E., Pinfield, S., 2014. Lay summaries of open access journal articles: engaging with the general public on medical research. Learned Publishing 27, 173–184.

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

OpenAI, 2023a. Gpt-4 technical report. arXiv:2303.08774.

OpenAI, 2023b. Gpt-4 technical report. ArXiv abs/2303.08774.

Ormazabal, A., Artetxe, M., Labaka, G., Soroa, A., Agirre, E., 2022. Principled paraphrase generation with parallel corpora. arXiv preprint arXiv:2205.12213 .

Osborne, F., Motta, E., 2015. Klink-2: Integrating multiple web sources to generate semantic topic networks, in: International Workshop on the Semantic Web, pp. 408–424. URL: https://api.semanticscholar.org/CorpusID:8677807.

Paakkari, L., Okan, O., 2020. Covid-19: health literacy is an underestimated problem. The Lancet Public Health 5, e249–e250.

Pagnoni, A., Balachandran, V., Tsvetkov, Y., 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. arXiv preprint arXiv:2104.13346 .

Paice, C., 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases, in: SIGIR, pp. 172–191.

Pampari, A., Raghavan, P., Liang, J., Peng, J., 2018. emrqa: A large corpus for question answering on electronic medical records. arXiv preprint arXiv:1809.00732 .

Papineni, K., Roukos, S., Ward, T., Zhu, W.J., 2002. Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318.

Paravastu, S.C.V., Horne, M., Dodd, P.D.F., 2016. Endovenous ablation therapy (laser or radiofrequency) or foam sclerotherapy versus conventional surgical repair for short saphenous varicose veins. Cochrane Database of Systematic Reviews .

Parker, R.M., Williams, M.V., Weiss, B.D., Baker, D.W., Davis, T.C., Doak, C.C., Doak, L.G., Hein, K., Meade, C.D., Nurss, J., et al., 1999. Health literacy-report of the council on scientific affairs. Jama-Journal of the American Medical Association 281, 552–557.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems 32.

Pattisapu, N., Prabhu, N., Bhati, S., Varma, V., 2020. Leveraging social media for medical text simplification, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 851–860.

Peng, H., Parikh, A.P., Faruqui, M., Dhingra, B., Das, D., 2019. Text generation with exemplar-based adaptive decoding. arXiv preprint arXiv:1904.04428 .

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations, in: Walker, M.A., Ji, H., Stent, A. (Eds.), NAACL-HLT, Association for Computational Linguistics. pp. 2227–2237.

Pirolli, P., Card, S., 1999. Information foraging. Psychological review 106, 643.

Pitcher, N., Mitchell, D., Hughes, C., 2022. Template and guidance for writing a cochrane plain language summary.

Pitler, E., Nenkova, A., 2008. Revisiting readability: A unified framework for predicting text quality, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii. pp. 186–195.

Pivovarov, R., Elhadad, N., 2015. Automated methods for the summarization of electronic health records. Journal of the American Medical Informatics Association 22, 938–947.

Plaven-Sigray, P., Matheson, G.J., Schiffler, B.C., Thompson, W.H., 2017. The readability of scientific texts is decreasing over time. eLife 6. URL: https://api.semanticscholar.org/CorpusID:3629355.

Plaza, L., 2014. Comparing different knowledge sources for the automatic summarization of biomedical literature. Journal of biomedical informatics 52, 319–328.

Portenoy, J., Radensky, M., West, J.D., Horvitz, E., Weld, D.S., Hope, T., 2021. Bursting scientific filter bubbles: Boosting innovation via novel author discovery. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems , 1–13URL: https://api.semanticscholar.org/CorpusID:237485597.

Prabhakaran, V., Hutchinson, B., Mitchell, M., 2019. Perturbation sensitivity analysis to detect unintended model biases, in: Conference on Empirical Methods in Natural Language Processing, pp. 1–14.

Pruksachatkun, Y., Phang, J., Liu, H., Htut, P.M., Zhang, X., Pang, R.Y., Vania, C., Kann, K., Bowman, S.R., 2020. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? ArXiv abs/2005.00628.

PubMed, a. Searching results from web of science core collection for 'medical'. https://www.webofscience.com/wos/woscc/summary/4e31b58f-ef9a-4d79-9a13-9dd0fc1a2ba7-353cc5dd/relevance/1. Accessed: 2022-05-03.

PubMed, b. Total number of papers in pubmed. https://pubmed.ncbi.nlm.nih.gov/?term=all%5Bsb%5D+. Accessed: 2022-05-03.

Pushparajah, D.S., Manning, E., Michels, E., Arnaudeau-Bégard, C., 2018. Value of developing plain language summaries of scientific and clinical articles: a survey of patients and physicians. Therapeutic innovation & regulatory science 52, 474–481.

Qenam, B., Kim, T.Y., Carroll, M.J., Hogarth, M., 2017. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. Journal of medical Internet research 19, e417.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., 2018. Improving Language Understanding by Generative Pre-Training. Technical Report. OpenAI.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21, 5485–5551.

Rakedzon, T., Segev, E., Chapnik, N., Yosef, R., Baram-Tsabari, A., 2017. Automatic jargon identifier for scientists engaging with the public and science communication educators. PLoS ONE 12. URL: https://api.semanticscholar.org/CorpusID:24435125.

Rasooli, M.S., Tetreault, J.R., 2015. Yara parser: A fast and accurate dependency parser. Computing Research Repository arXiv:1503.06733. URL: http://arxiv.org/abs/1503.06733. version 2.

Redish, J., 2000. Readability formulas have even more limitations than klare discusses. ACM Journal of Computer Documentation (JCD) 24, 132–137.

Reiter, E., Belz, A., 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Computational Linguistics 35, 529–558.

Ribeiro, M.T., Guestrin, C., Singh, S., 2019. Are red roses red? evaluating consistency of question-answering models, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6174–6184.

Ridder, I.D., 2002. Visible or invisible links?, in: CHI Extended Abstracts, pp. 624–625. URL: https://api.semanticscholar.org/CorpusID:12632765.

Roberts, K., Simpson, M.S., Voorhees, E.M., Hersh, W.R., 2015. Overview of the trec 2015 clinical decision support track., in: TREC, pp. 1–16.

Robertson, S., Zaragoza, H., et al., 2009. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval 3, 333–389.

Roby, W.B., 2005. ”what’s in a gloss?” a commentary on lara l. lomicka’s ”to gloss or not to gloss”: An investigation of reading comprehension online. language learning & technology, vol. 1, no. 2. University of Hawaii National Foreign Language Resource Center URL: https://api.semanticscholar.org/CorpusID:31192898.

Rush, A.M., Chopra, S., Weston, J., 2015. A neural attention model for abstractive sentence

summarization, in: Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., Marton, Y. (Eds.), EMNLP, The Association for Computational Linguistics. pp. 379–389.

Sai, A.B., Dixit, T., Sheth, D.Y., Mohan, S., Khapra, M.M., 2021. Perturbation checklists for evaluating nlg evaluation metrics. arXiv preprint arXiv:2109.05771 .

Sai, A.B., Mohankumar, A.K., Khapra, M.M., 2022. A survey of evaluation metrics used for nlg systems. ACM Computing Surveys (CSUR) 55, 1–39.

Salatino, A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E., 2018. The computer science ontology: A large-scale taxonomy of research areas, in: International Workshop on the Semantic Web, pp. 187–205. URL: https://api.semanticscholar.org/CorpusID:49481128.

Salo, M., Haapio, H., Passera, S., 2016. Putting financial regulation to work: Using simplification and visualization for consumer-friendly information, in: Networks. Proceedings of the 19th International Legal Informatics Symposium IRIS, pp. 399–406.

Samsonovich, A., Jong, K.A.D., Kitsantas, A., Peters, E.E., Dabbagh, N., Kalbfleisch, M.L., 2008. Cognitive constructor: An intelligent tutoring system based on a biologically inspired cognitive architecture (bica), in: Artificial General Intelligence, p. 311. URL: https://api.semanticscholar.org/CorpusID:1898659.

Sarkar, K., Nasipuri, M., Ghose, S., 2011. Using machine learning for medical document summarization. International Journal of Database Theory and Application 4.

Scarton, C., Paetzold, G., Specia, L., 2018. Text simplification from professionally produced corpora, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 1–12.

Schoch, S., Yang, D., Ji, Y., 2020. “this is a problem, don’t you agree?” framing and bias in human evaluation for natural language generation, in: Proceedings of the 1st Workshop on Evaluating NLG Evaluation, pp. 10–16.

Schvaneveldt, R.W., 1990. Pathfinder associative networks: studies in knowledge organization. URL: https://api.semanticscholar.org/CorpusID:62565850.

Scialom, T., Lamprier, S., Piwowarski, B., Staiano, J., 2019. Answers unite! unsupervised metrics for reinforced summarization models. arXiv preprint arXiv:1909.01610 .

See, A., Liu, P.J., Manning, C.D., 2017. Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1073–1083.

Sen, P.K., 1968. Estimates of the regression coefficient based on kendall's tau. Journal of the American statistical association 63, 1379–1389.

Shardlow, M., 2014. A survey of automated text simplification. International Journal of Advanced Computer Science and Applications 4.

Shuster, K., Poff, S., Chen, M., Kiela, D., Weston, J., 2021. Retrieval augmentation reduces hallucination in conversation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic. pp. 3784–3803. URL: https://aclanthology.org/2021.findings-emnlp.320, doi:10.18653/v1/2021.findings-emnlp.320.

Silveira, S.B., Branco, A., 2012. Enhancing multi-document summaries with sentence simplification, in: Proceedings on the International Conference on Artificial Intelligence (ICAI), The Steering Committee of The World Congress in Computer Science, Computer .... p. 1.

Simpson, M.S., Voorhees, E.M., Hersh, W., 2014. Overview of the trec 2014 clinical decision support track. Technical Report. LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., Feldman, S., 2022. Scirepeval: A multi-format benchmark for scientific document representations. ArXiv abs/2211.13308, 1–16.

Smith, R., Snow, P., Serry, T., Hammond, L., 2021. The role of background knowledge in reading comprehension: A critical review. Reading Psychology 42, 214–240.

Soroya, S.H., Farooq, A., Mahmood, K., Isoaho, J., Zara, S.e., 2021. From information seeking to information avoidance: Understanding the health information behavior during a global health crisis. Information processing & management 58, 102440.

Srikanth, N., Li, J.J., 2020a. Elaborative simplification: Content addition and explanation generation in text simplification. arXiv preprint arXiv:2010.10035 .

Srikanth, N., Li, J.J., 2020b. Elaborative simplification: Content addition and explanation generation in text simplification, in: Findings, pp. 1–8. URL: https://api.semanticscholar.org/CorpusID:224803326.

Stephens, D.W., Krebs, J.R., 1986. Foraging theory. Princeton University Press.

Stoll, M., Kerwer, M., Lieb, K., Chasiotis, A., 2022. Plain language summaries: A systematic review of theory, guidelines and empirical research. Plos one 17, e0268789.

Strober, M.H., 2006. Habits of the mind: Challenges for multidisciplinary engagement. Social Epistemology 20, 315 – 331. URL: https://api.semanticscholar.org/CorpusID:15709099.

Sudore, R.L., Schillinger, D., 2009. Interventions to improve care for patients with limited health literacy. Journal of clinical outcomes management: JCOM 16, 20.

Sugawara, S., Stenetorp, P., Inui, K., Aizawa, A., 2020. Assessing the benchmarking capacity of machine reading comprehension datasets, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8918–8927.

Sulem, E., Abend, O., Rappoport, A., 2018. Bleu is not suitable for the evaluation of text simplification. arXiv preprint arXiv:1810.05995 .

Szmuda, T., Özdemir, C., Ali, S., Singh, A., Syed, M.T., Słoniewski, P., 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.

Talbot, J., Aronson, J.K., 2011. Stephens' detection and evaluation of adverse drug reactions: principles and practice. John Wiley & Sons.

Tanaka-Ishii, K., Terada, H., 2011. Word familiarity and frequency. ArXiv abs/1806.03431. URL: https://api.semanticscholar.org/CorpusID:47019416.

Teufel, S., Moens, M., 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. Comput. Linguistics 28, 409–445.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

Uman, L.S., 2011. Systematic reviews and meta-analyses. Journal of the Canadian Academy of Child and Adolescent Psychiatry 20, 57.

Vilhena, D.A., Foster, J.G., Rosvall, M., West, J.D., Evans, J.A., Bergstrom, C.T., 2014. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. Sociological Science 1, 221–238. URL: https://api.semanticscholar.org/CorpusID:27876915.

Wallace, B.C., Saha, S., Soboczenski, F., Marshall, I.J., 2021. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. AMIA Summits on Translational Science Proceedings 2021, 605.

Wang, B., Xie, Q., Pei, J., Tiwari, P., Li, Z., et al., 2021a. Pre-trained language models in biomedical domain: A systematic survey. arXiv preprint arXiv:2110.05006 .

Wang, M., Wang, M., Yu, F., Yang, Y., Walker, J., Mostafa, J., 2021b. A systematic review of automatic text summarization for biomedical literature and ehrs. Journal of the American Medical Informatics Association .

Wang, M., Wang, M., Yu, F., Yang, Y., Walker, J., Mostafa, J., 2021c. A systematic review of automatic text summarization for biomedical literature and ehrs. Journal of the American Medical Informatics Association 28, 2287–2297.

Wang, Y., Deng, J., Sun, A., Meng, X., 2022. Perplexity from plm is unreliable for evaluating text quality. arXiv preprint arXiv:2210.05892 .

Wilson, M., 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. Behavior Research Methods, Instruments, & Computers 20, 6–10. URL: https://api.semanticscholar.org/CorpusID:62652458.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al., 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 .

Woodsend, K., Lapata, M., 2011. Wikisimple: Automatic simplification of wikipedia articles, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 927–932.

Wright, D., Wadden, D., Lo, K., Kuehl, B., Cohan, A., Augenstein, I., Wang, L.L., 2022. Generating scientific claims for zero-shot scientific fact checking. arXiv preprint arXiv:2203.12990 .

Wu, C., Wu, F., Huang, Y., Xie, X., 2021. User-as-graph: User modeling with heterogeneous graph pooling for news recommendation, in: International Joint Conference on Artificial Intelligence, pp. 1624–1630. URL: https://api.semanticscholar.org/CorpusID:237101184.

Wubben, S., Krahmer, E., van den Bosch, A., 2012. Sentence simplification by monolingual machine translation, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1015–1024.

Wudarczyk, O.A., Kirtay, M., Kuhlen, A.K., Rahman, R.A., Haynes, J.D., Hafner, V.V., Pischedda, D., 2021. Bringing together robotics, neuroscience, and psychology: Lessons learned from an interdisciplinary project. Frontiers in Human Neuroscience 15. URL: https://api.semanticscholar.org/CorpusID:232387364.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C., 2016. Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics 4, 401–415.

Xu, W., Portanova, J., Chander, A., Ben-Zeev, D., Cohen, T., 2020. The centroid cannot hold: Comparing sequential and global estimates of coherence as indicators of formal thought disorder, in: AMIA Annual Symposium Proceedings, American Medical Informatics Association. p. 1315.

Yadav, D., Desai, J., Yadav, A.K., 2022. Automatic text summarization methods: A comprehensive review. arXiv preprint arXiv:2204.01849 .

Yang, L., Hu, J., Qiu, M., Qu, C., Gao, J., Croft, W.B., Liu, X., Shen, Y., Liu, J., 2019. A hybrid retrieval-generation neural conversation model, in: Proceedings of the 28th ACM international conference on information and knowledge management, pp. 1341–1350.

Yuan, W., Neubig, G., Liu, P., 2021. Bartscore: Evaluating generated text as text generation. Advances in Neural Information Processing Systems 34, 27263–27277.

Zesch, T., Müller, C., Gurevych, I., 2008. Extracting lexical semantic knowledge from wikipedia and wiktionary., in: LREC, pp. 1646–1652.

Zhang, J., 2002. Representations of health concepts: a cognitive perspective. Journal of biomedical informatics 35 1, 17–24. URL: https://api.semanticscholar.org/CorpusID:4632497.

Zhang, J., Hamilton, W., Danescu-Niculescu-Mizil, C., Jurafsky, D., Leskovec, J., 2017. Community identity and user engagement in a multi-community landscape, in: Proceedings of the international AAAI conference on web and social media, pp. 377–386.

Zhang, L., Negrinho, R., Ghosh, A., Jagannathan, V., Hassanzadeh, H.R., Schaaf, T., Gormley, M.R., 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. arXiv preprint arXiv:2109.12174 .

Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y., 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 .

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y., 2020. Bertscore: Evaluating

text generation with bert, in: International Conference on Learning Representations, pp. 1–8. URL: https://openreview.net/forum?id=SkeHuCVFDr.

Zhang, X., Li, J.K., Chi, P.W., Chandrasegaran, S.K., Ma, K.L., 2023. Concepteva: Concept-based interactive exploration and customization of document summaries. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems , 1–16URL: https://api.semanticscholar.org/CorpusID:257901215.

Zhang, Y., Ding, D.Y., Qian, T., Manning, C.D., Langlotz, C.P., 2018. Learning to summarize radiology findings, in: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, pp. 204–213.

Zhang, Y., Merck, D., Tsai, E., Manning, C.D., Langlotz, C., 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5108–5120.

Zhao, Y., Tian, Z., Yao, H., Zheng, Y., Lee, D., Song, Y., Sun, J., Zhang, N.L., 2022. Improving meta-learning for low-resource text classification and generation via memory imitation. arXiv preprint arXiv:2203.11670 .

Zhong, Y., Jiang, C., Xu, W., Li, J.J., 2020. Discourse level factors for sentence deletion in text simplification, in: Proceedings of the AAAI conference on artificial intelligence, pp. 9709–9716.

Appendix A

SUPPLEMENTARY MATERIALS FOR APPLS

A.1 Round-trip translation for oracle extractive hypothesis

We use round-trip translation to introduce lexical variation into our oracle extractive summaries. This is important when computing metrics such as SARI, which exhibit degenerate behavior when the hypothesis is an extractive subset of the source. We examine two languages for round-trip translation: German and Russian. By employing the BLEU score as a performance metric for the round-trip generated text relative to the original source, we find that the English-German-English (en-de-en) translation sequence yields superior BLEU scores (Figure A.1), and therefore, select the en-de-en sequence to produce the oracle extractive hypothesis for our testbed.

To scrutinize the introduced variation through this extractive and round-trip translation pipeline, we evaluate the BLEU score. As depicted in Figure A.2, the BLEU score for the oracle extractive hypothesis is lower than that of the oracle extractive summary. This suggests the successful introduction of text variations. Augmented by human evaluation results in Table A.1, with 152 out of 198 raters indicating comparable simplification levels

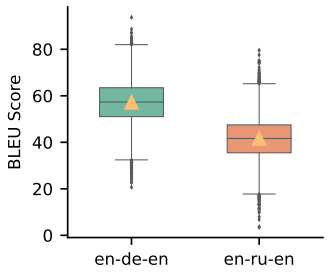


Figure A.1: BLEU scores of round-trip translation for English-German-English (en-de-en) and English-Russian-English (en-ru-en) in CELLS oracle extractive hypotheses.

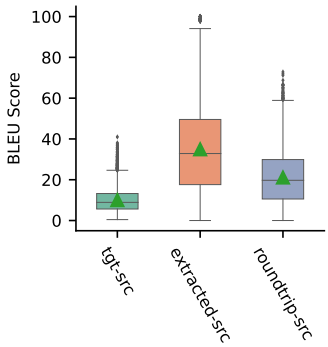


Figure A.2: Comparison of BLEU scores between oracle extractive summary (extracted) and oracle extractive hypothesis (roundtrip), using the scientific abstract (src) as the reference for BLEU calculation.

Type	Unmatched	Criteria	Str. Agree	Agree	Neutral	Disagree	Str. Disagree
Round Trip Translation	1	Simplification	12	27	152	7	0
		Informativeness	188	4	3	4	0
		Faithfulness	155	6	4	20	14
		Coherence	30	11	156	2	0
GPT Simplification	1	Simplification	143	35	16	5	0
		Informativeness	86	35	5	43	30
		Faithfulness	88	27	6	45	33
		Coherence	146	43	9	1	0

Table A.1: Counts of human evaluation ratings on each matched sentence for each criteria. For round trip translation and GPT simplification, there are a total of 400 ratings (2 annotators rating 200 pairs each). Overall, we see that round trip translation maintains strong faithfulness to the original, does not remove important information, and remains equally simple and coherent (shown by a majority of neutral ratings for the simplification and coherence criteria). For GPT simplification, we see that the simplification perturbation leads to substantially more simple text, while also maintaining faithfulness and informativeness.

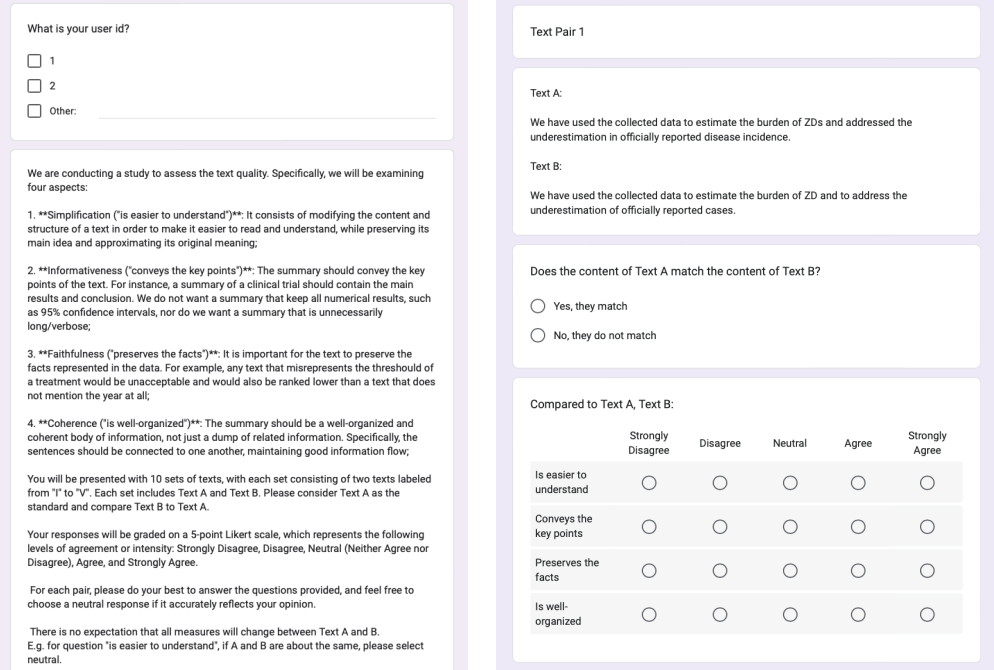


Figure A.3: An example human evaluation task for assessing GPT-simplified summary quality.

between the oracle extractive hypothesis and its extractive counterparts, we conclude that our extractive and round-trip translation approach successfully introduces lexical variation in our oracle extractive summaries without altering their simplicity level.

A.2 Details of human evaluation

To validate the quality of oracle extractive hypotheses and GPT-simplified summaries, we randomly select 100 summary pairs from each corpus for human evaluation. Each pair in the oracle extractive hypotheses consists of an oracle extractive sentence and its respective en-de-en round-trip-translation sentence. Similarly, each pair in the GPT-simplified summaries contains a hypothesis chunk along with its corresponding GPT-simplified summary chunk.

Each pair is reviewed by two independent annotators. Annotators were hired through UpWork and have Bachelors and Doctorate degrees in the biological sciences. In the evaluation, the text pairs are labeled as Text A and Text B, without any indication that either text

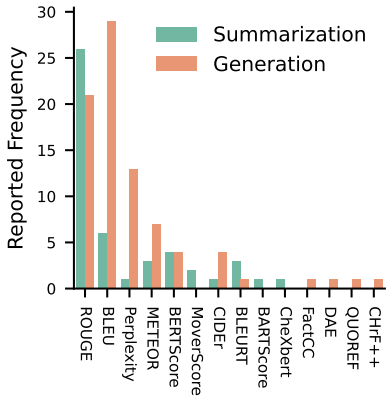


Figure A.4: Most common evaluation metrics reported in ACL'22 summarization and generation long papers.

is generated. The annotators are first asked to assess whether the content of Text A matches the content of Text B, where a match is defined as containing the same relation tuples. If the texts match, the annotators further evaluate Text B in relation to Text A, assessing whether Text B encapsulates key points (informativeness), is more comprehensible (simplification), maintains factual integrity (faithfulness), and exhibits a well-structured layout (coherence). All facets are assessed using a 1-5 Likert scale (1-strongly disagree, 5-strongly agree). Representative questions can be found in Figure A.3. This research activity is exempt from institutional IRB review.

A.3 Empirical Study of Evaluation Metrics Reported in ACL 2022 Publications

Our study undertakes a comprehensive analysis of scores reported in the long papers of ACL 2022 to identify the most prevalently reported metrics in summarization and simplification tasks. We primarily concentrate on tasks related to generation, summarization, and simplification. Our inclusion criteria are: 1) long papers with 'generat,' 'summar,' or 'simpl' in the title; and 2) papers that report scores for both the current model and at least one baseline model in the main text. We exclude scores from ablation studies.

Of the 601 long papers accepted to ACL 2022, 109 satisfy our inclusion criteria, which

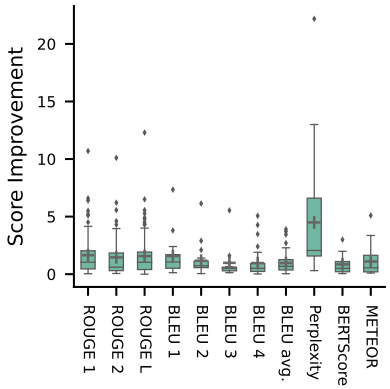


Figure A.5: Distributions of reported metric improvements over baseline (absolute value) reported in ACL'22 summarization and generation long papers.

we categorize into 31 summarization and 78 generation papers, with no qualified papers related to simplification tasks. Considering the significance of simplification in PLS, we expanded our search to all ACL 2022 papers, including long, short, system demonstration, and findings papers. This led to the identification of 2 out of 22 papers with 'simpl' in the title that reported SARI scores. As illustrated in Figure A.4, the five most frequently reported automated evaluation metrics are ROUGE, BLEU, GPT-PPL, METEOR, and BERTScore.

This investigation provides insight into the current adoption of evaluation metrics in natural language generation, summarization, and simplification tasks. We observe that a majority of papers employ the same metrics across these tasks, and the reported improvements are often relatively small compared to the overall ranges for each measure. We also underscore the difficulty of interpreting changes in some of these metrics, especially model-based metrics, which lack grounding to lexical differences in text such as n-gram overlap.

By presenting the reported score differences from ACL papers, we hope to contextualize the metric changes observed through testing in our meta-evaluation testbed. Median reported improvements for the most commonly reported metrics and SARI are: ROUGE (+0.89), BLEU (+0.69), PPL (-2.06), METEOR (+0.50), BERTScore (+0.55), and SARI (+1.71), as shown in Figure A.5. We report the median of BERTScore values and deltas as

	Ref. Free	Ref. Provided
Informativeness (↓)		
Delete sentence	**-61.09**	**-22.33**
Add out-of-domain sent	**-19.91**	**-34.88**
Add in-domain sent	**-7.32**	-6.99
Add definition (↑)	-0.6	12.64
Simplification (↑)	-15.98	-11.46
Coherence (↓)	-0.17	-1.06
Faithfulness (↓)		
Number swaps	0.45	-0.44
Entity swaps	**-5.48**	-7.24
Synonyms verb swaps	-4.02	-10.54
Antonyms verb swaps	**-6.05**	**-8.27**
Negate sentence	**-18.59**	**-21.94**

Table A.2: Overall score derived from the prompt-based evaluation for two settings: reference-free and reference-provided. **Bolded** values indicate statistical significance in the correct direction with Bonferroni-Holm correction for multiple hypothesis testing (Holm, 1979).

reported in these publications, without considering the usage of different models or settings.

A.4 LLM Prompt-Based Evaluation

a. Reference Free Prompt:
Imagine you are a human annotator now. You will evaluate the quality of generated plain lanugage summary written for a scientific literature abstract. **Please follow these steps:**
1. Carefully read the scientific literature abstract, and be aware of the information it contains.
2. Read the proposed generated plain langauge summary.
3. Compared to the scientific abstract, rate the summary on four dimensions: informativeness, simplification, coherence, and faithfulness. Assign a score for each aspect and provide an overall score. You should rate on a scale from 0 (worst) to 100 (best).
4. You do not need to explain the reason. Only provide the scores.

Definitions are as follows:
-**Informativeness**: measures the extent to which a plain language summary encapsulates essential elements such as methodologies, primary findings, and conclusions from the original scientific text. An informative summary efficiently conveys the central message of the source material, avoiding the exclusion of crucial details or the introduction of hallucinations (i.e., information present in the summary but absent in the scientific text), both of which could impair reader comprehension.
-**Simplification**: encompasses the rendering of information into a form that non-expert audiences can readily interpret and understand. This criterion prioritizes the use of simple vocabulary, casual language, and concise sentences that minimize excessive jargon and technical terminology unfamiliar to a lay audience.
-**Coherence**: pertains to the logical arrangement of a plain language summary. A coherent summary guarantees an unambiguous and steady progression of ideas, offering information in a well-ordered fashion that facilitates ease of comprehension for the reader. We conjecture that the original sentence order reflects optimal coherence.
-**Faithfulness**: denotes the extent to which the plain language summary aligns factually with the source scientific text, in terms of its findings, methods, and claims. A faithful summary should not substitute information or introduce errors, misconceptions, and inaccuracies, which can misguide the reader or misrepresent the original author's intent. Faithfulness emphasizes the factual alignment of the summary with the source text, while informativeness gauges the completeness and efficiency of the summary in conveying key elements.

The scientific abstract and the generated plain language summary are given below:
Scientific abstract: {}
Generated plain language sumamry:{}

b. Reference Provided Promt:
Imagine you are a human annotator now. You will evaluate the quality of generated summary written for a scientific literature abstract. **Please follow these steps:**
1. Carefully read the scientific abstract and plain language summary written by human, and be aware of the information it contains.
2. Read the proposed genereated summary.
3. Compared to the scientific abstract and human-written plain language summry, rate the generated summary on four dimensions: informativeness, simplification, coherence, and faithfulness. Assign a score for each aspect and provide an overall score. You should rate on a scale from 0 (worst) to 100 (best).
4. You do not need to explain the reason. Only provide the scores.

Definitions are as follows:
-**Informativeness**: measures the extent to which a plain language summary encapsulates essential elements such as methodologies, primary findings, and conclusions from the original scientific text. An informative summary efficiently conveys the central message of the source material, avoiding the exclusion of crucial details or the introduction of hallucinations (i.e., information present in the summary but absent in the scientific text), both of which could impair reader comprehension.
-**Simplification**: encompasses the rendering of information into a form that non-expert audiences can readily interpret and understand. This criterion prioritizes the use of simple vocabulary, casual language, and concise sentences that minimize excessive jargon and technical terminology unfamiliar to a lay audience.
-**Coherence**: pertains to the logical arrangement of a plain language summary. A coherent summary guarantees an unambiguous and steady progression of ideas, offering information in a well-ordered fashion that facilitates ease of comprehension for the reader. We conjecture that the original sentence order reflects optimal coherence.
-**Faithfulness**: denotes the extent to which the plain language summary aligns factually with the source scientific text, in terms of its findings, methods, and claims. A faithful summary should not substitute information or introduce errors, misconceptions, and inaccuracies, which can misguide the reader or misrepresent the original author's intent. Faithfulness emphasizes the factual alignment of the summary with the source text, while informativeness gauges the completeness and efficiency of the summary in conveying key elements.

The scientific abstract, plain language summary, and generated summary are given below:
Scientific abstract: {}
Plain language summary: {}
Generated summary: {}

Figure A.6: Prompts used for LLM evaluation. (a): Reference-free; (b) Reference-provided.

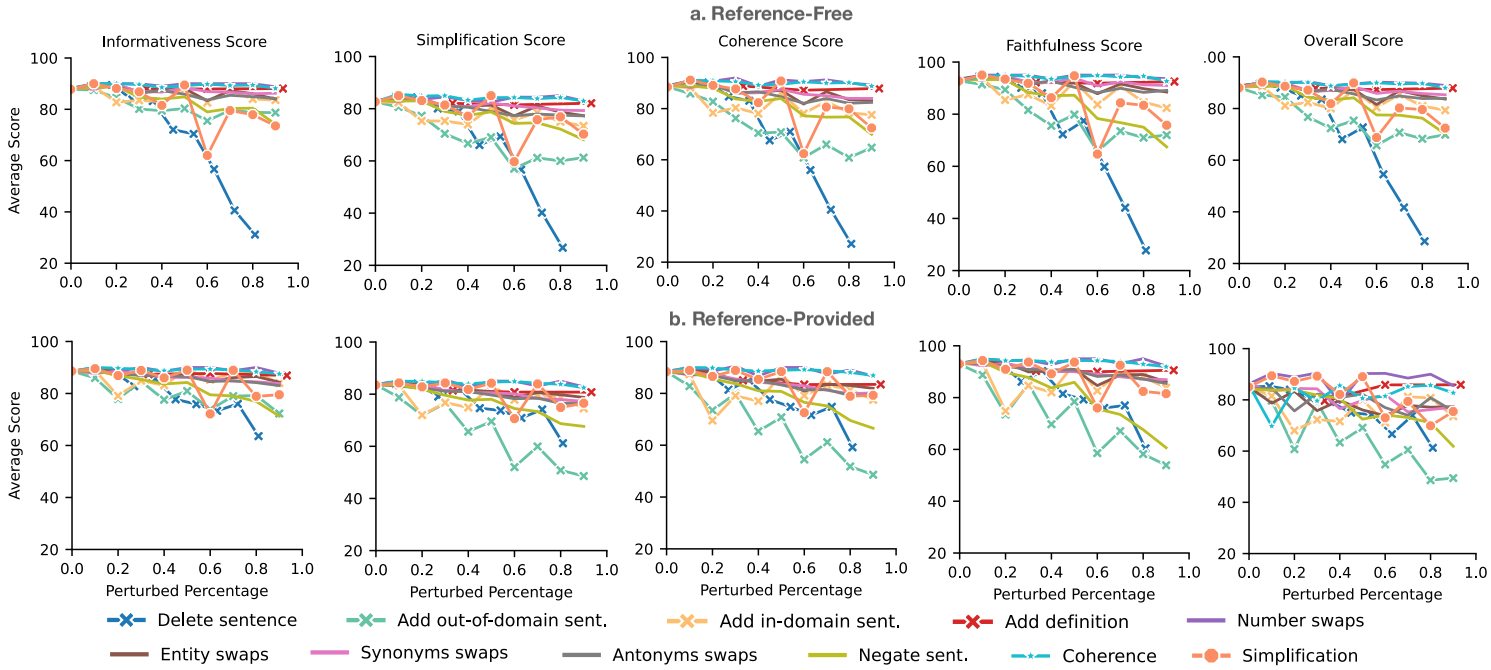


Figure A.7: Prompt-based evaluation scores for four criteria - informativeness, simplification, coherence, and faithfulness - along with an overall score. (a): Reference free; (b) Reference provided. Notably, prompt-based scores exhibit a reverse correlation with simplification perturbation (i.e., scores diminish as text simplifies) and demonstrate insensitivity towards coherence and faithfulness perturbations, except in instances of sentence negation.

We use GPT-3 for LLM evaluation. The generation process is configured with a temperature parameter of 0, a maximum length of 100, and a penalty value of 0. For each input, the top-ranked text is selected as the GPT-simplified output. Example prompts used for evaluation are provided in Figure A.6.

Figure A.7 shows the results for GPT-3 LLM evaluation, for both the reference-free and reference-provided settings. Though the evaluation is sensitive to some perturbations (deletion, addition, negation), it is insensitive to other perturbations (coherence, swaps) and sensitive to simplification in the inverse direction as would be expected (simplification score drops when more source text is replaced by simplified text). Additionally, the LLM evaluation is generally unable to distinguish between the four criteria, as most perturbations lead to the same score trends for simplification, coherence, faithfulness, and to a lesser degree informativeness. These patterns are similar to those observed in the overall score, indicating that the LLM evaluation as performed is not useful for providing facet-based judgments.

We also observe that in the reference-provided setting, scores for some perturbations are much higher (e.g., deletion) while others are much lower (e.g., add out-of-domain) than in the reference-free setting. The lack of a reference point or a way to normalize these scores makes it impossible to compare them across settings or datasets.

A.5 Additional perturbation results for PLABA

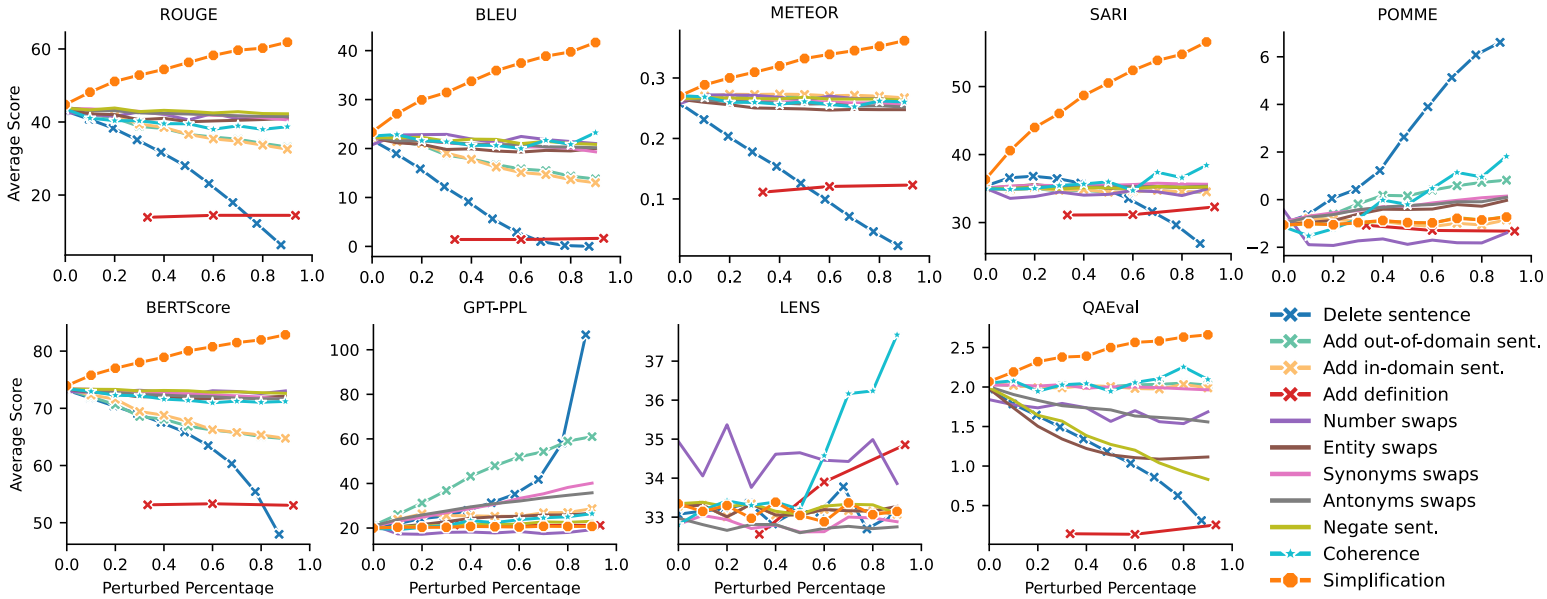


Figure A.8: Average scores of existing metrics and newly developed POMME score for perturbed texts in the PLABA dataset. Scores are averaged in 10 bins by perturbation percentage. Markers denote perturbations associated with our four defined criteria.

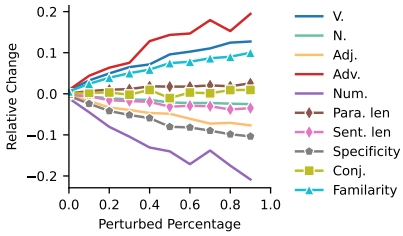


Figure A.9: Relative change of each lexical feature with respect to perturbations in the PLABA dataset. Different markers represent lexical feature categories.

We present full perturbation results on PLABA (Attal et al., 2023) in Figure A.8. The trends for many perturbations are in the same direction as in CELLS. While many metrics now show a desirable reversed trend to simplification (increasing), we point out that this is inconsistent performance relative to CELLS and is due to the high n-gram overlap between the hypothesis and targets in this case (we perturb by replacing source sentences with round-trip translated target sentences to form hypotheses, which only introduces minor lexical variation). Adding text, especially definitions, dramatically decreases many of these metrics due to the similar lengths of source and target texts in PLABA, again pointing to the n-gram and length sensitivities of most of these metrics.

The impact of simplification perturbations on lexical features in the PLABA dataset is shown in Figure A.9. Most trends are similar to CELLS, though paragraph length increases with higher perturbation percentage. In PLABA's target construction scheme, the target simplified texts are slightly longer than the source abstracts.

A.6 POMME *score for Llama- and Claude-simplified text*

In addition to using GPT-3 (Brown et al., 2020a) to produced simplified text for the simplification perturbation, we also test two other LLMs: Llama 2 (Touvron et al., 2023) and Claude. In Figure A.10, we show that POMME score changes when perturbing using the simplified text generated by all three models. Similar score changes are observed for all three models, demonstrating that POMME is consistently responsive to text simplicity and

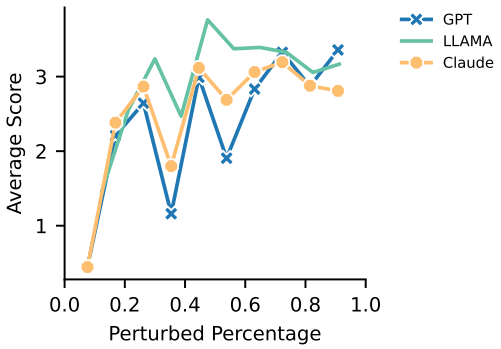


Figure A.10: Variation in POMME scores for simplification perturbations created by GPT-3, Llama2, and Claude on the CELLS dataset.

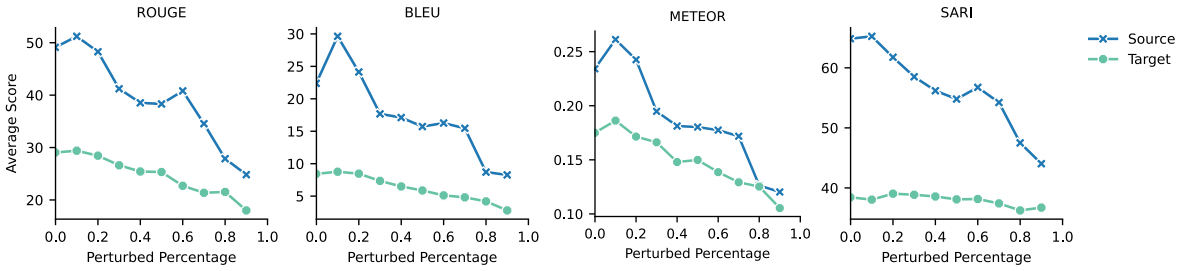


Figure A.11: Average scores of ROUGE, BLEU, METEOR, and SARI scores calculated using either the source text (complex) or target text (simple) as reference for simplification perturbations on the CELLS dataset. A metric sensitive to text simplicity should move in opposing directions under these two settings. However, metrics decrease uniformly in both settings, suggesting that they are not sensitive to text simplicity.

not specifically to the characteristics of GPT-simplified text.

A.7 Reversing source and target texts for simplification perturbation

To illustrate that existing metrics are not sensitive to text simplicity but rather to length and n-gram overlap, we present metric scores computed when swapping source and target for simplification perturbations (Figure A.11). When target text is used as reference, we start

with the oracle extractive hypothesis and increase perturbation percentage by swapping in simpler text, going from more complex to more simple text. When source text is used as reference, we reverse the original source and target, starting with simple text and swapping in the oracle extractive hypothesis, thereby moving from more simple to more complex text. A metric sensitive to text simplification should move in opposite directions in these two settings as perturbation percentage increases. However, these metric scores uniformly decrease under both settings, regardless of the reference, demonstrating that these metrics are not responsive to simplification but more so to text length and n-gram overlap. We do not report performance of BERTScore and QAEval under this setting due to the higher cost of computing these model based metrics.

Appendix B

SUPPLEMENTARY MATERIALS FOR PERSONALIZED JARGON

B.1 Initial Task Definition

For our initial study, we focus on reading full scientific abstracts because we wanted to understand if term familiarity was an important part of our broader envisioned setting of researchers reading abstracts outside of their domain. Further, our goal in providing a naive personalized abstract was to probe what transformations are feasible with current models that researchers respond positively to.

One abstract was from the medical domain, a domain not within any participant's dominant expertise, and one abstract was from a domain related to but distinct from a participant's specialization (i.e., psychology or linguistics). Personalized abstracts were generated using *text-davinci-003* model from OpenAI (OpenAI, 2023b; Brown et al., 2020a).

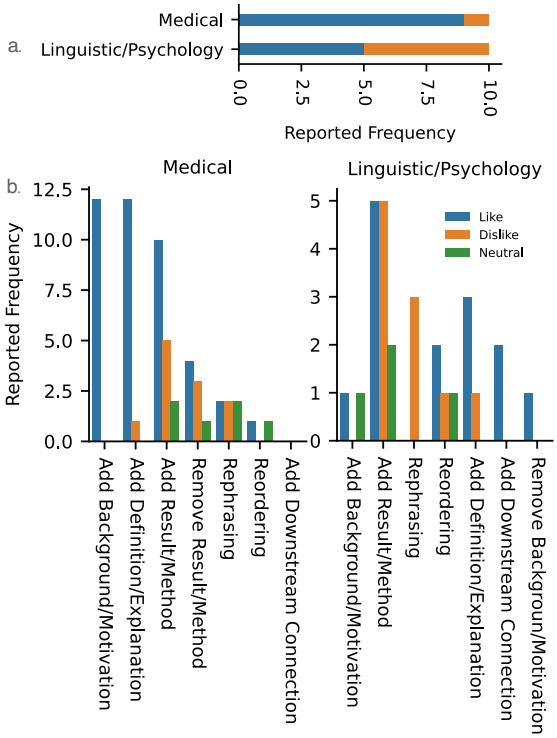


Figure B.1: Formative study results: domain-specific attitudes towards (a) personalized abstracts and (b) transformations of personalized abstracts. Medical abstracts are distant from the annotators' background, whereas linguistic or psychology abstracts are closer. Legend is shared for both plots.

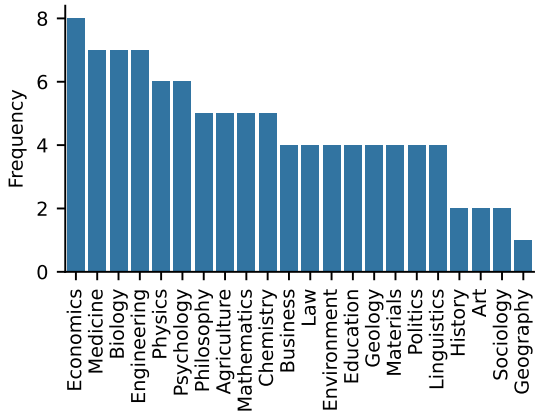


Figure B.2: Number of papers categorized by domain in the annotation dataset.

Modification	Original abstract sentence	Personalized abstract sentence	Human annotation
Add Background/ Motivation	Undiagnosed chronic kidney disease (CKD) is a common and usually asymptomatic disorder that causes a high burden of morbidity and early mortality worldwide.	Early detection of CKD is crucial for avoiding renal replacement therapy, **with an estimated 4.6% of total mortality worldwide**	*Like. I didn't realize how big the problem was, now I do.*
Add Definition/ Explanation	Distributional semantics provides multidimensional, graded, empirically induced word representations that successfully capture many aspects of meaning in natural languages, as shown by a large body of research in computational linguistics; yet, its impact in theoretical linguistics has so far been limited.	Distributional semantics is based on the Distributional Hypothesis, which states that similarity in meaning results in similarity of linguistic distribution.	*Like. Good to introduce what distributional semantics is.*
Add Result/ Method	We developed a deep learning model for CKD screening from routinely acquired ECGs.	We also used **Local Interpretable Model-agnostic Explanations (LIME)** to identify which ECG segments were particularly used in the identification of CKD, which focused mostly on QRS complexes and PR intervals.	*Marking as neutral because this isn't in the abstract so I'm unsure if this is right. But if it is, I really like that it's highlighting more of the CS methodological stuff.*
Rephrasing	Undiagnosed chronic kidney disease (CKD) is a common and usually asymptomatic disorder that causes a high burden of morbidity and early mortality worldwide.	Chronic kidney disease (CKD) is a **major global health burden**.	*Like. simpler language.*
Reordering	Distributional semantics provides multidimensional, graded, empirically induced word representations that successfully capture many aspects of meaning in natural languages, as shown by a large body of research in computational linguistics; yet, its impact in theoretical linguistics has so far been limited.	This survey provides an overview of the literature on distributional semantics, with a focus on methods and results that are of relevance for theoretical linguistics. Distributional semantics is based on the Distributional Hypothesis, which states that similarity in meaning results in similarity of linguistic distribution.	*Like. Not new but re-ordered; I like having this sentence at the start, as the first sentence in the original abstract is long and complicated.*
Add Downstream Connection	-	The paper also discusses how distributed representations can be used to generate dictionary definitions of words, and how they can be augmented with simple mathematical or logical expressions.	*Like. These additional details help make the study more concrete, and I understand them because this is my area. Also this connects to work I especially know because I've done some myself (generating dictionary definitions).*

Table B.1: Example sentence pair for the original abstract and personalized abstract. Human annotation results for their attitudes and reasons towards the modifications. The personalized abstract was generated using the abstract personalization prompt in Table B.2.

B.2 GPT Prompts

Task	Prompt	Model	Max length	Temp.
Abstract personalization	You are tasked with the role as a scientific writer to generate an personalized abstract for an individual reader. To do this effectively, consider including relevant background information or motivations related to the subject matter, provide necessary definitions or explanations to elucidate complex concepts, and incorporate significant methodological or result-oriented details. However, ensure that any additional information included is directly relevant and can be traced back to the provided content. The paper needed to be personalized is:{ }; The reader's publications are: { }; The reader's references are:{}; The personalized abstract is: .	text-davinci-003	500	0
Top-10 significant terms	Please review the following scientific paper abstract. Your task is to identify all scientific-related word/phrases within the text and then rank these word/phrases in descending order based on their significance within the abstract itself. Retain the first 10 word/phrases:.	text-davinci-003	100	0
Familiarity classification	Your job is to estimate how much the reader knows about an entity. You will be provided with the entity, the abstract where the entity came from, and related data about either the reader or the abstract. Entity: {} Abstract:{} Related Data:{} Here's how to gauge the reader's familiarity: - 0: The reader knows this subject well and can describe it to others. - 1: The reader has either encountered this subject before but knows little about it, or has never come across it at all. Based on the information provied, determine the familiarity score, either 0 or 1:	gpt-4	100	0
Definition needs classification	Your job is to estimate whether the reader might need an additional definition to fully grasp the entities mentioned in a given abstract. You will be provided with the entity, the abstract where the entity came from, and related data about either the reader or the abstract. Definition of definition/explanation: provides key information on the term independent of any context (e.g., a specific scientific abstract). A definition answers the question, Ẅhat is/are [term]?̈ Entity: {} Abstract:{} Rel:{}. Provide the estimation whether additional information is needed in a list in the order of the entity. The estimation should be either 0(no) or 1(yes). No need to mention the entity:	gpt-4	100	0
Background needs classification	Your job is to estimate whether the reader might need additional background to fully grasp the entities mentioned in a given abstract. You will be provided with the entity, the abstract where the entity come from, and related data about either the reader or the abstract. Definition of background/motivation: introduces information that is important for understanding the term in the context of the abstract. Background can provide information about how the term relates to overall problem, significance, and motivation of the abstract. Entity: {} Abstract:{} Rel:{}. Provide the estimation whether additional information is needed in a list in the order of the entity. The estimation should be either 0(no) or 1(yes). No need to mention the entity:	gpt-4	100	0
Example needs classification	Your job is to estimate whether the reader might need additional example to fully grasp the entities mentioned in a given abstract. You will be provided with the entity, the abstract where the entity come from, and related data about either the reader or the abstract. Definition of example: offers specific instances that help illustrate the practical application or usage of the term within the abstract. Entity: {} Abstract:{} Rel:{}. Provide the estimation whether additional information is needed in a list in the order of the entity. The estimation should be either 0(no) or 1(yes). No need to mention the entity:	gpt-4	100	0

Table B.2: GPT-4 prompts and configurations.

All prompts used for GPT-4 experiments are shown in Table B.2.

Model	F1	Recall	Precision
Majority Baseline	$44.4_{\pm 1.8}$	$100.0_{\pm 0.0}$	$28.6_{\pm 1.5}$
Oracle			
Majority	$50.7_{\pm 2.8}$	$42.7_{\pm 2.9}$	$62.3_{\pm 3.4}$
Nearest-neighbor	$54.3_{\pm 2.4}$	$60.4_{\pm 2.9}$	$49.3_{\pm 2.7}$
Lasso			
Mixed	$13.7_{\pm 3.6}$	$24.0_{\pm 6.2}$	$9.6_{\pm 2.7}$
Individual	$56.4_{\pm 2.8}$	$45.5_{\pm 3.0}$	$74.1_{\pm 3.4}$
GPT			
Baseline	$47.7_{\pm 1.8}$	$97.3_{\pm 1.0}$	$31.6_{\pm 1.6}$
Self-defined Subfield	$48.4_{\pm 1.9}$	$92.9_{\pm 1.6}$	$32.7_{\pm 1.6}$
Context-enhanced	$47.5_{\pm 1.8}$	$97.0_{\pm 1.1}$	$31.5_{\pm 1.6}$

Table B.3: Mean model performance ($\pm$std) on **additional definition** prediction in the test set (200 entities). The **bold** value indicates the best performing model for each category. Standard deviation is calculated by bootstrapping.

Model	F1	Recall	Precision
Majority Baseline	$37.2_{\pm 1.7}$	$100.0_{\pm 0.0}$	$22.8_{\pm 1.3}$
Oracle			
Majority	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Nearest-neighbor	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Lasso			
Mixed	$15.0_{\pm 5.1}$	$10.5_{\pm 3.8}$	$26.3_{\pm 8.7}$
Individual	$48.1_{\pm 3.3}$	$36.9_{\pm 3.2}$	$69.3_{\pm 4.0}$
GPT			
Baseline	$40.0_{\pm 1.9}$	$95.5_{\pm 1.3}$	$25.3_{\pm 1.5}$
Self-defined Subfield	$39.3_{\pm 1.9}$	$93.1_{\pm 1.7}$	$24.9_{\pm 1.5}$
Context-enhanced	$38.6_{\pm 1.9}$	$96.9_{\pm 1.2}$	$24.1_{\pm 1.4}$

Table B.4: Mean model performance ($\pm$std) on **additional background** prediction in the test set (200 entities). The **bold** value indicates the best performing model for each category.

Model	F1	Recall	Precision
Majority Baseline	$31.5_{\pm 1.8}$	$100.0_{\pm 0.0}$	$18.7_{\pm 1.2}$
Oracle			
Majority	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Nearest-neighbor	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Lasso			
Mixed	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$
Individual	$28.7_{\pm 3.9}$	$18.0_{\pm 2.8}$	$71.7_{\pm 6.6}$
GPT			
Baseline	$31.9_{\pm 1.8}$	$97.4_{\pm 1.2}$	$19.1_{\pm 1.2}$
Self-defined Subfield	$32.4_{\pm 1.8}$	$91.3_{\pm 2.1}$	$19.7_{\pm 1.3}$
Context-enhanced	$32.5_{\pm 1.8}$	$97.4_{\pm 1.1}$	$19.5_{\pm 1.3}$

Table B.5: Mean model performance ($\pm$std) on **additional example** prediction in the test set (200 entities). The **bold** value indicates the best performing model for each category.

Annotator		**Familiarity**		**Definition**		**Background**		**Example**	
ID	Sub-field	Neighbor	Sub-field	Neighbor	Subfield	Neighbor	Subfield	Neighbor	Subfield
1	CV	10	CV	5	Security	8	NLP	4	NLP
2	Networking	3	NLP	9	Networks	3	NLP	11	CV
3	NLP	11	CV	4	NLP	11	CV	2	Networking
4	NLP	11	CV	3	NLP	11	CV	1	CV
5	Security	2	Networking	2	Networking	11	CV	2	Networking
6	Theory	3	NLP	9	Networks	3	NLP	9	Networks
7	NN	8	NLP	9	Networks	4	NLP	10	CV
8	NLP	11	CV	3	NLP	11	CV	11	CV
9	Networks	7	NN	2	Networking	4	NLP	6	Theory
10	CV	3	NLP	3	NLP	8	NLP	5	Security
11	CV	3	NLP	4	NLP	8	NLP	2	Networking

Table B.6: Nearest-neighbor for each annotator.

Entity sentence	Domain	Ann. subdomain	Fam.	Information	Total Fam.
We study auctions for selling a limited supply of a single **commodity** in the case where the supply is known in advance and the case it is unknown and must be instead allocated in an online fashion.	Economics	CS Theory	Familiar	Example	9
Inference with the graphical model for de novo peptide sequencing estimates **posterior probabilities** for amino acids rather than scores for single symbols in the sequence.	Engineering	Computer Vision	Familiar	Example	7
... to communicate foreign policy goals and decisions, construct a **strategic narrative** of Indian foreign policy and counter narratives inimical to Indian interests.	Political Science	NLP	Familiar	None	6
Several parameters obtained from the experimental results were compared and analyzed, including the **load-bearing capacity**, stiffness, ductility, energy dissipation, and failure characteristics of the specimens.	Engineering, Materials Science	CS theory	Unfamiliar	Definition	5
Agents with identical **linear time-invariant dynamics** are considered.	Mathematics	Neural Networks	Familiar	None	2
Self-directed learning is a necessary skill for students and workers to remain lifelong learners.	Education	NLP	Unfamiliar	Definition, Motivation	9

Table B.7: Sample of entities with sentence context, their familiarity ratings, and information needs. Entities are bolded within the sentence. Total familiarity count indicates how many annotators rated an entity as familiar.