Original Research

Health text simplification: An annotated corpus for digestive cancer education and novel strategies for reinforcement learning

Md Mushfiqur Rahman [*,1], Mohammad Sabik Irbaz[1], Kai North, Michelle S. Williams, Marcos Zampieri, Kevin Lybarger

George Mason University, United States of America

ARTICLE INFO

ABSTRACT

Objective: The reading level of health educational materials significantly influences the understandability and accessibility of the information, particularly for minoritized populations. Many patient educational resources surpass widely accepted standards for reading level and complexity. There is a critical need for high-performing text simplification models for health information to enhance dissemination and literacy. This need is particularly acute in cancer education, where effective prevention and screening education can substantially reduce morbidity and mortality.

Methods: We introduce *Simplified Digestive Cancer* (SimpleDC), a parallel corpus of cancer education materials tailored for health text simplification research, comprising educational content from the American Cancer Society, Centers for Disease Control and Prevention, and National Cancer Institute. The corpus includes 31 web pages with the corresponding manually simplified versions. It consists of 1183 annotated sentence pairs (361 train, 294 development, and 528 test). Utilizing SimpleDC and the existing Med-EASi corpus, we explore Large Language Model (LLM)-based simplification methods, including fine-tuning, reinforcement learning (RL), reinforcement learning with human feedback (RLHF), domain adaptation, and prompt-based approaches. Our experimentation encompasses Llama 2, Llama 3, and GPT-4. We introduce a novel RLHF reward function featuring a lightweight model adept at distinguishing between original and simplified texts when enables training on unlabeled data.

Results: Fine-tuned Llama models demonstrated high performance across various metrics. Our RLHF reward function outperformed existing RL text simplification reward functions. The results underscore that RL/RLHF can achieve performance comparable to fine-tuning and improve the performance of fine-tuned models. Additionally, these methods effectively adapt out-of-domain text simplification models to a target domain. The best-performing RL-enhanced Llama models outperformed GPT-4 in both automatic metrics and manual evaluation by subject matter experts.

Conclusion: The newly developed SimpleDC corpus will serve as a valuable asset to the research community, particularly in patient education simplification. The RL/RLHF methodologies presented herein enable effective training of simplification models on unlabeled text and the utilization of out-of-domain simplification corpora.

1. Introduction

Accessibility and comprehensibility of patient medical educational materials are critical to improving health literacy and patient health [1]. Patient educational materials, including cancer-related materials, are essential to provide patients with the information necessary to make informed healthcare decisions. Health literacy is strongly influenced by the clarity and understandability of such materials and plays a key role in health outcomes [2]. It affects various aspects of patient health, including reducing risk factors, encouraging timely screening, and facilitating proactive health management. To ensure comprehensibility for a broad audience, the American Medical Association (AMA) and the National Institutes of Health (NIH) recommend a reading level of sixth grade or lower for patient educational materials [3,4]. Despite the critical role of these educational materials in patient health, there remains a substantial discrepancy between the recommended and actual readability observed in practice [5]. Health educational materials from prominent sources are often at high school or college reading levels [5].

This reading complexity poses a barrier, particularly for people with lower literacy or limited health knowledge, preventing them from fully understanding and taking action on essential health information. The accessibility and comprehensibility of educational materials are particularly concerning for disseminating cancer information, where clear and understandable guidance impacts patient education and decision-making [6,7]. Our focus on text-based cancer educational materials stems from their pivotal role in patient education and the pressing need to improve their accessibility to diverse populations.

Advancements in Natural Language Processing (NLP) and Large Language Models (LLMs) offer a promising avenue to reconcile the disparity in reading levels in existing medical education materials through automatic simplification, to improve public health literacy [8]. LLMs have demonstrated proficiency in reducing linguistic complexity and preserving medical meaning [9,10]. Finding the optimum balance of simplification and information retention is challenging, especially when the stakes of miscommunication are high [11]. Developing and evaluating data-driven text simplification models requires parallel corpora of original and simplified text. Creating these parallel corpora is extremely labor-intensive and costly. In the health domain, these challenges are exacerbated by the need for domain-specific medical expertise. To address these challenges, robust methods are needed to create high-performing simplification models with minimal training data.

This paper makes significant contributions to the simplification of health information and text simplification more broadly. We focus on patient education materials related to digestive cancer, given the prevalence and mortality of these cancers [12,13]. We introduce *Simplified Digestive Cancer (SimpleDC)*, a novel text simplification corpus comprising patient educational materials from three prominent sources: National Cancer Institute (NCI), American Cancer Society (ACS), and Centers for Disease Control and Prevention (CDC). Although prior research has explored the simplification of health information, we are unaware of prior research focused on patient educational materials. SimpleDC is unique in its concentration on digestive cancers and includes parallel sentences (original and simplified versions), where simplifications were generated by a team of nurse oncologists and a nurse practitioner with medical and patient education backgrounds. We explore supervised fine-tuning (SFT) and reinforcement learning (RL) approaches using Llama 2 [14] and Llama 3 [15] and introduce a novel approach for reinforcement learning from human feedback (RLHF). We also explore prompt-based methods using GPT-4 [16,17] and introduce a novel self-correction strategy that utilizes emergent reasoning capabilities. Our evaluation includes standard automatic metrics (e.g. SARI, BLEU, BERTScore, and FKGL) and a human comparison of LLM and human-generated simplifications. Automatic and manual evaluations demonstrate that the proposed RLHF reward function outperforms existing text simplification reward functions, complements SFT, and facilitates domain adaptation. Our findings advance the simplification of health text and enable the automatic creation of more accessible health information, ultimately contributing to improved health literacy. We provide the new SimpleDC corpus, trained models, and code to the research community.[2],[3]

Statement of Significance

Problem	Existing patient education materials are frequently presented at a high reading level, reducing accessibility for individuals with lower literacy.
What is already known	Existing health text simplification corpora are limited, and there are knowledge gaps for low-resource text simplification related to RL, RLHF, and LLMs.
What this paper adds	We present a novel text simplification corpus, SimpleDC, and introduce an innovative RLHF reward function for training LLM simplification models. The findings demonstrate the RLHF approach surpasses existing methods in performance, complements supervised fine-tuning, and supports domain adaptation.

2. Related work

Text simplification involves the transformation of texts to improve understandability and accessibility to broad audiences. This section presents the scope of text simplification, health text simplification corpora, text simplification methods, and related RL research.

2.1. Scope of text simplification

Text simplification has been explored in many domains (news articles, literature, scientific articles, etc.) where complex ideas must be communicated clearly and concisely [18–22]. This research spans multiple languages, addressing the unique challenges and strategies needed for multilingual simplification to ensure clarity for linguistically diverse audiences [23–25]. Text simplification can be subdivided into three tasks: lexical simplification, sentence-level simplification, and paragraph/document-level simplification [26–28]. Lexical simplification commonly involves substituting words or phrases, including medical terminology [29–33]. However, word- or phrase-level editing limits simplification scope and cannot accommodate more comprehensive rephrasing or restructuring. Sentence- and document-level simplification enables more comprehensive text restructuring to enhance overall clarity and accessibility and typically requires parallel text corpora (complex and simplified text) for model training and evaluation [34].

2.2. Health text simplification

Health text simplification is challenging due to medical jargon, complex sentence structures, and lengthy explanations. It has been explored through several tasks, such as generating patient-friendly summaries of clinical notes [35–37], making biomedical literature accessible to laypeople, and simplifying educational materials for patients [8,38–42]. Text simplification can focus on improving readability or making stylistic modifications; however, simplification of health information requires rigorous attention to meaning preservation and medical accuracy, given the critical implications of miscommunication [8,32]. This challenge is amplified by the diverse needs of patients, which encompasses varying levels of literacy, language background, and other factors that require a nuanced approach to achieve clarity and accuracy.

Due to the labor-intensive nature of annotation, there are limited health-focused text simplification corpora [9,40,43,44]. The Medical dataset for Elaborative and Abstractive Simplification (Med-EASi) dataset comprises medical text from SIMPWIKI that was simplified by a team of medical experts and layperson crowd workers at the sentence-level [43]. The Autocomplete for Medical Text Simplification (AutoMeTS) dataset consists of automatically aligned sentences

from English Wikipedia and Simple English Wikipedia for medical topics [40]. Devaraj et al. provided a manually annotated parallel corpus of biomedical literature with simplified text created by domain experts [9]. We introduce a new health text simplification corpus, SimpleDC, which is based on educational patient materials from prominent information sources (ACS, CDC, and NCI) and explicitly focuses on the cancer domain.

2.3. Text simplification approaches

Early text simplification work used heuristics and traditional machine learning techniques, including pattern-based paraphrasing [45, 46]. Contemporary approaches conceptualize text simplification similar to machine translation, employing neural language models to transform sentences into simpler counterparts [8,47]. Document-level simplification adopts a similar machine translation approach [48,49]. Essentially all state-of-the-art text simplification models [50,51] are based on LLMs, including: (i) encoder–decoder architectures, like Text-to-Text Transfer Transformer (T5) [52] and Fine-tuned LAnguage Net on T5 (FlanT5) [53] and (ii) decoder-only models such as Large Language Model Analysis (Llama) variants [14,15,54] and Generative Pre-trained Transformer (GPT) [17] variants, like GPT-4 [16].

LLM simplification models are predominantly created through SFT; however, there has been some limited exploration of prompt-based methods and reinforcement learning. Instruction tuning allows LLMs to follow natural language instructions, facilitating prompt-based approaches such as in-context learning, chain-of-thought (CoT) learning [55,56], and self-correction [57,58]. These methods rely on the inherent language understanding and reasoning capabilities of LLMs [55, 59,60]. CoT learning enables models to articulate step-by-step reasoning, similar to human problem solving, improving response accuracy and interpretability [55,56]. Self-correction allows models to iteratively refine their outputs, improving the precision and reliability of their responses [58]. We present novel approaches for adapting these techniques to our text simplification task.

2.4. Reinforcement learning for simplification

RL has emerged as a promising approach for text simplification [44, 61,62]. Many studies employ a mix of simplicity metrics and linguistic features to construct their reward functions. Zhang et al. [63] introduced an RL approach, DRESS, which includes a reward function focused on simplicity, fluency, and meaning preservation. Nakamachi et al. explored the use of rewards related to grammaticality, meaning preservation, and simplicity [61]. Similarly, Yanamoto et al. introduced an RL approach based on the difference between estimated and target difficulty levels, using a novel reward calculation method [62]. Alkaldi et al. presented an RL paradigm that refines reading levels in a controlled and iterative process [64]. Luo et al. demonstrated the use of controllability in the biomedical sector [65]. The TESLEA RL framework integrates heuristic and data-driven reward models to develop LLM-based text simplification models [44] using three rewards: (1) readability (FKGL), (2) relevance (cosine similarity between sentence embeddings), and (3) lexical simplicity (Zipf frequency). In our baseline experimentation, we implemented a TESLEA-inspired approach that includes the readability and relevance rewards but omits the Zipf frequency-based lexical simplicity reward to avoid penalizing medically significant but infrequent words. We introduce a new RL approach that combines the heuristic readability reward (FKGL) with a data-driven RLHF reward for assessing simplification.

3. Materials and methods

3.1. Data

3.1.1. SimpleDC

We introduce *SimpleDC*, which is a meticulously curated parallel corpus, specifically designed for text simplification research within the medical domain. It comprises patient educational information from three high-reputation institutions: ACS, CDC, and NCI. These organizations were chosen because of their expertise and comprehensive coverage of health topics, including cancer. There are many types of cancer, each with different characteristics. The available cancer educational materials on the ACS, CDC, and NCI websites exceed our annotation budget, so we focused on the subset of cancers associated with the digestive system to create a corpus that spans many cancer types that share similar terminology and topics (e.g. anatomy or symptoms). The focus on digestive cancers is also motivated by the prevalence and mortality of these cancers and the role that prevention and screening play in the outcomes [66]. SimpleDC responds to the need to make medical information more accessible and comprehensible to a broad audience, especially minoritized populations.

We collected text from the ACS, CDC, and NCI web pages containing patient educational content for eight digestive cancers: anal, bile duct, colorectal, gastrointestinal, liver, pancreatic, small intestine, and stomach. For these cancer types, we collected the text from web pages presenting introductory cancer information (e.g. About, Prevention, Screening). The collected text was cleaned and processed, including the removal of extraneous white space, adjusting punctuation, removing rich-text formatting, and parsing into sentences. We developed annotation guidelines for the simplification task and trained a team of subject matter experts with expertise in both cancer and patient education, including two nurse oncologists and a nurse practitioner. For each web page, we created a two-column document, where the left and right sides contained the original text with each sentence on a separate line. The annotators modified the sentences in the right column based on the annotation guidelines to create the simplified versions. This approach allowed the annotators to focus on sentence-level simplification while considering the entire web page for context. Table 1 presents annotation examples for each source, together with the Flesch-Kincaid Grade Level (FKGL) scores for each example [67]. FKGL is a rule-based readability formula that assesses text complexity based on the average sentence length and syllable count per word, yielding scores indicative of the US school grade levels required for comprehension.

Since text simplification is a generative task without categorical labels, traditional inter-annotator agreement (IAA) metrics, like Cohen's Kappa, are not well suited. Instead, we evaluated annotation quality by having two annotators create the simplified text and then having a third annotator assess their preference for simplifications. In the initial training round, two annotators independently generated simplifications for each web page. The generated texts from the two annotators were marked *A* and *B* randomly. To assess annotation quality, a third annotator reviewed the simplifications relative to the original text and indicated their preferred simplification as one of four choices – *A* is better; *B* is better; *Both* are good; or *None* is good. The distribution of the third annotator's preferences was as follows: the first annotator was preferred for 34% of samples, the second annotator was preferred for 32% of samples, both annotators were considered effective for 26% of samples, and neither simplification was deemed adequate for 8% of samples. SimpleDC comprises 31 annotated web pages (14 train, 7 development, and 10 test), including 1183 annotated sentence pairs (361 train, 294 development, and 528 test). The development and test sets were doubly annotated by two annotators and adjudicated by the third annotator to create a robust evaluation set. The training set was singly annotated. Table 2 gives a detailed overview of the SimpleDC dataset. Fig. 1 presents the distribution of the FKGL reading level for

Table 1

Annotation examples and Flesch-Kincaid Grade-Level (FKGL) scores.

Source	Original text	Simplified text	Orig. FKGL	Simp. FKGL
ACS	The anus is the opening at the lower end of the intestines.	The anus is the opening where bowel movements come out of your body.	4.4	3.6
CDC	Screening can find precancerous polyps–abnormal growths in the colon or rectum–that can be removed before they turn into cancer.	Screening can help find growths in your colon or rectum called precancerous polyps. They can be removed before they turn into cancer.	10.7	5.2
NCI	Primary liver cancer is a disease in which malignant (cancer) cells form in the tissues of the liver.	A cancer that starts in your liver is called primary liver cancer.	10.3	6.8

Table 2

Statistics of the proposed Simple DC dataset.

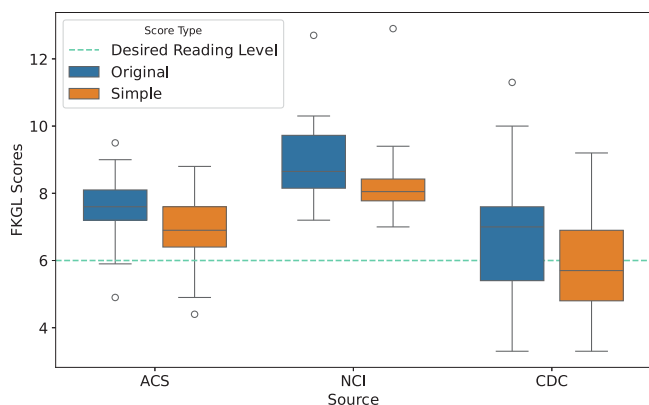|	Num. samples	Num. of webpages	Original FKGL		Simplified FKGL		Annotation type
			Mean	Median	Mean	Median	
Train	361	14	9.6	9.1	7.8	8.0	Single
Val	294	7	7.7	7.2	7.1	7.0	Double
Test	528	10	8.1	7.6	7.4	7.5	Double
Overall	1183	31	8.5	8	7.4	7.6	–

Fig. 1. Boxplot representing the FKGL scores of Original and Simple Text of the SimpleDC dataset (grouped by source).

the original and simplified text for different data sources. The median reading level varies 7.0–9.2 across sources, and the reduction in FKGL from simplification ranged 0.6–1.0.

We collected an additional 1395 web pages from ACS, CDC, and NCI related to digestive cancer that were not annotated/simplified. We randomly sampled 1000 of these unannotated web pages for use in the RL procedures described in Section .

3.1.2. Med-EASi

Our experimentation utilized the Med-EASi [43] corpus, which includes simplifications for medical texts annotated through a combination of expert, layperson, and AI-generated contributions. The more complex medical texts are annotated by experts, while simpler texts are handled by layperson crowd-workers, often assisted by AI-generated suggestions. It includes annotations for specific edit types: elaboration, replacement, deletion, and insertion. There are 1979 original-simplified sentence pairs, covering a wide range of medical topics. The average FKGL of the original and simplified samples is approximately 13th and 10th grades, respectively. The introduction of Med-EASi includes simplification results based on fine-tuning T5-large [52]. Experimentation included a conventional machine translation approach using T5, as well as a controlled simplification approach that utilized the edit types.

3.2. Simplification models

Our exploration of SimpleDC and Med-EASi focused on state-of-the-art decoder-only LLMs, including Llama 2 [14], Llama 3 [15]

and GPT-4 [16]. Using Llama models, we explored zero-shot, SFT, and RL/RLHF approaches. Using GPT-4, we explored prompt-based strategies, including a novel self-correction strategy.

3.2.1. Supervised fine-tuning (SFT)

As a baseline, we trained Llama models on original-simplified sentence pairs using conventional SFT with cross-entropy loss. SFT experimentation included SimpleDC and Med-EASi.

3.2.2. Reinforcement learning

We explored RL strategies to incorporate unlabeled in-domain text into training. Using Proximal Policy Optimization (PPO) [68], we investigated RL reward functions that include metrics related to reading level assessment (i.e., FKGL), topical relevance (semantic similarity), and a data-driven approach to identify simplified text (referred to as Original vs. Simplified below). This exploration includes RLHF [69]. Fig. 2 presents an overview of the RL pipeline, each reward metric is described below. RL frameworks commonly use reward values in the range $[0, 1]$, where higher values indicate more desirable outcomes and lower values indicate less desirable outcomes. For each reward metric presented, individual and composite rewards are scaled to this range.

Readability Reward: FKGL is a heuristic function to assess reading level. Although an imperfect metric, FKGL provides an automatic way to assess the reading grade level of the text. The TESLEA approach [44] includes a normalization strategy to convert FKGL scores to the range $[-1, 1]$, based on the target reading level. Our reading level reward, R_{FKGL}, which was inspired by the TESLEA approach, was calculated by normalizing the FKGL score by the reading level target and applying a sigmoid function:

$$R_{FKGL} = sigmoid\left(\frac{6.5 - FKGL(generated)}{6.5}\right)$$

The normalized reading score (fraction within the sigmoid) yields positive values for simpler text (more desirable) and negative values for more complex text (less desirable). The sigmoid function maps the normalized reading scores to the range $[0, 1]$, where higher values indicate lower reading levels.

Relevance Reward: The relevance reward, R_{Rel}, is intended to quantify the semantic relationship between the simplified and original texts, where a higher reward indicates greater semantic similarity. We follow the TESLEA approach [44] for relevance scoring: (1) the original and simplified texts are mapped to separate vector representations using an encoder and (2) semantic similarity is assessed based on cosine similarity between these vectors. BioSentVec [70] was selected as the encoder, as it has been extensively trained on biomedical literature.
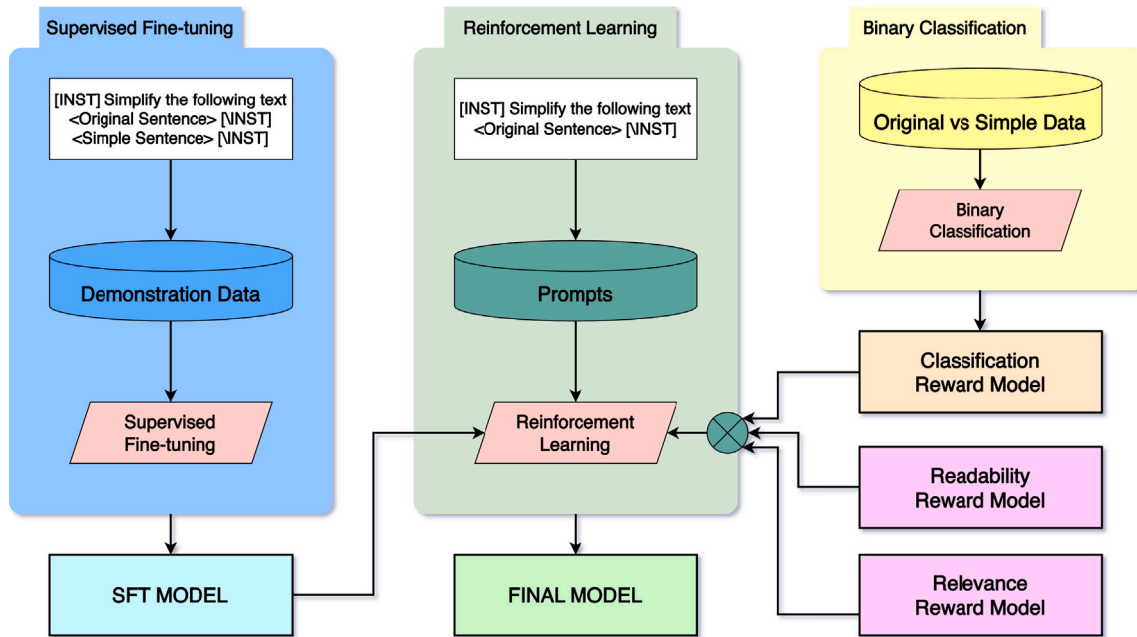
Fig. 2. Reinforcement learning from human feedback for medical text simplification.

Cosine similarity falls in the range $[0, 1]$, where higher values indicate higher semantic correspondence. The relevance reward incentivizes the preservation of semantic relatedness between the original and simplified texts. We used the pre-trained BioSentVec without any training on SimpleDC.

Original vs. Simplified (OvS) Reward: We introduce a new RL reward, R_{OvS}, based on the likelihood that a sample has been simplified. Using SimpleDC training data, we trained a binary classifier to label samples as original or simplified using BioMed-RoBERTa base [71]. Similar to InstructGPT [72], we utilize a classifier trained on human-annotated data; however, in our formulation, the classifier distinguishes between original and simplified texts. This classifier provides a probabilistic reward signal reflecting human judgment, capturing annotators' preferences and is considered RLHF. The *OvS* classifier achieved 75.1% accuracy on the SimpleDC development set. This performance demonstrates the ability to identify key features of the original and simplified texts. The reward value, R_{OvS}, is the predicted probability that the text is simplified ($P(simplified)$) providing a continuous reward signal ranging from $[0, 1]$, where higher values suggest better simplification. Our intuition behind this reward function is that it provides a mechanism for learning features of the original and simplified text and propagating this learning to unlabeled samples during RL.

Reward Aggregation: We calculated aggregated reward values from the individual metrics using the harmonic mean, H, to balance the contribution from each metric, as

$$H(x_1, x_2) = \frac{1}{\frac{1}{x_1} + \frac{1}{x_2}}$$

We developed two aggregated reward functions:

- $R_{FKGL+Rel}$: We combined the reading level reward, R_{FKGL}, and the relevance reward, R_{Rel}, to create an aggregated reward, $R_{FKGL+Rel} = H(R_{FKGL}, R_{Rel})$. This reward is intended to capture both text complexity and semantic similarity to yield text that is at the desired reading level and preserves the original meaning. It does not incorporate any learning from SimpleDC and is therefore considered RL.

- $R_{FKGL+OvS}$: We also combined the reading level reward, R_{FKGL}, and the Original vs. Simplified reward, R_{OvS}, to create an aggregated reward, $R_{FKGL+OvS} = H(R_{FKGL}, R_{OvS})$. This aggregated reward is intended to assess text complexity (R_{FKGL}) and identify linguistic features that differentiate the original and simplified texts (R_{OvS}) to produce text that is both at the desired reading level and emulates the simplification of SimpleDC. As this reward function incorporates a training signal derived from the SimpleDC annotations, it is considered RLHF.

In initial experimentation, we explored weighted averages, drawing inspiration from the TESLEA paper [44]. However, due to inconsistent outcomes with hyperparameters and lower performance, we chose to focus on experiments employing the harmonic mean.

3.2.3. Fine-tuning with reward models

We explored different pipelines for incorporating RL/RLHF, including only using RL/RLHF, combining SFT and RL/RLHF, and using RL/RLHF for domain adaptation:

- *RL/RLHF only*: A pre-trained LLM was trained using RL/RLHF with unlabeled in-domain data (digestive cancer text) to learn the target domain and simplification task without any SFT of the LLM on annotated in-domain data (SimpleDC). The goal of this exploration was to understand the achievable performance without any SFT.

- *In-domain SFT + RL/RLHF*: A pre-trained LLM was first trained through SFT on labeled in-domain data (SimpleDC train set), followed by continued training through RL/RLHF on unlabeled in-domain data (digestive cancer text). This experimentation was intended to assess whether RL/RLHF can continue to improve the model even after in-domain SFT.

- *Out-domain SFT + RL/RLHF*: A pre-trained LLM was first trained through SFT on labeled out-of-domain data (Med-EASi training set), followed by continued training through RL/RLHF unlabeled in-domain data (digestive cancer text).
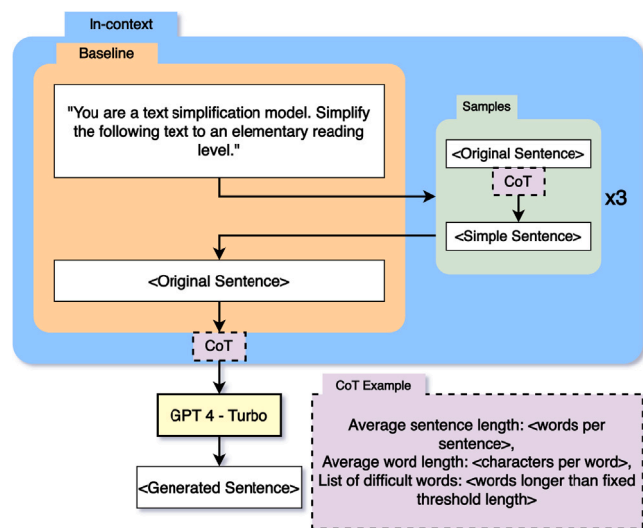
Fig. 3. Chain-of-Thought (CoT) prompting along with Baseline and In-context.

In this experimentation, the unlabeled in-domain text consisted of 1000 web pages from ACS, CDC, and NCI related to digestive cancer that were *not annotated* and did not overlap with SimpleDC.

In all RL experiments, training used PPO, where a Kullback–Leibler (KL) divergence constraint [73] was implemented to limit excessive deviation of the RL-trained LLM from the original LLM. This divergence constraint provides a regulatory mechanism to ensure the LLM maintains linguistic coherence and avoids overfitting to potentially incorrect or excessive scores from the reward model. This constraint was essential due to the inherent limitations of the reward model, which may not accurately capture the quality of the generated text.

3.2.4. Prompt-based models

Given our focus on developing simplification models with limited training data and the success of GPT-4 in many tasks, we explored prompt-based approaches for text simplification with GPT-4 [16], including in-context learning, CoT, and self-correction.

Zero-shot: In the zero-shot approach, GPT-4 was instructed to simplify the text to an elementary level of reading, to provide a baseline for evaluating more complex prompting strategies. The specific simplification prompt is included in the *Baseline* module in Fig. 3.

In-context Learning: The in-context variant used the same initial prompt as the zero-shot approach but was augmented with three representative examples from the SimpleDC training set. These examples were manually selected to illustrate the desired simplification style and complexity.

FKGL-Enhanced Prompt: Building on the initial prompt, we added a detailed description of the FKGL heuristic, indicating that lower reading levels are associated with shorter words and sentences, and provided the FKGL formula for grade-level calculations.

Chain-of-Thought (CoT): Our CoT approach builds on the *FKGL-Enhanced Prompt* approach by providing the model with additional sample-specific context, including the average word length, average sentence length, and a list of longer words (≥ 3 syllables) for each sample. This model aims to provide explicit guidance for the simplification process and anchor the simplification in commonly used readability standards. Fig. 3 illustrates the CoT approach.

Self-Correction (SC): Inspired by recent work on self-correction [57,58], we created the self-correction approach for simplification shown in Fig. 4. The intent was to both exploit the reasoning capabilities of the LLM and incorporate additional external learning signals to iteratively refine the output. Each generated simplification was scored for reading level using FKGL and semantic fidelity using BERTScore.

Acceptance criteria include: (1) FKGL score ≤ 6 or FKGL reduced by three grade levels from the original text and (2) BERTScore between the generated and original texts ≥ 0.95. If the simplification met the acceptance criteria, it was accepted; otherwise, the simplification was fed back into GPT-4 with instructions for improving the simplification. Each sample was processed iteratively at most three times.

3.2.5. Experimental paradigm

We utilized PyTorch [74] and Hugging Face's Transformers [75]. We used one Llama 2 variant: *Llama-2-7b-chat-hf* [4] and two Llama 3 variants: *Llama-3-8B-Instruct* [5] and *Llama-3-70B-Instruct*. [6] Experimentations were performed using NVIDIA A100 GPUs.

For the *Llama-2-7b-chat-hf* model, all experiments, including zero-shot, SFT, and RL/RLHF, were conducted using a single A100-80 GB GPU. For the *Llama-3-8B-Instruct* model, zero-shot and SFT experiments were also feasible using a single A100 GPU; however, RL/RLHF experimentation required two A100-80 GB GPUs. For the *Llama-3-70B-Instruct* model, zero-shot inference was conducted using four A100-80 GB GPUs without quantization; however, SFT required 4-bit quantization [76] with four A100-80 GB GPUs. Unfortunately, RL/RLHF experimentation with *Llama-3-70B-Instruct* was not feasible with our computational resources, even with the 4-bit quantization and utilization of four A100-80 GB GPUs.

We also used Weights and Biases [77] for tracking training and performance. For SFT, we utilized LORA [78] utilities from the Parameter-Efficient Fine-Tuning (PEFT) package [79,80], to reduce the time and space complexity of model training. Hyperparameters were selected based on computational constraints and performance optimization. We selected a batch size of 8 to balance memory usage and model updating. The learning rate was set to 2e−4, following preliminary experiments to find a balance between catastrophic forgetting and suboptimal convergence. Additional hyperparameters, such as weight decay (0.001) and maximum gradient norm (0.3), were included to prevent overfitting and maintain the stability of the training process. For RLHF, we specified a range of hyperparameters, including a learning rate of 1.41e−5 and PPO epochs of 2, reflecting a conservative approach to incremental learning. Our choice of batch and mini-batch sizes, set to 4, was influenced by the need to manage the computational load effectively while still capturing the nuances of gradient updates during training.

3.3. Evaluation

We utilized widely accepted evaluation metrics, each addressing specific aspects of text simplification quality, including:

1. **System output Against References and against the Input sentence (SARI)** [81]: SARI evaluates simplification quality by comparing system-generated output with both the original and reference texts. Outputs are assessed based the operations: *add* - words that are in the reference but not in the original text; *keep* - words from the original text are also in the reference, and *delete* - words from the original text that are not in the reference.
2. **Bilingual Evaluation Understudy (BLEU)** [82,83]: BLEU measures the precision of n-grams in the generated text against a reference by evaluating how many words in the generated text are also present in the reference text. We used 4-gram BLEU.
3. **BERTScore** [84]: BERTScore quantifies the semantic similarity between texts using contextual embeddings from the BERT model. It compares sentence-level embeddings of the generated text and reference texts, providing a measure of semantic comparison beyond surface-level text matching.
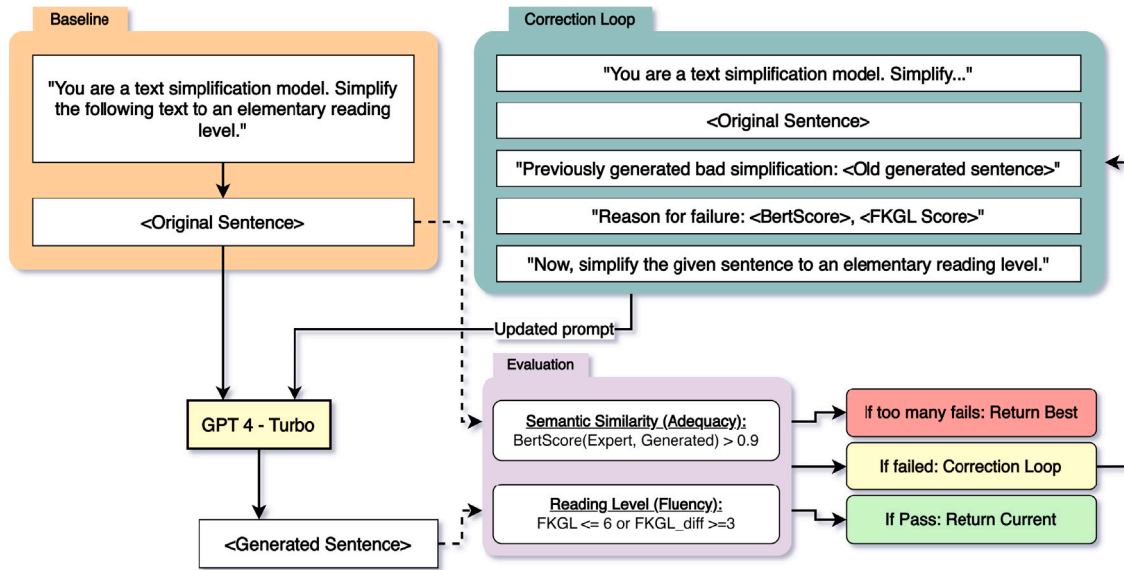
Fig. 4. Diagram of self-correction prompting.

4. **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** [85]: ROUGE focuses on n-gram overlap between the output and reference texts. ROUGE-N measures the overlap of n-grams, and ROUGE-L measures the longest common subsequence. We used ROUGE-L because it is less sensitive to minor differences in wording and is particularly useful for assessing fluency and coherency.

5. **Flesch-Kincaid Grade Level (FKGL)** [67]: FKGL assesses the reading level of text based on the average number of syllables per word and number of words per sentence. FKGL scores indicate the estimated grade level.

4. Results

4.1. Text simplification

Table 3 presents the performance on the withheld SimpleDC test set, evaluating: *Llama-2-7b-chat-hf*, *Llama-3-8B-Instruct*, *Llama-3-70B-Instruct*, and GPT-4 Turbo. In the zero-shot setting *(Exp. 01, 03, 05, and 15)*, *Llama-2-7b-chat-hf* achieved the highest SARI score, although GPT-4 achieved the lowest FKGL. SFT with the Llama models *(Exp. 02, 04, and 06)* markedly improved performance across SARI, BLEU, BERTScore, and ROUGE metrics, while reducing the FKGL by approximately two grade levels. These results underscore the role of SFT in understanding and learning the SimpleDC annotation schema and the annotators' stylistic nuances. Notably, both *Llama-2-7b-chat-hf* and *Llama-3-8B-Instruct* showed substantial performance improvements with SFT, with 47.92 and 52.04 SARI and 7.08 and 6.32 FKGL, respectively, indicating effective task training. *Llama-3-70B-Instruct* also demonstrated noteworthy improvements with SFT *(Exp. 06)*, achieving 46.73 SARI and the lowest reading level at 5.16 FKGL.

The success of the RL/RLHF varies by Llama variant and reward strategy. In RL/RLHF-only experiments, *Llama-3-8B-Instruct (Exp. 11–13)* consistently outperformed *Llama-2-7b-chat-hf (Exp. 07–09)* across multiple metrics, achieving higher SARI and lower FKGL. For *Llama-3-8B-Instruct*, the proposed RLHF reward function, $R_{FKGL+OvS}$, markedly increased the SARI score from 20.44 *(Exp. 11)* in the RL scenario to 55.01 *(Exp. 13)*. These findings emphasize the importance of integrating human feedback from the target task/data set through RLHF, as compared to the data set agnostic feedback of RL alone.

When combining SFT with RLHF for *Llama-2-7b-chat-hf (Exp. 10)*, the model achieved the highest performance, at 60.39 SARI, 0.70 BLEU, and 0.97 BERTScore. This configuration also yielded a relatively low reading level at 8.24 FKGL, demonstrating that SFT and RLHF can be complementary in enhancing text simplification capabilities. However, for *Llama-3-8B-Instruct*, combining SFT with RLHF *(Exp. 14)* also showed no discernable improvement over RLHF alone *(Exp. 13)*. In this configuration, the model achieved a SARI of 50.98, BLEU of 0.53, BERTScore of 0.91, and FKGL of 8.11.

For in-context learning, GPT-4 Turbo *(Exp. 15–19)* achieved lower scores across all adequacy metrics (SARI, BLEU, BERTScore, and ROUGE) compared to the Llama models *(Exp. 01–14)* but achieved the lowest FKGL. Lower performance does not necessarily indicate lower quality simplifications generated by GPT-4. Based on our qualitative review, GPT-4 edited the original text more extensively, resulting in a larger divergence between the original and simplified text. Additionally, the voice or writing style of GPT-4 appears to differ from that of our annotators, contributing to the lower performance for adequacy metrics. The FKGL-Enhanced prompting strategy *(Exp. 17)* achieved the best performance across adequacy metrics, and the CoT prompting strategy *(Exp. 18)* achieved the lowest FKGL. The Self-Correction approach *(Exp. 19)* achieved a balance between adequacy and reading level at 22.86 SARI and 4.48 FKGL.

4.2. Domain adaptation

We used the Med-EASi corpus to explore the role of RL/RLHF in domain adaptation by first conducting out-of-domain SFT with Med-EASi, followed by in-domain RL/RLHF training using unlabeled digestive cancer text. Table 3 *(Exp. 20–25)* presents the domain adaptation results. This experimentation evaluated the feasibility of leveraging out-of-domain corpora to enhance in-domain performance. In this domain adaptation setting, the performance gains from RL/RLHF vary by Llama variant and specific reward function. For *Llama-3-8B-Instruct*, continued in-domain RLHF using the $R_{FKGL+OvS}$ reward markedly improved performance over out-of-domain SFT on Med-EASi alone, increasing SARI from 38.73 to 51.06, BLEU from 0.43 to 0.57, and ROUGE from 0.60 to 0.72 *(Exp. 23 vs. Exp. 25)*. However, *Llama-3-8B-Instruct* with RLHF alone and no prior SFT on Med-EASI *(Exp. 13)* achieved better results

Table 3

In-domain and domain adaptation performance of Prompt-based, SFT, and RLHF experimentation on SimpleDC dataset. Source FKGL: 8.11, Target FKGL: 7.65. Higher scores indicate better performance for SARI, BLEU, BERTScore, and ROUGE ($\uparrow$), whereas lower scores are preferred for FKGL ($\downarrow$).

In-Domain							
Model name	Variant	Exp. No.	SARI ($\uparrow$)	BLEU ($\uparrow$)	BERT Score ($\uparrow$)	ROUGE ($\uparrow$)	FKGL ($\downarrow$)
Llama-2-7b-chat-hf	Zero-shot	01	28.43	0.15	0.90	0.48	9.08
	SFT	02	47.92	0.47	0.95	0.70	7.08
Llama-3-8B-Instruct	Zero-shot	03	25.15	0.11	0.90	0.46	8.96
	SFT	04	52.04	0.42	0.93	0.61	6.30
Llama-3-70B-Instruct	Zero-shot	05	24.49	0.02	0.84	0.12	8.59
	SFT	06	46.73	0.39	0.93	0.56	5.16
Llama-2-7b-chat-hf	RL ($R_{FKGL+Rel}$)	07	29.99	0.29	0.91	0.54	8.29
	RLHF (R_{OvS})	08	31.62	0.30	0.91	0.56	8.40
	RLHF ($R_{FKGL+OvS}$)	09	29.03	0.28	0.91	0.54	8.38
	SFT+RLHF ($R_{FKGL+OvS}$)	10	60.39	0.70	0.97	0.89	8.24
Llama-3-8B-Instruct	RL ($R_{FKGL+Rel}$)	11	20.44	0.11	0.88	0.32	7.38
	RLHF (R_{OvS})	12	42.45	0.37	0.91	0.51	8.02
	RLHF ($R_{FKGL+OvS}$)	13	55.01	0.59	0.93	0.80	8.24
	SFT+RLHF ($R_{FKGL+OvS}$)	14	50.98	0.53	0.91	0.73	8.11
GPT-4 Turbo	Zero-shot	15	22.03	0.14	0.92	0.43	5.71
	In-context	16	22.04	0.14	0.92	0.42	5.68
	FKGL-Enhanced	17	23.78	0.24	0.94	0.53	5.37
	CoT	18	20.97	0.18	0.93	0.46	4.71
	Self-Correction	19	22.86	0.21	0.93	0.51	4.84
Domain adaptation							
Model name	Variant	Exp. No.	SARI ($\uparrow$)	BLEU ($\uparrow$)	BERT Score ($\uparrow$)	ROUGE ($\uparrow$)	FKGL ($\downarrow$)
Llama-2-7b-chat-hf	None	20	35.15	0.52	0.94	0.64	8.49
	$R_{FKGL} + R_{Rel}$	21	45.71	0.63	0.96	0.78	7.50
	$R_{FKGL} + R_{OvS}$	22	74.41	0.73	0.97	0.89	7.74
Llama-3-8B-Instruct	None	23	38.73	0.43	0.93	0.60	6.36
	$R_{FKGL} + R_{Rel}$	24	49.97	0.45	0.92	0.59	7.27
	$R_{FKGL} + R_{OvS}$	25	51.06	0.57	0.91	0.72	8.39

overall, indicating the combined out-of-domain SFT and in-domain RLHF training were not complementary.

For *Llama-2-7b-chat-hf*, RL/RLHF alone *(Exp. 07–09)* was not as effective; however, the best performance across all models was achieved by combining out-of-domain SFT on Med-EASi with in-domain RLHF with the $R_{FKGL+OvS}$ reward. For this configuration, continued training through in-domain RLHF increased SARI from 35.15 to 74.41, BLEU from 0.52 to 0.73, and ROUGE from 0.64 to 0.89 and decreased FKGL from 8.49 to 7.74 *(Exp. 20 vs Exp. 22)*. The $R_{FKGL+Rel}$ reward *(Exp. 21)* also led to improvements over out-of-domain SFT alone but was less effective compared to the $R_{FKGL+OvS}$ reward. The domain adaptation results highlight that the top-performing domain-adapted LLM *(Exp. 22)* achieves comparable performance to the best in-domain models *(Exp. 10, 13, 04 and 14)*. These findings affirm the potential of combining SFT with RL/RLHF to enhance performance while minimizing annotation data requirements and associated costs.

4.3. Human evaluation

The above assessment focuses on adequacy and fluency, where adequacy refers to the preservation of the original meaning, and fluency assesses the readability of the simplification. Automatic adequacy and fluency metrics provide useful information; however, these metrics do not holistically assess the quality of simplifications [86]. To provide a comprehensive evaluation, we manually evaluated human- and

Table 4

Human evaluation of Human-generated simplification and AI-generated simplification.

Preservation of meaning								
Source	Human vs Llama 2				Human vs GPT-4			
	Human	AI	Both	None	Human	AI	Both	None
CDC	0	4.35	95.65	0	62.32	10.14	21.74	5.80
NCI	8.89	12.22	78.89	0	56.67	15.56	27.78	0
ACS	11.11	7.19	81.70	0	50.98	6.54	41.18	1.31
Overall	8.01	8.01	83.97	0	55.13	9.94	33.02	1.93
Understandability								
Source	Human vs Llama 2				Human vs GPT-4			
	Human	AI	Both	None	Human	AI	Both	None
CDC	18.84	1.45	75.36	4.35	43.48	42.03	8.70	5.80
NCI	30	17.78	46.67	5.56	35.56	48.89	14.44	1.11
ACS	19.61	10.46	65.36	4.58	35.95	56.86	5.23	1.96
Overall	22.4	10.6	62.2	4.8	37.5	51.3	8.7	2.6

AI-generated simplifications. Manual evaluations were performed by three nurse practitioners with extensive knowledge of patient care and education, who were not involved in the SimpleDC annotation. The nurse practitioners reviewed a subset of the test samples in an *A-B* comparison, to evaluate the preservation of medical meaning and understandability for a broad audience. In this A-B comparison, the

Table 5
Error Analysis of the best Llama 2 model (Exp 22) and the best GPT-4 model (Exp 19).

Source	Unchanged from original			Insertions in new			Deletions from original		
	Human	Llama 2	GPT-4	Human	Llama 2	GPT-4	Human	Llama 2	GPT-4
CDC	92.1%	98.2%	53.9%	9.3%	1.3%	57.3%	7.9%	1.8%	46.1%
NCI	72.8%	81.2%	41.5%	24.7%	13.2%	65.5%	27.2%	18.8%	58.5%
ACS	96.4%	92.9%	51.9%	2.9%	3.6%	60.8%	3.6%	7.1%	48.1%
Overall	88.6%	90.7%	49.3%	10.6%	5.9%	61.4%	11.4%	9.3%	50.7%

reviewers could indicate their preference for *A*, *B*, *Both*, or *None*. The reviewers were blinded to which samples, A or B, were human- or AI-generated, and the order (A vs. B assignments) was shuffled. The evaluation included: (1) human vs. the best performing Llama (*Exp. 22*) and (2) human vs. the best performing GPT-4 (*Exp. 19*). The manual evaluation comprised a total of 104 sentence pairs, including 23 from CDC, 30 from NCI, and 51 from ACS.

Table 4 presents the results of the human evaluation, which demonstrate human and Llama simplifications consistently preserve the meaning of the original text and use understandable language. This further validates the results in Table 3, specifically that the trained Llama model successfully learns the simplification task and emulates human simplification. The reviewers preferred human simplifications over GPT-4 simplifications to preserve meaning. The reviewers preferred the readability of GPT-4 over human simplification, suggesting a trade-off between fluency and adherence to the original content. This nuanced evaluation highlights the complexity of balancing adequacy and fluency in text simplification tasks, underscoring the need for a multifaceted approach in assessing model performance.

4.4. Error analysis

Despite GPT-4's advanced language capabilities and significantly larger model size, the trained Llama models outperformed GPT-4 in both automated metrics and manual evaluations. This unexpected result can be attributed to differences in task interpretation. The simplification task was designed to improve the accessibility of the text for patients without altering its meaning. To better understand the editing performed by human annotators and LLMs to create the simplified versions, we calculated the proportion of tokens that are unchanged (not edited), inserted, and deleted from the original text. Table 5 presents a summary of the unchanged, inserted, and deleted words in the SimpleDC test set, normalized by the number of tokens in the original text (comparing with the Llama 2 model outputs from *Exp. 22*). In creating SimpleDC, the annotators edited the NCI content more heavily than the ACS and CDC content because the NCI content started at a higher reading grade level (median 7.6 for ACS, 7.0 for CDC, and 8.3 for NCI). The Llama 2 model performed very light editing, deleting 9.3% of existing tokens and inserting 5.9% new tokens. In contrast, GPT-4 made many more edits, deleting 50.7% of existing tokens and inserting 61.4% new tokens. This discrepancy highlights the importance of task alignment between human expectations and AI model interpretations.

The GPT-4 performance was negatively impacted in the automatic and manual evaluations by over-generation. This tendency to add new text contrasts sharply with the more conservative approach of human annotators and the Llama models, which introduced new content at rates of approximately 10.6% and 5.9% respectively, with average word counts of 14.24 and 14.05. Although these insertions can enhance understandability by adding explanatory content or simplifying complex ideas, they can also stray from the original message. For example, the GPT model simplified the sentence "The main functions of the liver include the following:" into a 30+ word sentence ("The liver has some big jobs ... help break down fat in food".). While this type of generation

may be helpful for the audience, it does not align with the objective of the task.

The analysis underscores the differing abilities of the models to adapt to the inherent reading complexity of the source materials, notably between human annotators and trained Llama 2 and GPT-4 models. Specifically, both human annotators and the trained Llama 2 model demonstrated a nuanced ability to modulate their simplification efforts in response to the reading level complexity of the sources (ACS, NCI, and CDC). This adaptability is indicative of a strategic approach to text simplification, where the extent of modifications is calibrated based on the initial complexity of the source material. Conversely, the GPT-4 strategy appeared relatively uniform across sources, indicating a lack of sensitivity to the nuanced demands of the simplification task. Such uniformity may imply a limitation in prompt-based approaches to simplification.

5. Discussion

Our research explores the effectiveness of LLMs for simplifying educational health information, contrasting the performance of SFT, RL, RLHF, and in-context learning approaches. We explore the role of model training for specific tasks, revealing how GPT models, despite their general prowess, require targeted adjustments to excel in specialized contexts. We introduce a novel reward model designed to optimize simplification by focusing on readability, relevance, and fidelity to the original text's intent. This new model incorporates an automatic reading-level metric (FKGL) and the innovative Original vs. Simplified (OvS) reward informed by human feedback. Our findings demonstrate how RLHF can improve the quality of text simplifications, making complex medical information more accessible and understandable to patients. The integration of human feedback into the reinforcement learning process, through the OvS reward, marks a significant advancement in aligning AI-generated content with human expectations. Our study advances AI-driven text simplification for healthcare communication and demonstrates the potential for task-specific LLMs to bridge the gap between medical expertise and patient understanding.

Our results indicate that foundational LLMs, with their unique architectures and training protocols, respond differently to continued training. Specifically, Llama-2 and Llama-3 exhibited varied responses to SFT and RLHF. For Llama-2, SFT improved performance across all metrics compared to zero-shot; however, RLHF did not significantly alter performance relative to zero-shot. Combining SFT and RLHF for Llama-2 resulted in improved performance relative to either method used individually in both in-domain and domain adaptation experiments. In contrast, for Llama-3, both SFT and RLHF independently improved performance, with in-domain RLHF training yielding the best results. However, combining SFT and RLHF for Llama-3 did not enhance performance beyond that of RLHF alone. The architectural and training specifics for Llama-3 are not yet available, making it difficult to fully understand the differences between Llama-2 and Llama-3 and the reasons for these performance discrepancies.

The ability to accurately simplify complex medical texts without losing essential information is crucial for improving patient understanding and engagement. Our research illustrates the potential of using AI

to assist in this process but also highlights the challenges in ensuring the simplifications are appropriate and accessible to all patients. This emphasizes the need for a collaborative approach that combines technical innovation with input from healthcare professionals to ensure the simplifications meet the diverse needs of the patient population. By leveraging AI's capabilities alongside expert human feedback, we can better tailor health information to enhance patient comprehension and empowerment.

6. Conclusions

In this paper, we present SimpleDC, a carefully curated human-annotated text simplification corpus for digestive cancer education. The rich annotations of the dataset have been instrumental in training effective simplification models. We explore a range of learning approaches, including SFT, RL, RLHF, and prompt-based approaches. We introduced a new RLHF reward function, $R_{FKGL+OvS}$, which outperforms an existing RL reward function for text simplification. In terms of model performance, we observed that Llama models with task-specific training outperformed GPT-4. This highlights the impact of high-quality task-specific annotated data in improving model performance and simplification quality, including balancing comprehensibility and medical accuracy. Our results also demonstrate that RL/RLHF can be used to incorporate unlabeled text into training to improve simplification performance: (1) as a standalone training step, (2) in conjunction with SFT, and (3) as a domain adaptation technique. Although this represents a significant advancement, our findings also underscore the vast potential for future research in this area. Our work provides a robust foundation for further explorations, aiming to refine and expand the capabilities of NLP models to simplify complex medical information for enhanced patient understanding. Future research could explore integrating multimodal data sources, such as medical images or patient educational videos, alongside textual information to further enhance the simplification models' ability to provide comprehensive and accessible health education. Additionally, investigating the incorporation of feedback mechanisms from actual patients and healthcare professionals into the model training process could significantly improve the relevance and usability of simplified medical content.

7. Data availability and legal considerations

We provide the new SimpleDC corpus to the research community to advance health text simplification.[7] In developing SimpleDC, we utilized publicly available text data from the NCI, ACS, and CDC websites with the goal of advancing text simplification and public health communication. Our use of this data is in alignment with fair use principles, as the data was employed exclusively for academic and research purposes, without any commercial intent. We minimized data usage, extracting only the portions necessary to support our work on text simplification and public health communication. Furthermore, our use of the data does not adversely affect the market value or potential of the original content. The text data from the NCI, ACS, and CDC was specifically used to improve NLP methods that enhance the accessibility of health information, with all sources carefully attributed. SimpleDC is made publicly available, with detailed source attribution, ensuring full transparency and traceability.

CRediT authorship contribution statement

Md Mushfiqur Rahman: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Mohammad Sabik Irbaz:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kai North:** Writing – review & editing, Methodology, Investigation, Conceptualization. **Michelle S. Williams:** Writing – review & editing, Supervision, Data curation, Conceptualization. **Marcos Zampieri:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Kevin Lybarger:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability and legal considerations

We provide the new SimpleDC corpus to the research community to advance health text simplification.[8] In developing SimpleDC, we utilized publicly available text data from the NCI, ACS, and CDC websites with the goal of advancing text simplification and public health communication. Our use of this data is in alignment with fair use principles, as the data was employed exclusively for academic and research purposes, without any commercial intent. We minimized data usage, extracting only the portions necessary to support our work on text simplification and public health communication. Furthermore, our use of the data does not adversely affect the market value or potential of the original content. The text data from the NCI, ACS, and CDC was specifically used to improve NLP methods that enhance the accessibility of health information, with all sources carefully attributed. SimpleDC is made publicly available, with detailed source attribution, ensuring full transparency and traceability.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used Chat-GPT and Writefull to solicit editorial feedback regarding writing clarity and proofreading. All the scientific content and data interpretation remained solely the authors' contributions. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Appendix. Case study of simplification outputs

Table A.6 presents text simplification examples from Llama-2 and Llama-3 models using various fine-tuning and reinforcement learning approaches. Both Llama-2-7b and Llama-3-8b models generally produce concise and accurate simplifications. However, the Llama-3-70b model, while generating seemingly more readable outputs, tends to add additional information not present in the original text. These results underscore the importance of model choice and configuration in text simplification, highlighting the need for precise model selection and fine-tuning in domain-specific applications.

---

[7] https://github.com/mushfiqur11/simpledc-dataset.git

[8] https://github.com/mushfiqur11/simpledc-dataset.git

Table A.6
Simplification outputs of the same sentence by different models.

Source/Model	Exp. No.	Simplification output
Original	–	As food is digested, it passes from the stomach to the small intestine.
Target Simplification	–	As food is broken down in your body, it passes from the stomach to the section of the gut called the small intestine.
Llama-2-7b with SFT (SimpleDC) + RLHF (SimpleDC)	10	Food is digested as it passes from the stomach to the small intestine.
Llama-2-7b with SFT (MedEASi) + RLHF (SimpleDC)	22	At this point, food passes from the stomach to the colon.
Llama-3-8b with SFT (SimpleDC) + RLHF (SimpleDC)	14	Food is digested and it passes from the stomach to the small intestine.
Llama-3-8b with SFT (MedEASi) + RLHF (SimpleDC)	25	As food is digested, it goes from the stomach to the small intestine.
Llama-3-70b with SFT (SimpleDC)	06	When you eat food, it goes from the stomach to the small intestine. This is called the small bowel. It is like a long, thin, winding road.
GPT-4 with Self-correction	19	When we eat, the food goes from our stomach to our small intestine.
Original	–	Basal cell carcinomas are a type of skin cancer that can develop in the perianal skin.
Target Simplification	–	Basal cell cancer is a kind of skin cancer that can sometimes start in the perianal skin.
Llama-2-7b with SFT (SimpleDC) + RLHF (SimpleDC)	10	Basal cell carcinomas are a type of skin cancer that can develop in the perianal skin.
Llama-2-7b with SFT (MedEASi) + RLHF (SimpleDC)	22	Basal cell cancer is a type of skin cancer that can develop in the perianal skin.
Llama-3-8b with SFT (SimpleDC) + RLHF (SimpleDC)	14	Basal cell cancers are a type of skin cancer that can develop in the perianal skin.
Llama-3-8b with SFT (MedEASi) + RLHF (SimpleDC)	25	Basal cell carcinomas are a type of skin cancer that can develop in the perianal skin.
Llama-3-70b with SFT (SimpleDC)	06	Basal cell carcinomas are a kind of skin cancer that can grow in the perianal skin. This is called anal melanosis.
GPT-4 with Self-correction	19	Basal cell carcinomas are a kind of skin cancer. They can grow on the skin around the bottom.

References

[1] D.A. Kindig, A.M. Panzer, L. Nielsen-Bohlman, et al., Health literacy: a prescription to end confusion, 2004, Available from: https://www.ncbi.nlm.nih.gov/books/NBK216035/.

[2] R.J.W. Cline, K.M. Haynes, Consumer health information seeking on the Internet: the state of the art, Health Educ. Res. 16 (6) (2001) 671–692, http://dx.doi.org/10.1093/her/16.6.671, arXiv:https://academic.oup.com/her/article-pdf/16/6/671/9809432/160671.pdf.

[3] D.B. Friedman, L. Hoffman-Goetz, A systematic review of readability and comprehension instruments used for print and web-based cancer information, Health Educ. Behav. 33 (3) (2006) 352–373, http://dx.doi.org/10.1177/1090198105277329, PMID: 16699125.

[4] D.R. Hansberry, A. John, E. John, N. Agarwal, S.F. Gonzales, S.R. Baker, A critical review of the readability of online patient education resources from RadiologyInfo.org, Am. J. Roentgenol. 202 (3) (2014) 566–575, http://dx.doi.org/10.2214/AJR.13.11223, PMID: 24555593.

[5] P. Jindal, J.C. MacDermid, Assessing reading levels of health information: uses and limitations of flesch formula., Educ. Health Change Learn. Prac. 30 (1) (2017) Available from: https://pubmed.ncbi.nlm.nih.gov/28707643.

[6] B. Rimer, W.L. Jones, M.K. Keintz, R.B. Catalano, P.F. Engstrom, Informed consent: a crucial step in cancer patient education, Health Educ. Q. (1984) 30–42, Available from: https://pubmed.ncbi.nlm.nih.gov/11658652.

[7] M.E. Mills, K. Sullivan, The importance of information giving for patients newly diagnosed with cancer: a review of the literature, J. Clin. Nurs. 8 (6) (1999) 631–642, http://dx.doi.org/10.1046/j.1365-2702.1999.00296.x.

[8] L. van den Bercken, R.J. Sips, C. Lofi, Evaluating neural text simplification in the medical domain, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 3286–3292, http://dx.doi.org/10.1145/3308558.3313630.

[9] A. Devaraj, I. Marshall, B. Wallace, J.J. Li, Paragraph-level simplification of medical texts, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings

of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4972–4984, Available from: https://aclanthology.org/2021.naacl-main.395.

[10] Y. Feng, J. Qiang, Y. Li, Y. Yuan, Y. Zhu, Sentence simplification via large language models, 2023, arXiv preprint arXiv:2302.11957.

[11] R. McCabe, P.G. Healey, Miscommunication in doctor–patient communication, Top. Cogn. Sci. 10 (2) (2018) 409–424, http://dx.doi.org/10.1111/tops.12337.

[12] P. Rawla, T. Sunkara, A. Barsouk, Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors, Gastroenterol. Rev. 14 (2) (2019) 89–103, http://dx.doi.org/10.5114/pg.2018.81072.

[13] F. Teglia, P. Boffetta, Association between trends of mortality and incidence, survival and stage at diagnosis for six digestive and respiratory cancers in United States (2009–2013), Eur. J. Cancer Prev. 32 (2) (2023) 195–202, http://dx.doi.org/10.1097/CEJ.0000000000000766.

[14] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, et al., Llama 2: Open foundation and fine-tuned chat models, 2023, Available from: https://arxiv.org/abs/2307.09288.

[15] AI@Meta, Llama 3 model card, 2024, Available from: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[16] O.J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, et al., GPT-4, Technical Report, 2023, Available from: https://api.semanticscholar.org/CorpusID:257532815.

[17] T. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. 33, 1877–1901, Available from: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[18] Y. Cao, R. Shui, L. Pan, M.Y. Kan, Z. Liu, T.S. Chua, Expertise style transfer: A new task towards better communication between experts and laymen, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1061–1071, Available from: https://aclanthology.org/2020.acl-main.100.

[19] S. Nisioi, S. Štajner, S.P. Ponzetto, L.P. Dinu, Exploring neural text simplification models, in: R. Barzilay, M.Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 85–91, Available from: https://aclanthology.org/P17-2014.

[20] M. Maddela, F. Alva-Manchego, W. Xu, Controllable text simplification with explicit paraphrasing, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 3536–3553, Available from: https://aclanthology.org/2021.naacl-main.277.

[21] K. North, M. Zampieri, M. Shardlow, Lexical Complexity Prediction: An Overview, ACM Comput. Surv. 55 (9) (2022) http://dx.doi.org/10.1145/3557885.

[22] M. Shardlow, M. Cooper, M. Zampieri, CompLex — a new corpus for lexical complexity prediction from likert scale data, in: Proceedings of READI, 2020, Available from: https://aclanthology.org/2020.readi-1.9.

[23] M.J. Ryan, T. Naous, W. Xu, Revisiting non-english text simplification: A unified multilingual benchmark, 2023, arXiv preprint arXiv:2305.15678.

[24] A. Koptient, N. Grabar, Fine-grained text simplification in french: steps towards a better grammaticality, in: International Symposium on Health Information Management Research, 2020.

[25] S. Štajner, I. Calixto, H. Saggion, Automatic text simplification for spanish: Comparative evaluation of various simplification strategies, in: R. Mitkov, G. Angelova, K. Bontcheva (Eds.), Proceedings of the International Conference Recent Advances in Natural Language Processing, INCOMA Ltd. Shoumen, BULGARIA, Hissar, Bulgaria, 2015, pp. 618–626, Available from: https://aclanthology.org/R15-1080.

[26] N. Grabar, H. Saggion, Evaluation of automatic text simplification: Where are we now, where should we go from here, in: Y. Estève, T. Jiménez, T. Parcollet, M. Zanon Boito (Eds.), Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, ATALA, Avignon, France, 2022, pp. 453–463, Available from: https://aclanthology.org/2022.jeptalnrecital-taln.47.

[27] M. Shardlow, A survey of automated text simplification, Int. J. Adv. Comput. Sci. Appl. 4 (1) (2014) 58–70, http://dx.doi.org/10.14569/SpecialIssue.2014.040109.

[28] K. North, T. Ranasinghe, M. Shardlow, M. Zampieri, Deep learning approaches to lexical simplification: A survey, 2023, http://dx.doi.org/10.48550/arXiv.2305.12000.

[29] L. Ramadier, M. Lafourcade, Radiological text simplification using a general knowledge base, in: A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Springer International Publishing, Cham, 2018, pp. 617–627, Available from: https://link.springer.com/chapter/10.1007/978-3-319-77116-8_46.

[30] L. Zilio, L.B. Paraguassu, L.A.L. Hercules, G.L. Ponomarekano, L.P. Berwanger, M.J.B. Finatto, A lexical simplification tool for promoting health literacy, in: 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), 2020, Available from: https://aclanthology.org/2020.readi-1.11.

[31] B. Qenam, T.Y. Kim, M.J. Carroll, M. Hogarth, Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: Translation and evaluation, J. Med. Internet Res (ISSN: 1438-8871) 19 (12) (2017) e417, http://dx.doi.org/10.2196/jmir.8536, Available from: http://www.jmir.org/2017/12/e417/.

[32] S. Kandula, D. Curtis, Q. Zeng-Treitler, A semantic and syntactic text simplification tool for health content, in: AMIA Annual Symposium Proceedings, 2010, American Medical Informatics Association, 2010, p. 366, Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041424/.

[33] G. Leroy, J.E. Endicott, D. Kauchak, O. Mouradi, M. Just, et al., User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention, J. Med. Internet Res. 15 (7) (2013) e2569, Available from: https://www.jmir.org/2013/7/e144.

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

[35] J. Liang, C.H. Tsou, A. Poddar, A novel system for extractive clinical note summarization using EHR data, in: A. Rumshisky, K. Roberts, S. Bethard, T. Naumann (Eds.), Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 46–54, Available from: https://aclanthology.org/W19-1906.

[36] N. Kanwal, G. Rizzo, Attention-based clinical note summarization, in: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, SAC '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 813–820, http://dx.doi.org/10.1145/3477314.3507256.

[37] J.C. Feblowitz, A. Wright, H. Singh, L. Samal, D.F. Sittig, Summarization of clinical information: A conceptual model, J. Biomed. Inform. 44 (4) (2011) 688–699, Available from: https://www.sciencedirect.com/science/article/pii/S1532046411000591.

[38] Y. Guo, W. Qiu, Y. Wang, T. Cohen, Automated lay language summarization of biomedical scientific reviews, Proc. AAAI Conf. Artif. Intell. 35 (1) (2021) 160–168, http://dx.doi.org/10.1609/aaai.v35i1.16089.

[39] E. Abrahamsson, T. Forni, M. Skeppstedt, M. Kvist, Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language, in: S. Williams, A. Siddharthan, A. Nenkova (Eds.), Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations, PITR, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 57–65, Available from: https://aclanthology.org/W14-1207.

[40] H. Van, D. Kauchak, G. Leroy, AutoMeTS: The autocomplete for medical text simplification, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 1424–1434, Available from: https://aclanthology.org/2020.coling-main.122.

[41] R. Cardon, N. Grabar, Parallel sentence retrieval from comparable corpora for biomedical text simplification, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, INCOMA Ltd., Varna, Bulgaria, 2019, pp. 168–177, Available from: https://aclanthology.org/R19-1020.

[42] T. Sakakini, J.Y. Lee, A. Duri, R.F. Azevedo, V. Sadauskas, K. Gu, et al., Context-aware automatic text simplification of health materials in low-resource domains, in: E. Holderness, A. Jimeno Yepes, A. Lavelli, A.-L. Minard, J. Pustejovsky, F. Rinaldi (Eds.), Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, Online, 2020, pp. 115–126, Available from: https://aclanthology.org/2020.louhi-1.13.

[43] C. Basu, R. Vasu, M. Yasunaga, Q. Yang, Med-EASi: Finely annotated dataset and models for controllable simplification of medical texts, in: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, in: AAAI'23/IAAI'23/EAAI'23, AAAI Press, 2023, http://dx.doi.org/10.1609/aaai.v37i12.26649.

[44] A. Phatak, D.W. Savage, R. Ohle, J. Smith, V. Mago, Medical text simplification using reinforcement learning (TESLEA): Deep learning–based text simplification approach, JMIR Med. Inform. 10 (11) (2022) e38095, Available from: https://medinform.jmir.org/2022/11/e38095.

[45] K. Filippova, M. Strube, Dependency tree based sentence compression, in: M. White, C. Nakatsu, D. McDonald (Eds.), Proceedings of the Fifth International Natural Language Generation Conference, Association for Computational Linguistics, Salt Fork, Ohio, USA, 2008, pp. 25–32, Available from: https://aclanthology.org/W08-1105.

[46] K. Filippova, M. Strube, Sentence fusion via dependency graph compression, in: M. Lapata, H.T. Ng (Eds.), Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, 2008, pp. 177–185, Available from: https://aclanthology.org/D08-1019.

[47] M. Shardlow, R. Nawaz, Neural text simplification of clinical letters with a domain specific phrase table, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 380–389, Available from: https://aclanthology.org/P19-1037.

[48] R. Sun, H. Jin, X. Wan, Document-level text simplification: Dataset, criteria and baseline, in: M.-F. Moens, X. Huang, L. Specia, S.W.t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 7997–8013, Available from: https://aclanthology.org/2021.emnlp-main.630.

[49] Z. Sun, M. Wang, H. Zhou, C. Zhao, S. Huang, J. Chen, et al., Rethinking document-level neural machine translation, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3537–3548, Available from: https://aclanthology.org/2022.findings-acl.279.

[50] K.C. Sheang, H. Saggion, Controllable sentence simplification with a unified text-to-text transfer transformer, in: A. Belz, A. Fan, E. Reiter, Y. Sripada (Eds.), Proceedings of the 14th International Conference on Natural Language Generation, Association for Computational Linguistics, Aberdeen, Scotland, UK, 2021, pp. 341–352, Available from: https://aclanthology.org/2021.inlg-1.38.

[51] A. Farajidizaji, V. Raina, M. Gales, Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models, 2023, http://dx.doi.org/10.48550/arXiv.2309.12551, arXiv preprint arXiv:2309.12551:.

[52] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, et al., Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (1) (2020) Available from: https://dl.acm.org/doi/abs/10.5555/3455716.3455856.

[53] H.W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, et al., Scaling instruction-finetuned language models, 2022, arXiv Available from: https://arxiv.org/abs/2210.11416.

[54] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.A. Lachaux, T. Lacroix, et al., Llama: Open and efficient foundation language models, 2023, http://dx.doi.org/10.48550/arXiv.2302.13971, arXiv preprint arXiv:2302.13971.

[55] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, et al., Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, 35, Curran Associates, Inc., 2022, pp. 24824–24837, Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[56] J. Wang, J. Li, H. Zhao, Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 2717–2731, Available from: https://aclanthology.org/2023.findings-emnlp.179.

[57] T.H. Wu, L. Lian, J.E. Gonzalez, B. Li, T. Darrell, Self-correcting LLM-controlled diffusion models, 2023, http://dx.doi.org/10.48550/arXiv.2311.16090, arXiv preprint arXiv:2311.16090.

[58] L. Pan, M. Saxon, W. Xu, D. Nathani, X. Wang, W.Y. Wang, Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies, 2023, http://dx.doi.org/10.48550/arXiv.2308.03188, arXiv preprint arXiv:2308.03188.

[59] Y. Liu, T. Han, S. Ma, J. Zhang, Y. Yang, J. Tian, et al., Summary of chatgpt-related research and perspective towards the future of large language models, Meta-Radiol. (2023) 100017.

[60] S. Tian, Q. Jin, L. Yeganova, P.T. Lai, Q. Zhu, X. Chen, et al., Opportunities and challenges for ChatGPT and large language models in biomedicine and health, Brief. Bioinform. 25 (1) (2024) bbad493.

[61] A. Nakamachi, T. Kajiwara, Y. Arase, Text simplification with reinforcement learning using supervised rewards on grammaticality, meaning preservation, and simplicity, in: B. Shmueli, Y.J. Huang (Eds.), Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, Association for Computational Linguistics, Suzhou, China, 2020, pp. 153–159, Available from: https://aclanthology.org/2020.aacl-srw.22.

[62] D. Yanamoto, T. Ikawa, T. Kajiwara, T. Ninomiya, S. Uchida, Y. Arase, Controllable text simplification with deep reinforcement learning, in: Y. He, H. Ji, S. Li, Y. Liu, C.H. Chang (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online only, 2022, pp. 398–404, Available from: https://aclanthology.org/2022.aacl-short.49.

[63] X. Zhang, M. Lapata, Sentence simplification with deep reinforcement learning, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 584–594, Available from: https://aclanthology.org/D17-1062.

[64] W. Alkaldi, D. Inkpen, Text simplification to specific readability levels, Mathematics 11 (9) (2023) 2063, http://dx.doi.org/10.3390/math11092063.

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

[66] J. Li, Digestive cancer incidence and mortality among young adults worldwide in 2020: A population-based study, World J. Gastroint. Oncol. 14 (1) (2022) 278, http://dx.doi.org/10.4251/wjgo.v14.i1.278.

[67] J.P. Kincaid, R.P. Fishburne Jr., R.L. Rogers, B.S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Institute for Simulation and Training, University of Central Florida, 1975, Available from: https://stars.library.ucf.edu/istlibrary/56.

[68] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, 2017, http://dx.doi.org/10.48550/arXiv.1707.06347.

[69] P.F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017, Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[70] Q. Chen, Y. Peng, Z. Lu, BioSentVec: creating sentence embeddings for biomedical texts, in: 2019 IEEE International Conference on Healthcare Informatics, ICHI, IEEE, 2019, pp. 1–5, http://dx.doi.org/10.1109/ICHI.2019.8904728.

[71] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, et al., Don't stop pretraining: Adapt language models to domains and tasks, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8342–8360, http://dx.doi.org/10.18653/v1/2020.acl-main.740, Available from: https://aclanthology.org/2020.acl-main.740.

[72] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., Training language models to follow instructions with human feedback, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, 35, Curran Associates, Inc., 2022, pp. 27730–27744, Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[73] I. Csiszár, I-divergence geometry of probability distributions and minimization problems, Ann. Appl. Probab. (1975) 146–158, Available from: https://www.jstor.org/stable/2959270.

[74] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., PyTorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 32, Curran Associates, Inc., 2019, Available from: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[75] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al., Transformers: State-of-the-art natural language processing, in: Q. Liu, D. Schlangen (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45, Available from: https://aclanthology.org/2020.emnlp-demos.6.

[76] R. Jin, J. Du, W. Huang, W. Liu, J. Luan, B. Wang, et al., A comprehensive evaluation of quantization strategies for large language models, 2024, Available from: https://arxiv.org/abs/2402.16775.

[77] L. Biewald, Experiment tracking with weights and biases, 2020, Software available from wandb.com. Available from: https://www.wandb.com/.

[78] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., Lora: Low-rank adaptation of large language models, 2021, http://dx.doi.org/10.48550/arXiv.2106.09685, arXiv preprint arXiv:2106.09685.

[79] L. Xu, H. Xie, S.Z.J. Qin, X. Tao, F.L. Wang, Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment, 2023, http://dx.doi.org/10.48550/arXiv.2312.12148.

[80] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, PEFT: State-of-the-art parameter-efficient fine-tuning methods, 2022, https://github.com/huggingface/peft.

[81] Optimizing statistical machine translation for text simplification, Trans. Assoc. Comput. Linguist. 4 (2016) 401–415, Available from: https://www.aclweb.org/anthology/Q16-1029.

[82] K. Papineni, S. Roukos, T. Ward, W. jing Zhu, BLEU: a method for automatic evaluation of machine translation, 2002, pp. 311–318, http://dx.doi.org/10.3115/1073083.1073135.

[83] C.Y. Lin, F.J. Och, ORANGE: a method for evaluating automatic evaluation metrics for machine translation, in: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, COLING, Geneva, Switzerland, 2004, pp. 501–507, Available from: https://www.aclweb.org/anthology/C04-1072.

[84] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: International Conference on Learning Representations, 2020, Available from: https://openreview.net/forum?id=SkeHuCVFDr.

[85] C.Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81, Available from: https://www.aclweb.org/anthology/W04-1013.

[86] D. Kauchak, G. Leroy, Moving beyond readability metrics for health-related text simplification, IT Prof. 18 (3) (2016) 45–51, http://dx.doi.org/10.1109/MITP.2016.50, Available from: https://ieeexplore.ieee.org/document/7478479.