Investigating Large Language Models and Control Mechanisms to Improve Text Readability of Biomedical Abstracts

Zihao Li [†,1] Samuel Belkadi [†,2] Nicolo Micheletti [†,2]
Lifeng Han[*2] Matthew Shardlow[1] Goran Nenadic[2]
[1] Manchester Metropolitan University [2] The University of Manchester

zihao.li@stu.mmu.ac.uk
{samuel.belkadi, nicolo.micheletti}@student.manchester.ac.uk
{lifeng.han,g.nenadic}@manchester.ac.uk M.Shardlow@mmu.ac.uk
[†] *Co-first Authors* * *Corresponding Author*

Abstract—**Biomedical literature often uses complex language and inaccessible professional terminologies. That is why simplification plays an important role in improving public health literacy. Applying Natural Language Processing (NLP) models to automate such tasks allows for quick and direct accessibility for lay readers. In this work, we investigate the ability of state-of-the-art large language models (LLMs) on the task of biomedical abstract simplification, using the publicly available dataset for plain language adaptation of biomedical abstracts (PLABA). The methods applied include domain fine-tuning and prompt-based learning (PBL) on: 1) Encoder-decoder models (T5, SciFive, and BART), 2) Decoder-only GPT models (GPT-3.5 and GPT-4) from OpenAI and BioGPT, and 3) Control-token mechanisms on BART-based models. We used a range of automatic evaluation metrics, including BLEU, ROUGE, SARI, and BERTScore, and also conducted human evaluations. BART-Large with Control Token (BART-L-w-CT) mechanisms reported the highest SARI score of 46.54 and T5-base reported the highest BERTScore 72.62. In human evaluation, BART-L-w-CTs achieved a better simplicity score over T5-Base (2.9 vs. 2.2), while T5-Base achieved a better meaning preservation score over BART-L-w-CTs (3.1 vs. 2.6). We also categorised the system outputs with examples, hoping this will shed some light for future research on this task. Our codes, fine-tuned models, and data splits from the system development stage will be available at https://github.com/HECTA-UoM/PLABA-MU**

Index Terms—**Large Language Models, Text Simplification, Biomedical NLP, Control Mechanisms, Health Informatics**

I. INTRODUCTION

The World Health Organization (WHO) defines *health literacy* as: "the personal characteristics and social resources needed for individuals and communities to access, understand, appraise, and use information and services to make decisions about health" [1]. From this, the National Health Service (NHS) of the UK emphasises two key factors for achieving better health literacy [1], i.e., the individual's comprehension ability and the health system itself. The "system" here refers to the complex network of health information and sources which promote it.

These two factors are codependent. For instance, professionals write much healthcare information using complex language and terminologies without considering the readability of patients and the public in general. The health system must take into account the patient's ability to achieve health literacy. Scientific studies have reported a correlation between low health literacy, poorer health outcomes, and poorer use of health care services [2], [3]. Thus, Plain Language Adaptation (PLA) of scientific reports in the healthcare domain is valuable for knowledge transformation and information sharing for public patients so as to promote public health literacy [4]. Nowadays, there have been industrial practices on such tasks, which include the publicly available plain summaries of scientific abstracts from the American College of Rheumatology (ACR) Virtual Meeting 2020 offered by the medicine company Novartis.com [2].

The PLA task is related to text simplification and text summarisation, which are branches of the Natural Language Processing (NLP) field. This work investigates the biomedical domain PLA (BiomedPLA) using state-of-the-art large language models (LLMs) and control token methods [5]–[8] that have been proven to be effective in such tasks. Examples of BiomedPLA can be seen in Figure 1 from the PLABA2023 shared task[3]. We highlight some of the factors in colours of such tasks, including sentence simplification in grey (removing clause "which" in the first sentence example; separating into two sentences and removing bracket for the second example), term simplification in yellow ("pharyngitis" and "pharynx" into "throat"), paraphrasing and synonyms in green (e.g. "acute" into "sore" and "posterior" into "back" for synonyms), and summarisation (on overall text in certain situations). The LLMs we applied include advanced Encoder-Decoder models (T5, SciFive, BART) and Generative Pre-trained Transformers (BioGPT, ChatGPT). The methodologies we applied include fine-tuning LLMs, prompt-based learning (PBL) on GPTs, and control token mechanisms on LLMs (BART-base and BART-

large) with the efficient fine-tuning strategy. Using the publicly available PLABA (Plain Language Adaptation of Biomedical Abstracts) data set from [9], we demonstrate the capabilities of such models and carry out both quantitative and human evaluations of the model outputs. We also discuss the interesting findings from different evaluation metrics, their inconsistency, and future perspectives on this task, based on our primary work [10].

II. RELATED WORK

We first introduce recent developments in biomedical text simplification, then extend to broader biomedical LLMs related to this paper, followed by efficient training methodologies which we will apply to our work.

A. Biomedical Text Simplification

To improve the health literacy level for the general public population, [11] developed the first lay language summarisation task using biomedical scientific reviews. The key points for this task include an explanation of context knowledge and expert language simplification. The evaluation included quality and readability using quantitative metrics. [12] carried out a survey, up to 2021, on biomedical text simplification methods and corpora using 45 relevant papers on this task, which data covers seven natural languages. In particular, the authors listed some published corpora on English and French languages and divided them into comparable, nonparallel, parallel, thesaurus, and pseudo-parallel. The quantitative evaluation metrics mentioned in these papers include SARI, BLEU, ROUGE, METEOR, and TER, among which three of them are borrowed from the machine translation (MT) field i.e., BLEU, METEOR, and TER [13]. Very recently, [14] transferred lay language style from expert physicians' notes. They developed a comparable dataset from many non-parallel corpora on plain and expert texts. The baseline model applied for training is BART, with positive outcomes. [15] did a case study using chatGPT models on translating radiology reports into plain language for patient education. Detailed prompts were discussed to mitigate the GPT models to reduce "over-simplified" outputs and "neglected information".

B. Broader Biomedical LLMs

Beyond text simplification tasks, there have been active developments towards biomedical domain adaptation of LLMs in recent years. For instance, BioBERT used 4.5B words from PubMed and 13.5B words from PMC to do continuous learning based on the BERT pre-trained model. Then, it is fine-tuned in a task-specific setting on the following tasks: NER, RE, and QA [16]. In comparison, BioMedBERT [17] created new data sets called BREATHE using 6 million articles containing 4 billion words from different biomedical literature, mainly from NCBI, Nature Research, Springer Nature, and CORD-19, in addition to BioASQ, BioRxiv, medRxiv, BMJ, and arXiv. It reported better evaluation scores on QA data sets, including SQuAD and BioASQ, among other tested tasks, compared to other models.

BioMedLM 2.7B developed by Stanford Center for Research on Foundation Models (CRFM) and Generative AI company MOSAIC ML team [4] formerly known as PubMedGPT 2.7B [5] is trained on biomedical abstracts and papers using data from The Pile [18]. BioMedLM 2.7B claimed new state-of-the-art performance on the MedQA data set.

BioALBERT from [19] is based on the ALBERT [20] structure for training using biomedical data and reported higher evaluation scores on NER tasks of Disease, Drug, and Species, on several public data sets but much shorter training time in comparison to BioBERT. BioALBERT was also tested on broader BioNLP tasks using its different base and large models, including RE, Classification, Sentence Similarity, and QA by [21]

Afterwards, based on the T5 model structure [22], SciFive [23] was trained on PubMed Abstract and PubMed Central (PMC) data and claimed new state-of-the-art performance on biomedical NER and RE and superior results on BioAsq QA challenge over BERT and BioBERT. Similarly, BioBART [24] was developed recently based on the new learning structure BART model [25]. This work will examine T5, SciFive, and BART, leaving BioBART as a future work.

Other notable related works include a) the model comparisons in biomedical domains with different tasks by [26]–[28] on BERT, ALBERT, ELECTRA, PubMedBERT, and PubMedELECTRA; b) task- and domain-specific applications on QA by [29], on Medicines by [30], on radiology (RadBERT) by [31], concept normalisation by [32], abstract generation by [33]; c) language specific models such as in French [34] and Turkish [35]; and d) survey work by [36].

C. Efficient Training

Due to the computational cost of the extra large-sized PLMs, some researchers proposed efficient training, which factor we will also apply to our study. These include some previously mentioned works.

[37] proposed Parameter-Efficient Transfer Learning for NLP tasks using their Adapter modules. In this method, the parameters of the original PLMs are fixed, and a few trainable parameters are added for each fine-tuning task, between 2-4 % of the original parameter sizes. Using GLUE benchmark data, they demonstrated that efficient tuning with the Adapter modules can achieve similar high performances compared to the BERT models with full fine-tuning of 100% parameters. ALBERT [20] applied parameter reduction training to improve the speed of BERT model learning. The applied technique uses a factorisation of the embedding parameters, which are decomposed into smaller-sized matrices before being projected into the hidden space. They also designed a self-supervised loss function to model the inner-sentence coherence. This reduced the parameter sizes from 108M in the BERT base to 12M in the ALBERT base models. Addressing similar issues, [38] proposed *Prefix-tuning* method, which modifies only 0.1% of the full

---

[4]https://www.mosaicml.com/
[5]https://huggingface.co/stanford-crfm/BioMedLM

INPUT (ABSTRACT)	OUTPUT (ADAPTED)
Acute pharyngitis/tonsillitis, which is characterized by inflammation of the posterior pharynx and tonsils, is a common disease.	Sore throat/tonsillitis, or when the back of the throat or tonsils is inflamed, is common.
Several viruses and bacteria can cause acute pharyngitis; however, Streptococcus pyogenes (also known as Lancefield group A β-hemolytic streptococci) is the only agent that requires an etiologic diagnosis and specific treatment.	Many viruses and bacteria can cause short-term sore throat. However, group A strep, caused by Group A strep bacteria, is the only cause that must be identified based on signs and symptoms and treated.

Sentence structure simplification   Term simplification   Paraphrase / synonyms

Fig. 1: Examples from the PLABA dataset on Biomedical Sentences Adaptation.

parameters to achieve comparable performances using GPT-2 and BART for table-to-text generation and summarisation tasks.

Focus on the biomedical domain, [28] carried out fine-tuning stability investigation using the BLURB data set (Biomedical Language Understanding and Reasoning Benchmark) from [39]. Their findings show that freezing lower-level layers of parameters can be helpful for BERT-based model training, while re-initialising the top layers is helpful for low-resource text similarity tasks.

Instead of using Adapter modules [37] that require additional inference latency, [40] introduced Low-Rank Adaption (LoRA) that further reduces the size of trainable parameters by freezing the weights in PLMs and injects "trainable rank decomposition matrices" into every single layer of the Transformer structure for downstream tasks. The experiments were carried out on RoBERTa, DeBERTa, and GPTs that showed similar performances compared to the Adapter modules. We will apply LoRA for efficient fine-tuning on T5 and BioGPT in our work.

III. METHODOLOGIES

The overall framework of our experimental design is displayed in Figure 2. In the first step, we fine-tune selected LLMs including T5, SciFive, BioGPT, and BART, apply prompt-based learning for ChatGPTs, and optimise control mechanisms on the BART model. Then, we select the best performing two models using quantitative evaluation metrics SARI, BERTScore, BLEU and ROUGE. Finally, we chose a subset of the testing results of the two best-performing models for human evaluation.

A. Models

The models we investigated in our work include T5, SciFive, GPTs, BioGPT, BART, and Control Mechanisms; we will give more details below.

1) T5: T5 [22] used the same Transformer structure from [41] but framed the text-to-text learning tasks using the same vocabulary, sentence piece tokenisation, training, loss, and decoding. The pre-fixed tasks include summarisation, question answering, classification, and translation. The authors used the common crawl corpus and filtered to keep only natural text and de-duplication processing. They extracted 750GB of clean English data to feed into the model for multi-task pre-training. Different masking strategies are integrated into the T5 model to facilitate better performances of specific fine-tuning

tasks. It has demonstrated state-of-the-art results across a wide spectrum of natural language processing tasks, showcasing its remarkable capabilities in capturing nuanced semantics and generating simplified texts while upholding high levels of accuracy. Notably, it has been successfully employed in various fields such as Clinical T5 by [42] and [43].

In this paper, we fine-tuned three versions of T5, namely t5-small, t5-base, and t5-large, paired with their sentence-piece pre-trained tokenizer. Each is fine-tuned independently on the same dataset as the other models to provide comparable results. Note that we use the prompt "summarize:" as it is the closest to our task.

2) SciFive: Using the framework of T5, SciFive is a Large Language Model pre-trained on the biomedical domain and has demonstrated advanced performances on multiple biomedical NLP tasks [23].

Similarly to our work on T5, we fine-tuned two versions of SciFive, namely SciFive-base and SciFive-large, paired with their pre-trained tokenizer. Each is fine-tuned independently on the same dataset as the other models to provide comparable results. We again use the prompt "summarize:" for task-relation purposes.

3) OpenAI's GPTs: Given the remarkable performance demonstrated by OpenAI's GPT models in text simplification [44], we decided to apply simplifications using GPT-3.5-turbo and GPT-4 via its API[6], both models were acc.

4) BioGPT: BioGPT [45] is an advanced language model specifically designed for medical text generation. BioGPT is built upon GPT-3 but is specifically trained to understand medical language, terminology, and concepts. BioGPT follows the Transformer language model backbone and is pre-trained on 15 million PubMed abstracts. It has demonstrated a high level of accuracy and has great potential for applications in medicine. BioGPT is fine-tuned on the training and validation set, as with other encoder-decoder models (Figure 2).

5) BART: Like the default transformer structure, BART [25] aims to address the issues in BERT and GPT models by integrating their structures with a bi-directional encoder and an autoregressive decoder. In addition to the single token masking strategy applied in BERT, BART provides various masking strategies, including deletion, span masking, permutation, and

---

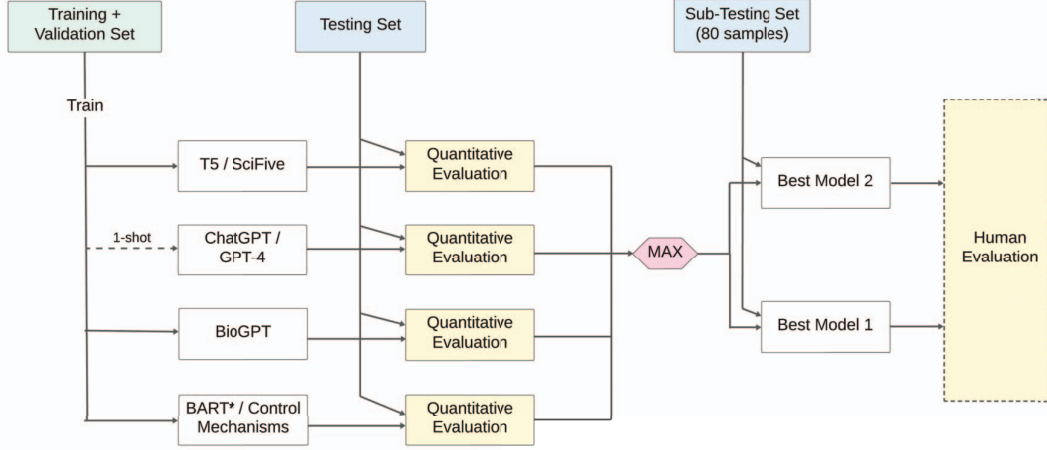[6]https://openai.com/blog/openai-api

Fig. 2: Model Development and Evaluation Pipeline. BART* is fine-tuned using Wikilarge data. MAX step chooses the two best-performing models according to the automatic evaluation results using SARI and BERTScore.

rotation of sentences. Compared to GPT models, BART provides both leftward and rightward context in the encoders.

6) Controllable Mechanisms: We applied the modified control token strategy in [8] for both BART-base and BART-large models. The training includes 2 stages, leveraging both Wikilarge training set [46] and our split of training set from PLABA [47].

The four attributes for control tokens (CTs) are listed below:
- <DEPENDENCYTREEDEPTH_x> (DTD)
- <WORDRANK_x> (WR)
- <REPLACEONLYLEVENSHTEIN_x> (LV)
- <LENGTHRATIO_x> (LR)

They represent 1) the syntactic complexity, 2) the lexical complexity, 3) the inverse similarity of input and output at the letter level, and 4) the length ratio of input and output respectively.

Before training, the four CTs are calculated and prepared for the 2 stage training sets. In both stages, we pick the best model in 10 epochs based on the training loss of the validation set. We applied the best model from the first stage as the base model and fine-tuned it on our PLABA training set for 10 epochs.

After fine-tuning, the next step is to find the optimal value of CTs. Following a similar process in MUSS [48], we applied Nevergrad [49] on the validation set to find the static optimal discrete value for DTD, WR, and LV. As for the LR, we applied the control token predictor to maximise the performance with the flexible value. The predictor is also trained on Wikilarge [46] to predict the potential optimal value for LR.

B. LoRA and LLMs

To evaluate bigger model architectures, we fine-tune FLAN-T5 XL [50] and BioGPT-Large, which have 3 billion and 1.5 billion parameters, respectively. FLAN-T5 XL is based on the pre-trained T5 model with instructions for better zero-shot and few-shot performance. To optimise training efficiency, and as our computational resources do not allow us to fine-tune the full version of these models, we employ the LoRA [40] technique, which allows us to freeze certain parameters, resulting in more efficient fine-tuning with minimal trade-offs.

C. Metrics

We decide to evaluate our models using four quantitative metrics, namely BLEU [51], ROUGE [52], SARI [53], and BERTScore [54], each offering unique insights into text quality. SARI and BERTScore are used from EASSE [55] package; BLEU and ROUGE metrics are imported from the Hugging Face[7] implementations.

While BLEU quantifies precision by assessing the overlap between n-grams in the generated text and references, ROUGE measures recall by determining how many correct n-grams in the references are present in the generated text. This combination makes them useful as an initial indicative evaluation for machine translation and summarisation quality.

In contrast, SARI goes beyond n-gram comparisons and evaluates fluency and adequacy in translations. It does this by considering precision (alignment with references), recall (coverage of references), and the ratio of output length to reference length. SARI's comprehensive approach extends its utility to broader evaluations of translation quality.

Finally, BERTScore delves into the semantic and contextual aspects of text quality. Using a pre-trained BERT model, it measures the similarity between word embeddings in the generated and reference texts. This metric provides insight

---

[7]https://github.com/huggingface/evaluate

into the semantic similarity and contextual understanding between generated and reference texts, making it akin to human evaluation. This metric does not quantify how good the simplification is but rather how much the meaning is preserved after simplification.

This comprehensive evaluation effectively addresses surface-level and semantic dimensions, resulting in a well-rounded and thorough assessment of the quality of machine-generated simplifications.

IV. EXPERIMENTS AND EVALUATIONS

PLABA data set is extracted from PubMed search results using 75 healthcare-related questions that MedlinePlus users asked. It includes 750 biomedical article Abstracts manually simplified into 921 adaptations with 7,643 sentence pairs in total. The dataset is publicly available via Zenodo [8].

A. Data Preprocessing and Setup

To investigate the selected models for training and fine-tuning, we divided the PLABA data into Train, Validation, and Test sets, aiming for an 8:1:1 ratio. However, in the final implementations, we found that there are only a few 1-to-0 sentence pairs, which might cause a negative effect in training the simplification models. Thus we eliminated all 1-to-0 sentence pairs. In addition, to better leverage the SARI score, we picked sentences with multi-references for validation and testing purposes. As a result, we ended up with the following sentence pair numbers according to the source sentences (5757, 814, 814).

B. Automatic Evaluation Scores

In this section, we list quantitative evaluation scores and some explanations for them. The results for T5 Small, T5 Base, T5 Large, FLAN-T5 XL with LORA, SciFive Base, SciFive Large, and BART models with CTs (BART-w-CTS) are displayed in Table I. Interestingly, the fine-tuned T5 Small model obtains the highest scores in both BLEU and ROUGE metrics including ROUGE-1, ROUGE-2, and ROUGE-L. The fine-tuned BART Large with CTs produces the highest SARI score at 46.54; while the fine-tuned T5 Base model achieved the highest BERTScore (72.62) with a slightly lower SARI score (44.10). The fine-tuned SciFive Large achieved the highest SARI score (44.38) among T5-like models, though it is approximately 2 points lower than BART Large with CTs.

The quantitative evaluation scores of GPT-like models are presented in Table II including GPT-3.5 and GPT-4 using prompts, and fine-tuned BioGPT with LoRA. GPT-3.5 reported relatively higher scores than GPT-4 on all lexical metrics except for SARI, and much higher score on BERTScore than GPT-4 (58.35 vs 46.99). In comparison, BioGPT-Large with LoRA reported the lowest SARI score (18.44) and the highest BERTScore (62.9) among these three GPT-like models. Comparing the models across Table I and Table II, the GPT-like models did not beat T5-Base on both SARI and BERTScore, and did not beat BART-w-CTs on SARI.
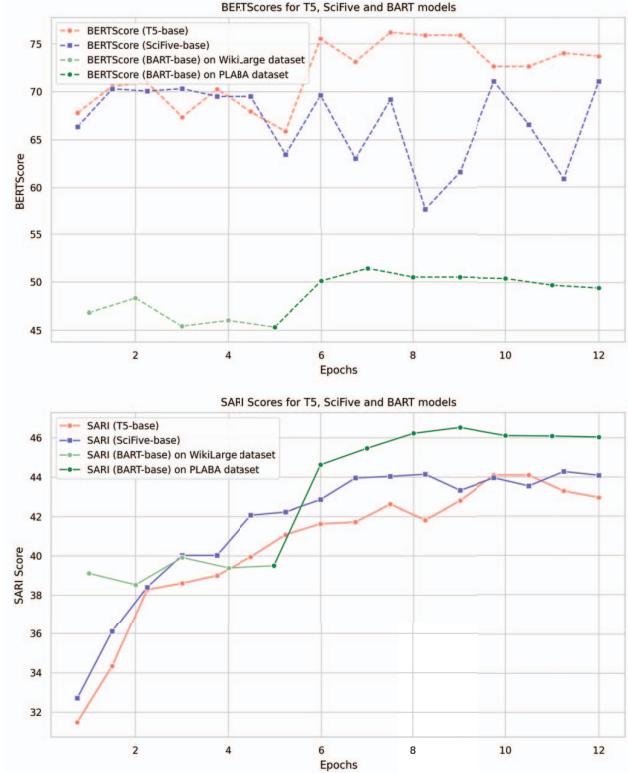
Fig. 3: Evaluation Scores of T5, SciFive and BART Models on the Extracted Testing Set

To look into the details of model comparisons from different epochs on the extracted testing set, we present the learning curve of T5, SciFive, BART-base on WikiLarge, and BART-base on PLABA data in Figure 3. We also present the learning curve of T5 Base and BART Base using different metrics in Figure 4.

Because the fine-tuned T5-Base model has the highest BERTScore (72.62) and also a relatively higher SARI score (44.10), we chose it as one of the candidates for human evaluation. The other candidate is the fine-tuned BART Large with CT mechanisms which has the highest SARI score (46.54) among all evaluated models. Note that, SciFive Large has results close to the ones of T5 Base. In this case, we selected the smaller model for human evaluation.

C. Human Evaluations

In human evaluation, we randomly sampled 80 sentences from the test set split and evaluated the corresponding outputs of BART-large with CTs and T5-base with anonymisation setting. In the human evaluation form, we randomly assigned the order of the two systems' outputs and put them beside the input sentence. Based on the comparison of input and output sentences, the annotators need to answer two questions: "To what extent do you agree the simplified sentence keeps the major information" and "To what extent do you agree the

Models	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	SARI	BERTScore
T5 Small	**49.86**	**65.94**	**48.60**	**63.94**	33.38	69.58
T5 Base	43.92	64.36	46.07	61.63	44.10	**72.62**
T5 Large	43.52	64.27	46.01	61.53	*43.70*	60.39
FLAN-T5 XL (LoRA)	44.54	63.16	45.06	60.53	43.47	67.94
SciFive Base	44.91	64.67	46.45	61.89	44.27	60.86
SciFive Large	44.12	64.32	46.21	61.41	*44.38*	72.59
BART Base with CTs	21.52	56.14	35.22	52.38	46.52	50.53
BART Large with CTs	20.71	54.73	32.64	49.68	**46.54**	50.16

TABLE I: Automatic Evaluations of Encoder-Decoder Models on Extracted Testing Set.

Models	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	SARI	BERTScore
GPT-3.5	20.97	50.07	24.72	43.12	42.61	58.35
GPT-4	19.50	48.36	23.34	42.38	43.22	46.99
BioGPT-Large (LoRA)	41.36	63.21	46.63	61.56	18.44	62.9

TABLE II: Quantitative Evaluations of GPTs and BioGPT-Large on the Extracted Testing Set
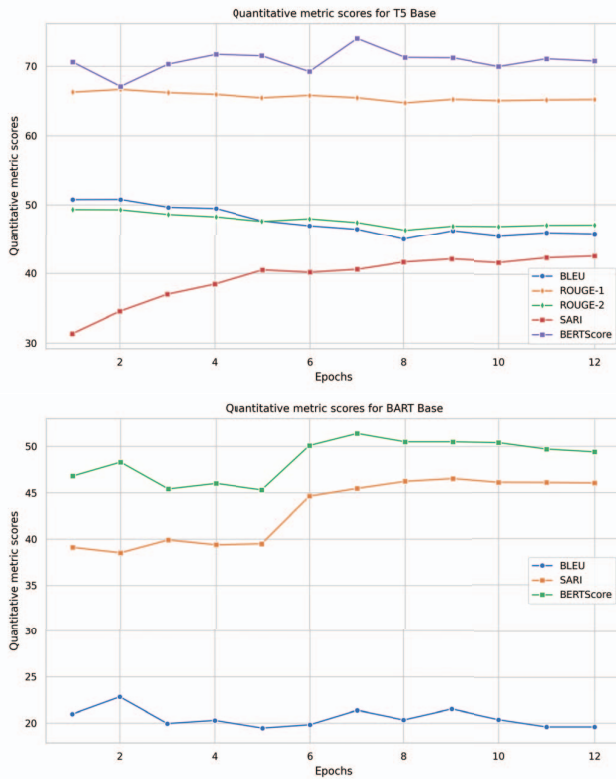

Fig. 4: Evaluation Scores of T5-base and BART-base on the Extracted Testing Set

annotators. The detailed human evaluation scores on the two selected models we evaluated are shown in Table III using the two designed criteria.

Based on the cross-annotation (overlaps), we calculated the inter-rater agreement level in Table IV using Cohen's Kappa on cross models and Table V using Krippendorff's Alpha with model-wise comparison. Both tables include the agreement levels on two sub-categories, namely "meaning preservation" and "text simplicity". Because there is no overlap in the annotation tasks between annotators (0, 3) and annotators (1, 2), we listed the agreement levels between the available pairs of annotations, i.e. (0, 1), (0, 2), (1, 3), and (2, 3). The Cohen's Kappa agreement level presented in Table IV shows the inter-rater agreement on the performance order of the two systems, whether one system is better than the other or tie. The Krippendorff's Alpha represented in Table V shows the annotation reliability over all 5-Likert options.

Based on the results from Table III, annotators show a different preference for the two systems. Despite the limited gap in the SARI score, BART with CTs shows a better capability to fulfil the simplification tasks. Yet regarding meaning preservation, the fine-tuned T5-base performs better, as the BERTScore tells among the comparisons in Table I.

From Table IV, Annotators 0 and 1 have the highest agreement on "meaning preservation" (score 0.583), while Annotators 1 and 3 have the highest agreement on "simplicity" (score 0.238) evaluation across the two models. This also indicates that it is not easy to evaluate the system performances against each other regarding these two criteria.

Model wise, Table V further shows that Annotators 0 and 1 agree more on the judgements of fine-tuned T5 Base model on both two criteria of "meaning preservation" (score 0.449) and "text simplicity" (score 0.386), in comparison to the BART model. On the fine-tuned BART Large model with CTs, these two annotators only agreed on the "meaning preservation" factor with a score of 0.441. This phenomenon also applies to Annotators 0 and 2 regarding the judgement of these two models. Interestingly, while Table IV shows that Annotators 1 and 3 have a better agreement on "simplicity" judgement over "meaning preservation" using Cohen's Kappa, Table V

simplified sentence is well simplified". The answer is limited to a 5-point Likert scale, from strongly disagree to strongly agree. There are 4 annotators in this human evaluation, one of the annotators is a native English speaker, and the others are fluent English users as a second language. Two annotators are final-year bachelor students, one annotator is a Masters candidate, and the last one is a Postdoctoral researcher. Each annotator evaluated 40 sentence pairs with 50% overlaps to make sure every sentence was evaluated twice by different

shows the opposite using Krippendorff's Alpha, i.e. it tells the agreement on "meaning preservation" of these two annotators instead. This shows the difference between the two agreement and reliability measurement metrics.

Model	Meaning Preservation	Simplicity
BART with CTs	2.625	2.900
T5-base	3.094	2.244

TABLE III: Human Evaluation Scores of Fine-tuned BART-Large with CTs and T5-Base Models on the Sampled Test Set

Annotator	Meaning preservation	Simplicity
0 & 1	0.583	0.138
0 & 2	0.238	0.126
1 & 3	0.008	0.238
2 & 3	-0.130	-0.014

TABLE IV: Cohen Kappa among annotators over 3 categories ordinal - win, lose, and tie.

Anno.	Model	Meaning preservation	Simplicity
0 & 1	T5-base	0.449	0.386
	BART w CTs	0.441	0.052
0 & 2	T5-base	0.259	0.202
	BART w CTs	0.200	0.007
1 & 3	T5-base	0.307	0.065
	BART w CTs	-0.141	-0.056
2 & 3	T5-base	-0.056	0.116
	BART w CTs	0.065	-0.285

TABLE V: Krippendorff's alpha among annotators (Anno.) over the 5-Likert scale from strongly agree to strongly disagree.

D. System Output Categorisation

We observe some interesting aspects of human evaluation findings by comparing the outputs of two models. 1) how to deal with the judgement on two models when one almost copied the full text from the source while the other did simplification but with introduced errors? 2) Abbreviation caused interpretation inaccuracy. This can be a common issue in the PLABA task. For instance, the source sentence "*A total of 157 consecutive patients underwent TKA (n = 18) or UKA (n = 139).*" is simplified by the T5-base model into "*A total of 157 consecutive patients underwent knee replacement or knee replacement.*" and by BART-w-CTs into "*A total of 157 patients had either knee replacement or knee replacement surgery.*" Both these two models produced repeated phrases "knee replacement or knee replacement" due to a lack of meaningful understanding of "TKA: total knee arthroplasty" and "UKA: Unicompartmental knee arthroplasty". A reasonable simplification here can be "157 patients had knee surgery".

We list more categories below by comparing outputs from the two evaluated models. These categories apply to some segment-level comparisons, but *not all the outputs*, e.g. in Figure 5.

- Both models simplified the abstract into the exact same output
- Both models produced hallucinations and similar outputs

- Both models cut half the sentence/meaning, but at different parts
- Both models cut complex sentences into multiple sentences, but BART adapted lay-language
- BART adapted using lay language but cut out some meaning
- BART generated simplification vs T5 generated nonsense
- T5 did little simplification but maintained good meaning.
- T5 cut meaning; BART did not but maintained the same complexity as the abstract
- BART shifted the meaning

V. DISCUSSIONS

Several aspects can be further improved or investigated:

- **On automatic metrics**. Based on the results depicted in Figure 4, we acknowledge that SARI stands out as a more reliable metric than BLEU and ROUGE-1/2 for assessing the quality of generated simplifications. During the early training epochs, the model outputs closely resemble the input texts, which can lead SARI to assign lower scores compared to BLEU and ROUGE-1/2. This occurs because these metrics might be satisfied by the mirroring of n-grams between inputs and generated outputs during simplification. However, it is essential to acknowledge that, in the context of simplification, the generated output typically remains relatively close to the input. As a result, BLEU and ROUGE may exhibit consistent scores throughout epochs and may not effectively evaluate the quality of the generated texts. In contrast, BERTScore offers a different perspective by focusing on meaning preservation after simplification instead of simplification quality. If the generated outputs are copies of the input texts, BERTScore may still yield high scores although the model indeed performs poorly. Therefore, we use the combination of both metrics — SARI for evaluating generation quality and BERTScore for assessing meaning preservation — in order to select the best-performing models. It would be very useful to develop a new automatic metric that can effectively reflect both text simplicity and meaning preservation.

- **On human evaluations**. Due to the various backgrounds of annotators and the lack of proper training material and methods, it is difficult to build a standardised scheme to help annotators choose from the 5-Likert options. This means that the definition of "Strongly Agree/Disagree" may vary among annotators. In addition, there is no method to normalise the scores to a unified level in order to avoid the effects caused by differences in subjective judgements. Thus, we evaluate the two systems simultaneously and allow the annotator to decide on the performance order and performance gap. To reflect on the agreement of these two aspects, we calculated the inter-rater agreement in Cohen Kappa and Krippendorff's alpha. As shown in Table IV, only annotators 0 and 1 show a decent agreement on meaning preservation, while the others show only limited agreement or even

Category	Abstract (Original)	Biomedical-T5-base	BART-L-w-CTs
Both models simplified the abstract into the exact same output	Vitamin B6 was associated with significantly greater reduction in TD symptoms compared to placebo (SMD = 1.41 ± 0.22; 95% CI, 0.98 to 1.85; z = 6.4; $P < .001$) in 2 trials conducted by the same research group.	Vitamin B6 was associated with significantly greater reduction in TD symptoms compared to sugar pills in 2 trials conducted by the same research group.	Vitamin B6 was associated with significantly greater reduction in TD symptoms compared to sugar pills in 2 trials conducted by the same research group.
Both models produced hallucinations and similar outputs	None of the studies observed the effects of oral versus depot antipsychotic drugs.	None of the studies observed the effects of antipsychotic drugs taken **by mouth or by mouth**.	None of the studies looked at the effects of antipsychotic drugs taken **by mouth or by mouth**.
Both models cut half the sentence/meaning, but at different parts	Angiotensin-converting enzyme (ACE) inhibitors and beta-blockers significantly decreased and hydrochlorothiazides significantly increased the plasma Hcy levels in hypertensive patients ($P < 0.05$).	Hydroxychlorothiazides significantly increased the blood Hcy levels in hypertensive patients.	Angiotensin-converting enzyme (ACE) inhibitors and beta-blockers significantly decreased blood Hcy levels in patients with high blood pressure.
Both models cut complex sentences into multiple sentences, but BART adapts lay-language	We **characterized** the content, dose, and progression of usual post-acute PT services following TKR, and examined associations of specific details of post-acute PT with patients' 6-month functional outcomes.	We characterized the content, amount, and progression of usual PT services following knee replacement surgery. We also examined links of specific details of PT with patients' 6-month functional outcomes following surgery.	We **looked at** the content, amount, and progression of PT services after knee replacement. We also looked at links between specific details of PT after surgery with patients' 6-month functional outcomes.
BART uses lay-language, T5 does not	The difference in muscle strength was **maintained** at 12-month follow-up.	The difference in muscle strength was maintained at 12-month follow-up appointments.	The difference in muscle strength was **still there** at the 12-month check-up.
BART adapts using lay-language, but cuts down some meaning	This review describes therapeutic applications of **antisense** techniques currently under investigation in CF.	This review summarizes therapeutic applications of antisense techniques currently being studied in CF.	This review describes **treatments** that may **help people with CF** that are being studied in the future.
BART generates simplification, T5 generates nonsense	Inflammation, including the breakdown of vitamin B6 in the body, could explain these findings.	UneoriUneoriUneori UneoriUneoriUneori[...]Uneori	Inflammatory processes, including vitamin B-6 catabolism, could explain such findings.

Fig. 5: Segment-level examples of categorised outputs from two models.

272

slight disagreement. Although human evaluation has long been the gold standard for evaluation tasks, there are few acknowledged works on a standard procedure to make it more explainable and comparable. In the future, a more standardised and unified process may be required.

- **On broader test sets**. In this work, we carried out an investigation using the PLABA data set. To be more fairly comparing selected models, we will include other related testing data on biomedical text simplification tasks. How to improve the number of references for the PLABA data set can be also explored.

VI. CONCLUSIONS AND FUTURE WORK

We have carried out an investigation into using LLMs and Control Mechanisms for the text simplification task on biomedical abstracts using the PLABA data set. Both automatic evaluations using a broad range of metrics and human evaluations were conducted to assess the system outputs. As automatic evaluation results show, both T5 and BART with Control Tokens demonstrated high accuracy in generating simplified versions of biomedical abstracts. However, when we delve into human evaluations, it becomes clear that each model possesses its unique strengths and trade-offs. T5 demonstrated strong performances at preserving the original abstracts' meaning, but sometimes at the cost of lacking simplification. By maintaining the core content and context of the input, it has proven to be over-conservative in some cases, resulting in outputs that very closely resemble the inputs therefore maintaining the abstract's complexity. On the other hand, BART-w-CTs demonstrated strong simplification performances to produce better-simplified versions. However, it has shown a potential drawback in reducing the preservation of the original meaning.

In future work, we plan to carry out investigations on more recent models including BioBART [24], try different prompting methods such as the work from [56], and design a more detailed human evaluation such as the work proposed by [57] with error severity levels might shed some light on this.

REFERENCES

[1] S. Dodson, S. Good, and R. Osborne, "Health literacy toolkit for low-and middle-income countries: A series of information sheets to empower communities and strengthen health systems," 2015. [Online]. Available: https://www.who.int/publications/i/item/9789290224754

[2] N. D. Berkman, S. L. Sheridan, K. E. Donahue, D. J. Halpern, and K. Crotty, "Low health literacy and health outcomes: an updated systematic review," *Annals of internal medicine*, vol. 155, no. 2, pp. 97–107, 2011.

[3] T. Greenhalgh, "Health literacy: towards system level solutions," 2015.

[4] A. T. McCray, "Promoting Health Literacy," *Journal of the American Medical Informatics Association*, vol. 12, no. 2, pp. 152–163, 03 2005. [Online]. Available: https://doi.org/10.1197/jamia.M1687

[5] D. Nishihara, T. Kajiwara, and Y. Arase, "Controllable text simplification with lexical constraint loss," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 260–266. [Online]. Available: https://aclanthology.org/P19-2036

[6] L. Martin, É. de la Clergerie, B. Sagot, and A. Bordes, "Controllable sentence simplification," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4689–4698. [Online]. Available: https://aclanthology.org/2020.lrec-1.577

[7] S. Agrawal, W. Xu, and M. Carpuat, "A non-autoregressive edit-based approach to controllable text simplification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, Aug. 2021, pp. 3757–3769. [Online]. Available: https://aclanthology.org/2021.findings-acl.330

[8] Z. Li, M. Shardlow, and S. Hassan, "An investigation into the effect of control tokens on text simplification," in *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*. Abu Dhabi, United Arab Emirates (Virtual): Association for Computational Linguistics, Dec. 2022, pp. 154–165. [Online]. Available: https://aclanthology.org/2022.tsar-1.14

[9] K. Attal, B. Ondov, and D. Demner-Fushman, "A dataset for plain language adaptation of biomedical abstracts," *Scientific Data*, vol. 10, no. 1, p. 8, 2023.

[10] BeeManc-team, "BeeManc at the PLABA track of TAC 2023: Investigating llms and controllable attributes for improving biomedical text readability," in *TAC2023*, B. Ondov, D. Demner-Fushman, and H. T. Dang, Eds. NIST, 2024.

[11] Y. Guo, W. Qiu, Y. Wang, and T. Cohen, "Automated lay language summarization of biomedical scientific reviews," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 160–168.

[12] B. Ondov, K. Attal, and D. Demner-Fushman, "A survey of automated methods for biomedical text simplification," *Journal of the American Medical Informatics Association*, vol. 29, no. 11, pp. 1976–1988, 2022.

[13] L. Han, A. Smeaton, and G. Jones, "Translation quality assessment: A brief survey on manual and automatic methods," in *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*. online: Association for Computational Linguistics, May 2021, pp. 15–33. [Online]. Available: https://www.aclweb.org/anthology/2021.motra-1.3

[14] L. Bacco, F. Dell'Orletta, H. Lai, M. Merone, and M. Nissim, "A text style transfer system for reducing the physician–patient expertise gap: An analysis with automatic and human evaluations," *Expert Systems with Applications*, vol. 233, p. 120874, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417423013763

[15] Q. Lyu, J. Tan, M. E. Zapadka, J. Ponnatapura, C. Niu, K. J. Myers, G. Wang, and C. T. Whitlow, "Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: results, limitations, and potential," *Visual Computing for Industry, Biomedicine, and Art*, vol. 6, no. 1, p. 9, 2023.

[16] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 09 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682

[17] S. Chakraborty, E. Bisong, S. Bhatt, T. Wagner, R. Elliott, and F. Mosconi, "BioMedBERT: A pre-trained biomedical language model for QA and IR," in *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 669–679. [Online]. Available: https://aclanthology.org/2020.coling-main.59

[18] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy, "The Pile: An 800gb dataset of diverse text for language modeling," *arXiv preprint arXiv:2101.00027*, 2020.

[19] U. Naseem, M. Khushi, V. Reddy, S. Rajendran, I. Razzak, and J. Kim, "Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–7.

[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[21] U. Naseem, A. G. Dunn, M. Khushi, and J. Kim, "Benchmarking for biomedical natural language processing tasks with a domain specific albert," *BMC bioinformatics*, vol. 23, no. 1, pp. 1–15, 2022.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[23] L. N. Phan, J. T. Anibal, H. Tran, S. Chanana, E. Bahadroglu, A. Peltekian,

and G. Altan-Bonnet, "Scifive: a text-to-text transformer model for biomedical literature," 2021.

[24] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "BioBART: Pretraining and evaluation of a biomedical generative language model," in *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 97–109. [Online]. Available: https://aclanthology.org/2022.bionlp-1.9

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "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*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[26] P. Lewis, M. Ott, J. Du, and V. Stoyanov, "Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. Online: Association for Computational Linguistics, Nov. 2020, pp. 146–157. [Online]. Available: https://aclanthology.org/2020.clinicalnlp-1.17

[27] S. Alrowili and V. Shanker, "BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA," in *Proceedings of the 20th Workshop on Biomedical Language Processing*. Online: Association for Computational Linguistics, Jun. 2021, pp. 221–227. [Online]. Available: https://aclanthology.org/2021.bionlp-1.24

[28] R. Tinn, H. Cheng, Y. Gu, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Fine-tuning large neural language models for biomedical natural language processing," *Patterns*, vol. 4, no. 4, p. 100729, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666389923000697

[29] S. Alrowili and V. Shanker, "Large biomedical question answering models with albert and electra." in *CLEF (Working Notes)*, 2021, pp. 213–220.

[30] N. H. Shah, D. Entwistle, and M. A. Pfeffer, "Creation and Adoption of Large Language Models in Medicine," *JAMA*, vol. 330, no. 9, pp. 866–869, 09 2023. [Online]. Available: https://doi.org/10.1001/jama.2023.14217

[31] A. Yan, J. McAuley, X. Lu, J. Du, E. Y. Chang, A. Gentili, and C.-N. Hsu, "Radbert: Adapting transformer-based language models to radiology," *Radiology: Artificial Intelligence*, vol. 4, no. 4, p. e210258, 2022.

[32] Y.-C. Lin, P. Hoffmann, and E. Rahm, "Enhancing cross-lingual biomedical concept normalization using deep neural network pretrained language models," *SN Computer Science*, vol. 3, no. 5, p. 387, 2022.

[33] J. Sybrandt and I. Safro, "Cbag: Conditional biomedical abstract generation," *Plos one*, vol. 16, no. 7, p. e0253905, 2021.

[34] A. Berhe, G. Draznieks, V. Martenot, V. Masdeu, L. Davy, and J.-D. Zucker, "AliBERT: A pre-trained language model for French biomedical text," in *BioNLP*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 223–236.

[35] H. Türkmen, O. Dikenelli, C. Eraslan, M. Calli, and S. Ozbek, "Harnessing the power of BERT in the Turkish clinical domain: Pretraining approaches for limited data scenarios," in *ClinicalNLP*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 161–170. [Online]. Available: https://aclanthology.org/2023.clinicalnlp-1.22

[36] B. Wang, Q. Xie, J. Pei, Z. Chen, P. Tiwari, Z. Li, and J. Fu, "Pre-trained language models in biomedical domain: A systematic survey," *ACM Comput. Surv.*, aug 2023, just Accepted. [Online]. Available: https://doi.org/10.1145/3611651

[37] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2790–2799.

[38] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *ACL (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 4582–4597. [Online]. Available: https://aclanthology.org/2021.acl-long.353

[39] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, oct 2021. [Online]. Available: https://doi.org/10.1145/3458754

[40] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Conference on Neural Information Processing System*, 2017, pp. 6000–6010.

[42] E. Lehman and A. Johnson, "Clinical-t5: Large language models built using mimic clinical text," 2023.

[43] Q. Lu, D. Dou, and T. Nguyen, "ClinicalT5: A generative language model for clinical text," in *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5436–5443. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.398

[44] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke, and M. Ingrisch, "Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports," 2022.

[45] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "Biogpt: Generative pre-trained transformer for biomedical text generation and mining," *Briefings in bioinformatics*, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252542956

[46] X. Zhang and M. Lapata, "Sentence simplification with deep reinforcement learning," *arXiv preprint arXiv:1703.10931*, 2017.

[47] K. Attal, B. Ondov, and D. Demner-Fushman, "A dataset for plain language adaptation of biomedical abstracts," *Scientific Data*, vol. 10, no. 1, p. 8, Jan. 2023. [Online]. Available: https://doi.org/10.1038/s41597-022-01920-3

[48] L. Martin, A. Fan, É. de la Clergerie, A. Bordes, and B. Sagot, "Multilingual unsupervised sentence simplification," *CoRR*, vol. abs/2005.00352, 2020. [Online]. Available: https://arxiv.org/abs/2005.00352

[49] J. Rapin and O. Teytaud, "Nevergrad - A gradient-free optimization platform," 2018, https://GitHub.com/FacebookResearch/Nevergrad.

[50] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," 2022. [Online]. Available: https://arxiv.org/abs/2210.11416

[51] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[52] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[53] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, "Optimizing statistical machine translation for text simplification," vol. 4, 2016, pp. 401–415. [Online]. Available: https://www.aclweb.org/anthology/Q16-1029

[54] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

[55] F. Alva-Manchego, L. Martin, C. Scarton, and L. Specia, "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*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 49–54. [Online]. Available: https://aclanthology.org/D19-3009

[56] Y. Cui, L. Han, and G. Nenadic, "MedTem2.0: Prompt-based temporal classification of treatment events from discharge summaries," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 160–183. [Online]. Available: https://aclanthology.org/2023.acl-srw.27

[57] S. Gladkoff and L. Han, "HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation," in *LREC2022*. Marseille, France: ELRA, Jun. 2022, pp. 13–21.