Evaluating LLMs for Targeted Concept Simplification for Domain-Specific Texts

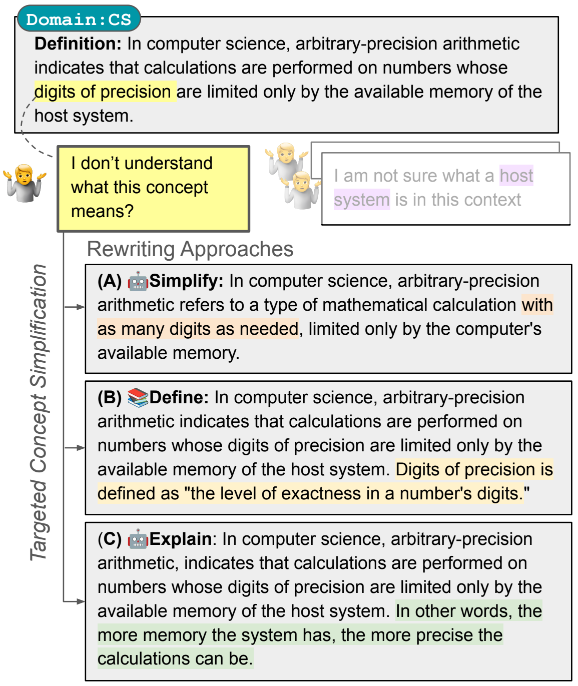
Sumit Asthana[†*] **Hannah Rashkin**[‡] **Elizabeth Clark**[‡]
Fantine Huot[‡] **Mirella Lapata**[‡]
[†]University of Michigan, Ann Arbor [‡]Google Deepmind
asumit@umich.edu
{hrashkin,eaclark,fantinehuot,lapata}@google.com

Abstract

One useful application of NLP models is to support people in reading complex text from unfamiliar domains (e.g., scientific articles). Simplifying the entire text makes it understandable but sometimes removes important details. On the contrary, helping adult readers understand difficult concepts in context can enhance their vocabulary and knowledge. In a preliminary human study, we first identify that lack of context and unfamiliarity with difficult concepts is a major reason for adult readers' difficulty with domain-specific text. We then introduce *targeted* concept simplification, a simplification task for rewriting text to help readers comprehend text containing unfamiliar concepts. We also introduce WikiDomains[1], a new dataset of 22k definitions from 13 academic domains paired with a *difficult concept* within each definition. We benchmark the performance of open-source and commercial LLMs, and a simple dictionary baseline on this task across human judgments of ease of understanding and meaning preservation. Interestingly, our human judges preferred explanations about the *difficult concept* more than simplification of the concept phrase. Further, no single model achieved superior performance across all quality dimensions, and automated metrics also show low correlations with human evaluations of concept simplification (~ 0.2), opening up rich avenues for research on personalized human reading comprehension support.

1 Introduction

Text simplification helps lay audiences understand challenging text by simplifying difficult terms, syntax, or discourse (Zhang and Lapata, 2017; Agrawal and Carpuat, 2023) or by adding content to elaborate on the text (Srikanth and Li, 2021). With advances in neural models, especially LLMs, sentence simplification has made considerable progress towards generating text at different reading grade levels (Kew et al., 2023). However, skilled adult readers face more challenges with lack of subject-matter knowledge (Guo et al., 2023). Supporting readers in understanding concepts they find personally difficult within a larger body of text not only expands their vocabulary, but also helps them develop a broader understanding of the topic (Kintsch, 1991; Van den Broek, 2010).



Definition: In computer science, arbitrary-precision arithmetic indicates that calculations are performed on numbers whose digits of precision are limited only by the available memory of the host system.

I don't understand what this concept means?

I am not sure what a host system is in this context

(A) Simplify: In computer science, arbitrary-precision arithmetic refers to a type of mathematical calculation with as many digits as needed, limited only by the computer's available memory.

(B) Define: In computer science, arbitrary-precision arithmetic indicates that calculations are performed on numbers whose digits of precision are limited only by the available memory of the host system. Digits of precision is defined as "the level of exactness in a number's digits."

(C) Explain: In computer science, arbitrary-precision arithmetic, indicates that calculations are performed on numbers whose digits of precision are limited only by the available memory of the host system. In other words, the more memory the system has, the more precise the calculations can be.

Figure 1: An example from the dataset, which consists of a *definition* and a potential *difficult concept* in the text that a reader may struggle with. The task is to rewrite the *definition* in a way that simplifies this concept for the reader. (a) Simplifies "digits of precision" to "as many digits as needed", (b) Adds the definition of "digits of precision", (c) Contextually explains that "digits of precision" refers to precision of calculations and how it relates to memory.

For example, in Figure 1, a person unfamiliar with the concept "digits of precision" will not un-

[*] Work done as student researcher at Google DeepMind.
[1] https://github.com/google-deepmind/wikidomains

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6208–6226
November 12-16, 2024 ©2024 Association for Computational Linguistics

derstand the definition of "arbitrary precision arithmetic". AI tools could help rewrite the text by lexically substituting "digits of precision" with "as many digits as needed" (simplifying) or by elaborating on the concept (defining or explaining). (a) Lexical simplification makes the definition understandable by reducing the overall complexity, perhaps losing some of the meaning. (b) Adding a definition of "digits of precision" may broaden the reader's vocabulary but does not explain its significance in the context of the overall definition (i.e. its implications for calculations and memory). (c) Providing a contextual explanation about "digits of precision" could more explicitly link the relation between memory and preciseness of calculations, enhancing comprehension (Van den Broek, 2010; Srikanth and Li, 2021).

In our study, we asked human raters to read definitions from 13 academic domains and identify the challenges in understanding them. We found that 50% of the reading difficulties arose from unfamiliar concepts, and annotators expressed the need for more context around them. Motivated by this, we present the new task of *targeted* concept simplification for rewriting text to support understanding of *difficult concept*s within the definitions' context. This task focuses on simplifying specific concepts that users struggle with, allowing for personalized reading support than simply rewriting an entire document at an easier reading level. Personalized support with *difficult concept*s can help readers receive more contextually-relevant information tailored to their background knowledge. For instance, a computer scientist reading a physics document might struggle with physics concepts but understand the mathematical terms, while someone without a math background might need help with mathematical terms (Guo et al., 2023).

To investigate this task, we collect a new dataset, WIKIDOMAINS, consisting of 22k definitions from Wikipedia. We collect definitions using article titles and leading statements from Wikipedia. Our definitions span 13 academic domains (e.g., business, education, etc., see Table 1) improving over existing datasets that are limited to a single domain (e.g., science) (August et al., 2022). We annotate a potential *difficult concept* in each definition using an automated heuristic (Biran et al., 2011).

We use this dataset to evaluate the performance of open-source and commercial LLMs on *targeted* concept simplification. We explore three methods for rewriting definitions: adding a dictionary definition of the *difficult concept*, prompting LLMs to simplify the *difficult concept*, and prompting LLMs to explain the *difficult concept* in context. We conduct human evaluations of all three approaches along three dimensions: 1) meaning preservation, 2) whether a reader who is unfamiliar with the *difficult concept* can understand the rewritten definition, and 3) whether the rewritten definition is easier to understand than the original. Our human evaluations demonstrate a clear preference towards strategies for contextual explanation of the *difficult concept* rather than lexical simplifications. However, we also find that LLMs need to improve further on dimensions of comprehension. Low to mild correlations of automated simplification metrics with human evaluations of comprehension and meaning preservation ($\sim$ 0.1-0.3) also indicate a need for better metrics to evaluate nuanced contextual explanations.

Domain	#Definitions
Food & Drink	1,403
Performing arts	322
Business & Economics	1,539
Politics & Government	2,267
Biology	7,200
Chemistry	957
Computing	2,083
Earth and Environment	1,314
Mathematics	1,747
Medicine & Health	2,939
Physics	741
Engineering	89
Technology	7
Total	22,561

Table 1: Domains and number of definitions in each domain in the WIKIDOMAINS dataset.

In summary, our main contributions include:

- Introducing targeted concept simplification as a task for supporting readers as they encounter difficult concepts in text.
- Analysis from an annotation study examining the difficulties humans face in reading and the possible utility of assistance in understanding difficult concepts.
- WIKIDOMAINS, a dataset of 22k challenging domain-specific definitions collected from Wikipedia with automatically-annotated difficult concepts.

- Human evaluations of the performance of open-source and commercial LLMs on our task across multiple quality dimensions, including analysis of different prompting strategies and automatic metrics.

2 Background

Cognitive Support and Human Reading Comprehension Successful reading comprehension is key to integrating new knowledge and fostering learning from text (Lorch Jr and van den Broek, 1997; Dunietz et al., 2020). Cognitive theories suggest that comprehension is a multi-stage process that primarily involves 1) constructing a local meaning representation of text such as concepts, facts, and their relations (Graesser et al., 1994), and 2) forming a schema and filling in gaps using background knowledge to create a "mental picture" of what the text is about (Kintsch and Van Dijk, 1978; Bartlett, 1995). Adult readers lacking domain knowledge can be supported by explicit cues, such as examples and explanations, to help them construct better mental representations of ideas from the text (Kintsch, 1991; Van den Broek, 2010).

Text Simplification Reducing reading-level complexity and syntax (Garbacea et al., 2021) in text simplification benefits specific audiences like students, second language learners, and individuals with dyslexia (Paetzold and Specia, 2016; Bingel et al., 2018), but may not enhance comprehension for general adult readers (Garbacea et al., 2021). Contextual explanations can enhance comprehension but findings from studies of elaborating events in news domains (Srikanth and Li, 2021) may not be the same as difficulty with concepts in academic texts. While Wikipedia and news corpora (Kauchak, 2013; Xu et al., 2015; Zhang and Lapata, 2017) have advanced text simplification, they focus more on syntax and discourse difficulties than on academic concepts. Similarly, lexicons are limited to medicine (Elhadad and Sutaria, 2007; Ong et al., 2007) and science concepts (August et al., 2022), highlighting the need for a multi-domain corpus to advance personalized simplification for a general audience.

Complex Terms and Jargon Lexical simplification systems (Paetzold and Specia, 2016) have been shown to benefit children, people with language impairments or medical jargon simplification (Fatima and Strube, 2023; Joseph et al., 2023).

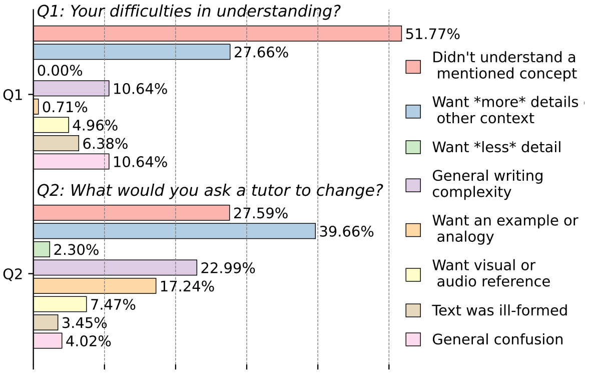


Figure 2: Results of annotator study: We asked annotators to read complex text for (1) what made the text difficult for them to understand and (2) how they would want a tutor to edit the text to help their understanding.

However, beyond lab studies, it is challenging to specify reader knowledge in large-scale evaluations. Proxies for audience knowledge include specialized lexicons (Paris, 1988; Elhadad and Sutaria, 2007), coarse indicators such as reading grade level (Agrawal and Carpuat, 2023), or binary indicators to denote science knowledge audience (August et al., 2022). Guo et al. (2023) highlighted the challenge of specifying audience knowledge at a finer level, suggesting the use of domain as a proxy for concept familiarity. Building on this, we provide a multi-domain corpus of challenging definitions to specify fine-grained audience levels. Unlike prior work on generating definitions (August et al., 2022) or simplifying all difficult concepts (Fatima and Strube, 2023), we focus on rewriting definitions to address specific concept difficulties, enabling readers to leverage their background knowledge and improve comprehension (Kintsch, 1991; Rello et al., 2013). While previous tools explored simple strategies like adding definitions for complex terms (Bingel et al., 2018), we evaluate LLMs that can provide contextual explanations (Srikanth and Li, 2021).

3 What AI assistance can benefit reading domain specific text

To better motivate the scope of this task, we investigate adult readers' difficulties with domain-specific text and what types of help they would want from an AI-tutor. We randomly selected a set of 900 text examples from Wikipedia-derived definitions spanning 13 domains (see Section 5 for details about definitions and domain selection). For each example, we ask a human annotator to respond in free-text to: 1) the reasons for difficulty (if any) when reading the definitions, and 2) what they would ask a tutor to change in the definition if they faced a difficulty. In Figure 2, we report recurring themes from annotator responses to both questions for cases where annotators had difficulty reading from these examples (categories were agreed on by the authors, see details in Appendix Section A).

These results suggest that specific *difficult concepts* used in the definitions were one of the most frequent reasons for reading difficulty (52% of definitions had such difficulty), and annotators frequently asked for help from a tutor with these concepts (28%). This indicates that our proposed task, targeted concept simplification, is an important subtask for simplification aimed to resolve a key challenge for lay adult readers. When asking a tutor for help, annotators also explicitly asked for **more** details on the *difficult concept* (rather than less). This suggests that contextual elaborations (Srikanth and Li, 2021) for these *difficult concept*s are a better alternative over lexical simplifications to support their comprehension and knowledge. Annotators also asked for examples/analogies (17%), visual/audio aids (8%), or identified general issues with the writing complexity (23%, e.g., an issue with general reading level or syntactic complexity), though to a lesser degree. The majority of our annotators had an above high-school level educational qualification (see Table 8 in Appendix), suggesting that unfamiliar concepts in context is a greater challenge for skilled adult readers than simply difficult words or syntax.

4 Task Definition

We present *targeted concept simplification*: a text simplification task focused on specific words or phrases that readers find difficult to understand. This setup allows for personalized and controlled rewriting of difficult concepts. Our initial user study (Section 3) shows that unfamiliar words or phrases often hinder comprehension.

The task of targeted concept simplification is to rewrite an input definition containing a concept c to make it understandable to someone unfamiliar with the concept. For example, Figure 1 shows the definition of the term "arbitrary precision arithmetic." The task is to rewrite the definition to help someone unfamiliar with the *difficult concept* "digits of precision." Possible approaches could involve replacing "digits of precision" with a simpler phrase like "as many digits as needed," explaining it within the definition, or perhaps even adding examples, analogies, or illustrations. The usefulness of each strategy will depend on its ability to complement the reader's existing knowledge about the topic Kintsch (1991). Unlike other text simplification tasks, our task targets simplifying concepts *difficult for the reader* rather than simplifying the entire text.

5 The WikiDomains Dataset

To support research on targeted concept simplification, we introduce a dataset of 22k definitions from 13 academic domains[2], where each *definition* is a 1–2 sentence explanation of a *term*[3]. Within each definition, we select a *difficult concept*—a word or phrase that could impede the reader's ability to comprehend the definition as whole. We take inspiration from August et al. (2022) who collected definitions from Wikipedia science glossaries; however, instead of glossaries, we directly collect definitions from Wikipedia articles of concepts spanning 13 *domains* (see Table 1 for list of domains).

To collect definitions, we start with Johnson (2021)'s dataset that contains all English Wikipedia articles with probabilities of belonging to high-level domains. These domains are broad academic topics (e.g., Physics, Economics) that Wikipedia editors identified through consensus (Asthana and Halfaker, 2018). We refer to each Wikipedia article title as a *term* and take the first sentence of its lead section as its *definition* (August et al., 2022). For every domain, we first select articles with domain assignment probabilities greater than a threshold δ_{domain}.[4] To filter out low-importance articles that could be named entities, unimportant places or things, we also excluded articles having a pagerank percentile score less than δ_{pr}.[5] Finally, we also excluded articles that were additionally members of domains related to named entities, events, or things (e.g., Biography). Table 1 summarizes the 13 domains and the number of articles in each domain that the final WikiDomains dataset contains (more details in Appendix B). We also provide

[2] https://github.com/google-deepmind/wikidomains

[3] We call the concept being explained by the original definition a *term* to avoid confusion with *difficult concepts* present within the definitions.

[4] Manual inspection of topic assignments for 50 Wikipedia articles suggested that a threshold of 0.7 was reasonable to identify articles belonging to a domain.

[5] We determined the threshold as 0.1 through a manual examination of 100 articles.

	train	dev	test
# definitions	15,873	3,384	3,304
avg # tokens	22.75	22.63	22.61
total # tokens	361,066	76,572	74,691
vocab size	42,356	15,576	14,911

Table 2: Statistics on WIKIDOMAINS definitions broken down by split; #tokens and vocabulary size are calculated by splitting the definitions on whitespace and removing punctuation.

each *term*'s lead section in the dataset for future research.

We select a training, development, and test split of 15,873/3,384/3,304 examples (see Table 2 for more details about the data splits.) We conduct our experiments in a zero- or few-shot setting without using the training or development data, but we publicly release the full set to facilitate future research.

5.1 Difficult Concept Identification

For each definition, we automatically label a potential *difficult concept* that could impede a reader's comprehension. Lay readers will be more familiar with concepts that are popularly mentioned across Wikipedia (e.g., "bacteria") than concepts that only occur in articles of a specific domain (e.g., "Phytosterol"). Thus, following prior work on approximating word difficulties using specificity-based measures (Biran et al., 2011), we use a domain-specificity measure to score concept difficulty for a lay audience.

First, we identify candidate concepts c mentioned in each *term*'s definition using Wikidata (Vrandečić and Krötzsch, 2014)[6]. We then order the candidates by a score of how specific they are to the *term*'s domain D_t. This is measured by the ratio of how many articles $\mathcal{A}$ the concept c appears in within this domain compared to across Wikipedia generally:

$$\frac{\sum_{\mathcal{A}\in\mathcal{D}_t} \mathbb{1}[c \in \mathcal{A}]}{\sum_{\mathcal{A}\in\mathcal{D}_{\text{all}}} \mathbb{1}[c \in \mathcal{A}]} \tag{1}$$

For each definition, we select one *difficult concept* out of the top-k identified candidates[7]. If we could not identify any *difficult concept* in the definition, we instead chose a difficult concept using the age of acquisition lexicon (Kuperman et al., 2012), which provides an average age when different words are acquired as a proxy for its difficulty.

6 Experiments

We explore the performance of existing NLP tools on targeted concept simplification and possible avenues for future improvement. More concretely, we investigate the following research questions:

RQ1: What is the performance of out-of-the-box NLP tools in this task?

RQ2: Which types of simplification strategies improve human understanding of *difficult concept*s and the definitions that they appear in?

RQ3: For *targeted* concept simplification, how do human evaluations compare to automatic metrics commonly used in text simplification?

We perform experiments on the WIKIDOMAINS test data created in Section 5. As an additional evaluation set, we also use the scientific definitions dataset from August et al. (2022) (SCIDEF) that contains definitions of science terms extracted from Wikipedia glossaries and MedQuAD (Ben Abacha and Demner-Fushman, 2019). We perform the same post-processing on SCIDEF as with WIKIDOMAINS to select a *difficult concept* within each definition.

6.1 Models

To explore the benchmark performance on this data, we selected four popular LLMs: GPT-4 (OpenAI, 2023), PaLM-2 (Anil et al., 2023), Falcon-40b (Almazrouei et al., 2023), and BLOOM-170b (BigScience Workshop, 2023). For the open-source models, we selected the instruct versions with the highest number of parameters available.

We also included a baseline approach of dictionary look-up (non-LLM) to compare to the LLMs. For this baseline, we looked up a definition of the *difficult concept* and simply appended it to the end of the original definition. We retrieved the definition from Wikidata (Vrandečić and Krötzsch, 2014), falling back on WordNet (Miller, 1994) if the term was not found in Wikidata (or stated that the *difficult concept*'s definition could not be found if both sources failed).

6.2 Simplification Strategies and Prompts

In our preliminary user study (Section 3), we found that users frequently indicated they would like more details and context, as well as more general breakdowns of writing complexity. These two

[6] We used the Wikidata extension in spaCy to identify concepts in definitions that have corresponding Wikidata entries.

[7] We chose k as 2 based on manual assessment of 100 definitions.

	Name	Description
Human Eval.	Meaning preservation ($\mathcal{H}_{\text{MP}}$)	Human evaluation of whether the rewritten definition preserves the meaning of the original definition (on a 5-point Likert scale; 5 = perfectly preserved).
	Rewrite understanding ($\mathcal{H}_{\text{RU}}$)	Human evaluation of whether a reader can understand the rewritten definition if they do not know the *difficult concept* (1 = yes, 0 = no).
	Rewrite easier ($\mathcal{H}_{\text{RE}}$)	Human evaluation of whether the rewritten definition is easier to understand than the original definition (1 = rewrite is easier; 0 = the original is easier or both are similar).
Automatic Eval.	Density	Density (Grusky et al., 2018) is a measure of how extractive the rewritten definition is from the original definition.
	BLEU-4	BLEU-4 score (Papineni et al., 2002) of the rewritten definition with respect to the original definition.
	BERTSCORE (BertSc)	BERTSCORE (Zhang* et al., 2020) of the rewritten definition with respect to the original definition.
	Change in length (ΔLen)	Average difference between the lengths (in number of tokens) of the rewritten and the original definition (positive means the rewritten definition is longer than the original).
	Change in age of acquisition (ΔAoA)	Average difference of the top-10 percentile of the age-of-acquisition (Kuperman et al., 2012) of the words between the rewritten and the original definition (positive means the rewritten definition uses less complex words).
	Change in Flesch ease (ΔFlesch)	Average difference of the Flesch reading ease (Flesch, 1948) between the rewritten and the original definition (positive means the rewritten definition is at an easier reading level than the original).

Table 3: Human and automatic metrics used to evaluate LLM rewritten text for concept simplification.

strategies also correspond to familiar approaches for general text simplification tasks that rely on elaboration (Srikanth and Li, 2021) and lexical changes (Paetzold and Specia, 2016), respectively.

We chose two different prompts for the LLMs that reflect these two simplification strategies. In our first prompt, we show the model the *term*, *definition*, and *difficult concept*. We instruct the model to rewrite the definition, "integrating an explanation" for the difficult concept ($explain$). The second prompt is similar, except we instruct the model to rewrite the definition "simplifying" the difficult concept word ($simplify$).

We chose the specific wording of the prompts for the two strategies after a small scale analysis of results with a few candidate prompts. We describe the candidate prompts, and the full phrasing of the final selected prompts in the Appendix (Section C). We report results using 3-shot settings for the LLMs.[8]

6.3 Human Evaluation

We asked human raters to rate the rewritten definitions along dimensions of meaning preservation and ease of understanding of the rewrites with respect to both the *difficult concept* and the original definition. Specifically, we asked them about (1) *meaning preservation*, denoted as $\mathcal{H}_{\text{MP}}$: how much does the rewrite preserve the meaning of the original definition on a Likert scale of 5; (2) *rewrite understanding*, denoted as $\mathcal{H}_{\text{RU}}$: If a reader is unfamiliar with the *difficult concept*, would they be able to understand the rewrite (Yes/No); (3) *rewrite easier*, denoted as $\mathcal{H}_{\text{RE}}$: Is the rewrite easier to understand than the original? These dimensions are summarized in the first three rows of Table 3. We obtain judgments from 3 human raters per example for 120 randomly selected examples from the WIKIDOMAINS dataset and 60 randomly selected examples from the SCIDEF dataset (2880 judgments in total). We provide exact wording of the question, their rationale and annotator background in the Appendix (Section D). In Appendix Table 14 we show Krippendorff's alpha agreement scores for each human evaluation dimension.

6.4 Automated Metrics

We investigate the utility of commonly used simplification automated metrics for our task and compare them to human judgments. Because our data are reference-less, we cannot use reference-based metrics like SARI (Xu et al., 2016). Instead, we estimate changes in complexity using the difference between the rewritten and the original definition in terms of: (1) age of acquisition (AoA; Kuperman et al. 2012), (2) Flesch reading ease (Flesch, 1948), and (3) token length. We also measure density (Grusky et al., 2018), which scores how extractive the rewritten defintion is from the original. Lastly, we use BLEU (Papineni et al., 2002) and

[8] We also experimented with a zero-shot settings with results in the Appendix.

		Model	$\mathcal{H}_{\text{MP}}$	$\mathcal{H}_{\text{RE}}$	$\mathcal{H}_{\text{RU}}$
WIKIDOMAINS		Baseline	4.66	0.31	0.79
	simplify	Bloomz	4.25	0.20	0.53
		Falcon	3.82	0.59	0.67
		PaLM2	**4.71**	0.12	0.46
		GPT4	4.43	0.29	0.75
	explain	Bloomz	4.53	0.43	0.69
		Falcon	4.30	0.55	0.71
		PaLM2	4.64	0.59	0.66
		GPT4	4.47	**0.75**	**0.82**
		Model	$\mathcal{H}_{\text{MP}}$	$\mathcal{H}_{\text{RE}}$	$\mathcal{H}_{\text{RU}}$
SCIDEF		Baseline	4.21	0.40	0.61
	simplify	Bloomz	4.58	0.10	0.46
		Falcon	4.24	0.25	0.57
		PaLM2	4.57	0.12	0.62
		GPT4	4.47	0.12	**0.88**
	explain	Bloomz	4.39	0.53	0.65
		Falcon	4.29	0.53	0.86
		PaLM2	**4.86**	0.18	0.54
		GPT4	4.09	**0.58**	0.87

Table 4: Human evaluations of LLM-generated rewrites for targeted concept simplification for the metrics $\mathcal{H}_{\text{MP}}$ (meaning preservation), $\mathcal{H}_{\text{RE}}$ (rewrite easier), and $\mathcal{H}_{\text{RU}}$ (rewrite understanding) in 3-shot setting.

	Prompt	$\mathcal{H}_{\text{MP}}$	$\mathcal{H}_{\text{RE}}$	$\mathcal{H}_{\text{RU}}$
WIKI DOMAINS	simplify	4.30	0.30	0.60
	explain	**4.48***	**0.57***	**0.72***
SCIDEF	simplify	**4.47**	0.15	0.64
	explain	4.41	**0.45***	**0.73***

Table 5: Comparison of the prompts – simplify and explain – for the human evaluation metrics. All results are significantly different (ttest, $p < 0.01$) marked by *, except for the comparison of meaning preservation on SCIDEF. Results are from the 3-shot setting.

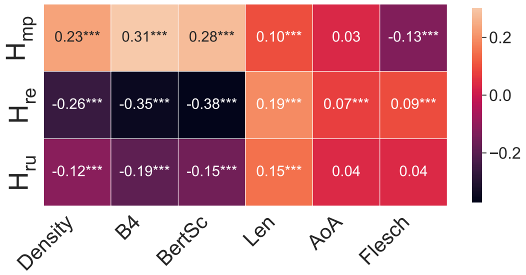


Figure 3: Pearson correlations between automated metrics and human evaluations ($^{***}: p < 0.005, ^{**}: p < 0.05, ^{*}: p < 0.01$).

BERTSCORE (Zhang* et al., 2020) to score the similarity of the rewritten definitions with respect to the original definition. Table 3 presents a full list of the human and automatic evaluations.

6.5 Model Rankings

Table 4 summarizes the evaluations of the rewrites based on human judgment according to the meaning preservation ($\mathcal{H}_{\text{MP}}$), whether the rewrite is easier to understand than the original ($\mathcal{H}_{\text{RE}}$), and whether the rewrite can be understood for someone unfamiliar with the *difficult concept* ($\mathcal{H}_{\text{RU}}$).

We observe that no model excels in all dimensions, though GPT-4 performs best on average. Different models have distinct strengths; for instance, PaLM2 excels in meaning preservation but its rewrites are rarely easier to understand. Additionally, the dictionary-lookup baseline performs comparably well to the LLM models.

Weaker scores on the $\mathcal{H}_{\text{RU}}$ and $\mathcal{H}_{\text{RE}}$ dimensions compared to the $\mathcal{H}_{\text{MP}}$ dimension across all models, indicates opportunities for future research to improve these scores.

6.6 Simplifying vs Explaining

We discuss the differences of the human evaluations for the two prompt strategies—$explain$ and $simplify$ (introduced in Section 6.2). Table 5 shows the comparison of the prompt strategies on human evaluation dimensions averaged across the four LLMs. Human raters clearly preferred rewrites where the model was asked to explain the *difficult concept* in both $\mathcal{H}_{\text{RE}}$ and $\mathcal{H}_{\text{RU}}$ judgments. On WIKIDOMAINS data, human raters also had a significant preference towards the "explain" strategy when judging meaning preservation (the difference in $\mathcal{H}_{\text{MP}}$ on SCIDEF was not significant). This aligns with some of our observations from our initial user study (Section 3), which found that humans preferred adding more context (40%) as opposed to simpler word substitutions (23%). This highlights that adding elaborative details is very important towards facilitating human understanding centered around difficult concepts.

6.7 Correlation between Human and Automated Evaluation

Figure 3 shows the correlations between automated metrics and human evaluations ($\mathcal{H}_{\text{MP}}$, $\mathcal{H}_{\text{RE}}$, $\mathcal{H}_{\text{RU}}$). We observe no single metric captures all the dimensions of human evaluations. BLEU-4, Density, and BERTSCORE show mild correlations with $\mathcal{H}_{\text{MP}}$ as they capture similarity of text. However, none of

(a) Lustre or luster is the way light interacts with the surface of a crystal, rock, or mineral. Mineral is defined as "naturally occurring usually inorganic substance that has a (more or less) definite chemical composition and a crystal structure."
(b) Quality control, or QC for short, is a process by which entities review the quality of all factors involved in production. Entity is defined as "something that exists in the identified universe.".

Table 6: Examples from the dictionary baseline, which appends a definition shown in blue. (a) Added definition has domain-specific jargon that may be unfamiliar to the reader. (b) Added definition is vague, not accounting for the context.

Economics: The Financial Stability Board (FSB) is an international body that monitors and makes recommendations about the ~~global financial system~~world's money. (**global financial system**) [PaLM2]
Biology: ~~Jungle is an area covered with dense vegetation dominated by large trees, often tropical~~A jungle is a region filled with thick plant life, often dominated by large trees, typically found in tropical areas. (**vegetation**) [GPT4]
Computing: Prolog is a logic programming language associated with artificial intelligence and computational linguistics. (**linguistics**) [Bloomz] [no change]

Table 7: Examples of concept simplification behavior for the *simplify* 3-shot prompt from three domains: Economics, Biology and Computing. The *difficult concept* is shown in **bold** at the end of the definition. Deletions are show in ~~red~~; additions in blue.

automated metrics correlate with either $\mathcal{H}_{\text{RE}}$ or $\mathcal{H}_{\text{RU}}$ which capture comprehension related to the target *difficult concept*. Even Flesch reading ease, which is commonly used in text simplification setups, is not adequate for measuring whether the rewrites are easier understood. This result calls for new metrics, beyond aggregate similarity measures, to evaluate comprehension at the semantic level of concepts.

In the Appendix (Table 15), we show the full automated metric scores for each model, which may be useful in characterizing some qualities of the outputs. For example, Bloomz and PaLM2 make relatively few changes to the text (low ΔLen), and GPT4 chose considerably easier words in the rewrite (high ΔAoA). GPT4's low meaning preservation rating suggests choosing easier words is not always desirable (Table 4). However, given the low correlations with human scores, we generally keep our observations about relative model rankings to the human judgment.

7 Discussion

We close our paper by discussing the research questions we posed in our experiments and how they may relate to future improvements on this task.

Can LLMs support Contextual Explanations of Difficult Text? Despite their instruction-following capabilities, human evaluations indicate that there's still considerable room for improvement at this task. Human judgments (Table 4) reveal that no model excels universally, each having its own strengths and weaknesses. All models tend to perform better at meaning preservation, though other dimensions may be more crucial for enhancing broader comprehension (Kintsch, 1991).

Our evaluations support prior findings that dictionary-based methods for simplification are limited by availability and their inability to personalize to the reader's context and background knowledge (August et al., 2022). Table 6 shows two examples from the dictionary baseline that are either too vague or too complex to be useful to a lay reader. However, we find that LLMs outperform the deterministic dictionary look-up baseline only by a small margin depending on the dimension of quality. In examples of output (Table 7), we can see failure cases where the models either over-simplify text beyond just the *difficult concept* or make no changes to the definition at all. LLMs have been found to be useful for reading-grade level simplifications (Agrawal and Carpuat, 2023), yet they seem to struggle with making fine-grained simplifications at the level of *difficult concept*s, calling for more careful tooling for targeted simplification. While more custom prompts may elicit desired simplifications from LLMs, we cannot expect lay audience to be familiar with such prompting (Zamfirescu-Pereira et al., 2023).

Strategies Supporting Readers in Understanding Difficult Text. Open-ended human feedback about reading difficulties (Section 3) as well as human judgment of differences between prompts (Section 6.6) support the idea that adult readers may prefer additional details and context addition to understand *difficult concept*s. This echoes prior cognitive science work that cues in text (e.g., explanations, examples, analogies) enable readers to effectively utilize their background knowledge for comprehension (Kintsch, 1991; Van den Broek, 2010).

Better Evaluations to Support Text Understanding. As shown in prior work (Alva-Manchego et al., 2021), we find that automated metrics cannot capture fine-grained differences in simplification. While there are some correlations between meaning preservation and BLEU-4 and BERTSCORE, we did not observe clear correlations of automated metrics with other dimensions of comprehension, such as alleviating difficulty with an unfamiliar concept. We observe that many of these metrics rely on brittle lexical scoring (Alva-Manchego et al., 2021), and it may be necessary for automated metrics to take more of the underlying concept structure of the rewrites into account in order to adequately judge whether the *difficult concept* has been explained sufficiently. A separate LLM to score these dimensions more reliably is an option (Wang et al., 2023; Chen et al., 2023; Gao et al., 2024); however, human judgment currently remains the best standard in this task.

8 Conclusion

To support comprehension of domain-specific text for adult readers, we introduced the task of *targeted* concept simplification to study fine-grained simplification of difficult concepts in context. Our human annotation study highlights the importance of aiding users' understanding of these concepts in domain-specific texts. We also introduced WIKIDOMAINS, a dataset of 22k definitions across 13 academic domains, to support this task. Our findings show a preference for strategies that add explanatory details over simplifying difficult concepts. Human evaluations of LLM rewrites indicate considerable room for improvement, especially for personalized help with difficult concepts.

9 Limitations

Difficulty with concepts varies based on personal knowledge. Thus, it is challenging to build large-scale evaluation corpora for domain-specific concepts. Our dataset of domain-specific concepts is a first step, providing a foundation for future work to study comprehension across domains.

While we used popular LLMs at the time of conducting the human evaluations, we also acknowledge that some of our LLMs may no longer be state-of-the-art when submitting the work. However, we will release our dataset and have described our experimental setup to promote reproducibility of the results with newer LLMs.

We evaluated our work by asking human raters to rate whether they can understand the definitions. However, human reading comprehension is also goal-directed, and different reading goals will evoke different needs for details (Dunietz et al., 2020). The details needed could differ depending on using the text for one's own understanding or using it for communicating it with other people about specific aspects. E.g., a lawyer communicating with engineers about the risks of a technology may need focus on the applications rather than just the understanding of technical concepts. Future work can evaluate how supporting readers with concept simplifications in documents (e.g., explanations, examples, analogies, and illustrations) help them develop a better understanding of the domain in pre-post tests.

10 Ethical Considerations

We extract our domain-specific definitions dataset from Wikipedia, which is publicly available and accessible to all. However, Wikipedia content has a Global North bias because of its editor base, and concepts in our domain-specific dataset will reflect this bias. We also acknowledge the broader educational implications of making definitions easier to understand, and that using LLMs could introduce false information. While in our work we did not observe instances of hallucinations, LLMs may introduce false information when rewriting entire documents or narratives, and we need robust measures to validate the faithfulness of rewritten definition in addressing concept difficulty and providing correct facts. While our evaluations attempt to provide initial insights into LLM's behavior with difficult concepts in domain-specific text, we also acknowledge that concept difficulty is a complex construct, and it can be dependent on a reader's age, educational, and professional background, which future evaluations should consider.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback and suggestions. We'd also like to thank Priyanka Agrawal and Slav Petrov for their feedback on the paper manuscript, as well as the other members of the Google DeepMind community for their guidance throughout the project. We also thank the anonymous individuals who participated in our human evaluations and rated the definitions.

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore. Association for Computational Linguistics.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification. *Computational Linguistics*, 47(4):861–889.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Sumit Asthana and Aaron Halfaker. 2018. With few eyes, all hoaxes are deep. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating scientific definitions with controllable complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8298–8317, Dublin, Ireland. Association for Computational Linguistics.

Frederic Charles Bartlett. 1995. *Remembering: A study in experimental and social psychology*. Cambridge university press.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):1–23.

BigScience Workshop. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.

Mehwish Fatima and Michael Strube. 2023. Cross-lingual science journalism: Select, simplify and rewrite summaries for non-expert readers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1843–1861, Toronto, Canada. Association for Computational Linguistics.

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32 3:221–33.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *ArXiv*, abs/2402.01383.

Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable prediction of text complexity: The missing preliminaries for text simplification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1086–1097, Online. Association for Computational Linguistics.

Arthur C Graesser, Murray Singer, and Tom Trabasso. 1994. Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Yue Guo, Joseph Chee Chang, Maria Antoniak, Erin Bransom, Trevor Cohen, Lucy Lu Wang, and Tal August. 2023. Personalized jargon identification for enhanced interdisciplinary communication. *ArXiv*, abs/2311.09481.

Isaac Johnson. 2021. Wikipedia Article Topics for All Languages (based on article outlinks).

Sebastian Joseph, Kathryn Kazanas, Keziah Reina, Vishnesh Ramanathan, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2023. Multilingual simplification of medical texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16662–16692, Singapore. Association for Computational Linguistics.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Walter Kintsch. 1991. The role of knowledge in discourse comprehension: A construction-integration model. In *Advances in psychology*, volume 79, pages 107–153. Elsevier.

Walter Kintsch and Teun A Van Dijk. 1978. Toward a model of text comprehension and production. *Psychological review*, 85(5):363.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44:978–990.

Robert F Lorch Jr and Paul van den Broek. 1997. Understanding reading comprehension: Current and future contributions of cognitive science. *Contemporary educational psychology*, 22(2):213–246.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. 2002. Getting access to what goes on in people's heads? reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction*, pages 101–110.

Ethel Ong, Jerwin Damay, Gerard Lojico, Kimberly Lu, and Dex Tarantan. 2007. Simplifying text in medical literature. *Journal of Research in Science, Computing and Engineering*, 4(1):37–47.

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

Gustavo Paetzold and Lucia Specia. 2016. Anita: An intelligent text adaptation tool. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 79–83, Osaka, Japan. The COLING 2016 Organizing Committee.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Cecile L. Paris. 1988. Tailoring object descriptions to a user's level of expertise. *Computational Linguistics*, 14(3):64–78.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, New York, NY, USA. Association for Computing Machinery.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137, Online. Association for Computational Linguistics.

Paul Van den Broek. 2010. Using texts in science education: Cognitive processes and knowledge representation. *Science*, 328(5977):453–456.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. volume 4, pages 401–415.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

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

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

A User study for understanding difficulty with definitions

As a preliminary study of reading difficulty, we asked annotators to read 900 concept definitions from WIKIDOMAINS and describe difficulties they have in understanding the definition text.

As shown in Figure 4, we asked each participant the following questions 1) "Please tell us the difficulties that you face in understanding the concept C from the definition," 2) "If you could ask a tutor to make changes to the definition to increase the knowledge and clarity of the concept for you or someone else, what would you ask them to change (add/edit/remove)." The first question attempts to understand the difficulties that lay people may face with domain specific definitions. The second question attempts to involve users in the thinking process of asking a tutor to rewrite the definition. Prior studies in human-centered research suggest that involving users in the task elicits better task-specific challenges than simply asking about the difficulties alone (Nielsen et al., 2002). We keep the task open-ended and ask for free-text responses to give annotators freedom to express any challenges in reading the material.

Following the completion of the study, two of the authors reviewed a random subset of 900 responses (450 responses to Q1 and 450 responses to Q2). They agreed that when annotators had issues with the reading material, it could generally fit into categories below:

- Didn't understand a mentioned concept: The annotator referenced a specific word or phrase that was mentioned in the text that hindered their understanding

- Want *more* details or other context: The annotator referenced missing details or additional background context that would have helped their understanding

- Want *less* detail: the annotator said that the definition text included unnecessary detail

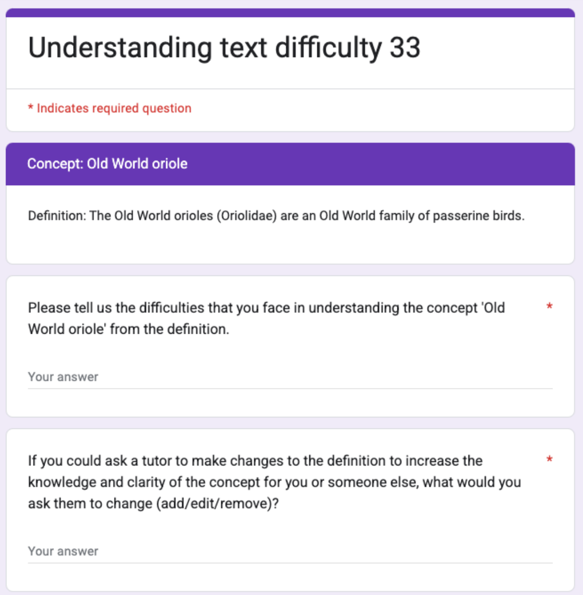
- General writing complexity: the annotator referenced the overall reading level, syntactic, or lexical complexity of the text

- Want an example or analogy: the annotator said that an example or analogy would be needed for their understanding

- Want visual or audio reference: that annotator said that they needed visual or auditory supplements for understanding the text

- Text was ill-formed: the annotator said the text was ill-formed in some way that made it difficult to read

- General confusion: annotator expressed general confusion without listing a specific pain point



Understanding text difficulty 33

* Indicates required question

Concept: Old World oriole

Definition: The Old World orioles (Oriolidae) are an Old World family of passerine birds.

Please tell us the difficulties that you face in understanding the concept 'Old World oriole' from the definition. *

Your answer

If you could ask a tutor to make changes to the definition to increase the knowledge and clarity of the concept for you or someone else, what would you ask them to change (add/edit/remove)? *

Your answer

Figure 4: Screenshot of an annotation example for understanding difficulties that readers face with domain specific text.

For the responses where an annotator identified a difficulty (which is 57% of the responses), each was labelled with one or more of the categories above (i.e., categories are not mutually exclusive). The results of this grouping is summarized in Figure 2.

A.1 User demographics for evaluations of difficulty with definitions

Table 8 summarizes the demographics of participants who evaluated the difficulty with domain-specific definitions.

Background	Percentage
4-year college degree	53%
Master's degree	12%
2-year college degree	18%
Some college	10%
High school	3%
Professional degree (MD, JD, etc)	2%
Doctoral degree (PhD)	2%

Table 8: Educational background of annotators for human evaluations of difficulty with definitions. Total number of annotators was 28.

B WikiDomains dataset construction

Editors on Wikipedia have voluntarily come together to form focus groups, called WikiProjects, dedicated to curating and improving articles in specific domains or interest areas, such as Economics, Chemistry, Literature (Asthana and Halfaker, 2018). Any Wikipedia editor can join different WikiProjects and participate in editing articles in that specific WikiProject. As part of the WikiProject effort, Wikipedia editors have annotated a large number of articles on Wikipedia with their WikiProject topic assignments and developed a hierarchical taxonomy of topics called the Wikiprojects directory[9]. The dataset contains articles from the entire Wikipedia annotated by domains (broad academic topics) derived from Wikiprojects. We use the topics in the first two levels of this categorization as domains because they represent broad domain categorizations.

B.1 Domain selection criteria

We selected domains where majority of articles related to names of academic concepts or processes in the domain. Thus, we needed to exclude articles about people, events, names of things (e.g., music albums). For example, the Biography domain contains biographical articles of famous personalities, and the Military domain contains articles on historical military conflicts. To identify such domains, the lead author manually assessed a random sample of 100 articles in each domain. If the number of articles in each domain that corresponded to named entities exceeded 50% of the assessed articles, we dropped that domain. This is because our work is focused on academically challenging concepts and how they are explained in terms of other concepts. While articles of named entities may contain challenging concepts, the concepts themselves and their explanations in the domain is not the focus of the article. E.g., an article on World War II will likely contain concepts like "diplomacy", but its explanation will not be the main focus of the article. We finally excluded the domains: Internet-culture, Literature, Religion, History, Geography, Military-and-warfare, Transportation, Society, Sports, Libraries and Information, Space, and STEM.STEM* (because this is a superset of the domains: Physics, Chemistry, Mathematics, Biology).

B.2 Dataset snapshot

Table 9 shows a snapshot of the WikiDomains dataset.

[9] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Directory

Term	Topic	Wikipedia lead section	Definition	Difficult concept
Electron gun	Physics	An electron gun...by the number of electrodes.	An electron gun (also called electron emitter) is an electrical component in some vacuum tubes that produces a narrow, collimated electron beam that has a precise kinetic energy.	electron
Vala (programming language)	Comp.	Vala is an ... in May 2006.	Vala is an object-oriented programming language with a self-hosting compiler that generates C code and uses the GObject system.	compiler

Table 9: Snapshot of the WIKIDOMAINS dataset

B.3 Difficult concept statistics

In roughly 85% of WIKIDOMAINS examples, we extracted the *difficult concept* using a ratio of how often the concept appears within this domain compared to Wikipedia overall (Equation 1). For those examples, the average computed ratio for the selected *difficult concept* is 0.9. In the remaining 15% of examples, the *difficult concept* was chosen using the age of acquisition lexicon (Kuperman et al., 2012). On average, each *difficult concept* contains 1.3 tokens.

C Prompts

We experimented with 4 candidate prompts for both $explain$ and $simplify$ prompts categories. Table 10 outlines these candidate prompts. To identify the best prompt, we applied the prompts to a set of 100 randomly sampled definitions from the WIKIDOMAINS and SCIDEF datasets, and the lead author manually assessed the goodness of the rewrites, assigning a binary label 1 or 0 to each of the rewrites, indicating whether the rewrite successfully addresses the concept difficult or not respectively. The prompts that we use in our experimental setup had the highest number of definitions where the rewrite was assessed as a good rewrite.

Table 11 details the “simplify” and “explain” prompts that we used in our study.

D Instructions for human evaluation

To evaluate LLM-generated definitions for their suitability for simplifying domain specific concepts, we show a human rater the following 1) the original definition, 2) a difficult concept c^d within the definition that we identified, 3) the LLM-rewritten definitions. We ask raters to answer the following 1) Please rate on a scale of 1-5 how much the REWRITE preserves the meaning of the original, 2) Can someone understand the definition if they do not know the difficult concept: X? (Yes/No), 3) Please rate which of the ORIGINAL and REWRITE are easier to understand? (Original/Rewrite/Both), 4) Please rate your level of familiarity with the concept.

Rationale for human evaluation questions We cannot control readers’ familiarity with the concept, therefore we rely on their understanding to determine someone’s ability to understand the definition without knowledge of the *difficult concept*. How much is a definition understandable to someone is dependent on their background knowledge. Therefore, by asking whether the REWRITE is easier to understand or the ORIGINAL definition, we rely on the annotator’s opinion of whether the rewritten definition gives them a better understanding of the topic.

Each annotator was presented with about 15 definitions to answer questions about, and the total annotation time per annotator was about 20-25 minutes. Before the annotation, we briefed the annotators about task and provided two examples to help them understand the task of concept simplification. Figure 5 shows screenshot of the annotation task.

We displayed a consent form to the participants detailing the study and that the risks would be no more than assessing definitions written by AI and gave them the option to leave the study at any time. We compensated the participants above the hourly minimum wage based on their demographic location. The study was approved by the internal ethics review team.

We collected the educational background of an-

Prompt Strategy	Prompt text
simplify	Rewrite the definition simplifying the concept: “cerebellum.”
simplify	Rewrite the definition making the concept simpler: “cerebellum.”
simplify	Rewrite the definition making the concept simpler: “cerebellum.”
simplify	Rewrite the definition simplifying difficulty with the concept: “cerebellum.”
explain	Rewrite the definition integrating an explanation for the concept: “cerebellum.”
explain	Rewrite the definition adding an explanation for the concept: “cerebellum.”
explain	Rewrite the definition providing an explanation of the concept: “cerebellum.”
explain	Rewrite the definition to add content that explains the concept: “cerebellum.”

Table 10: Candidate prompts that we explored

Prompt Strategy	Prompt text
simplify	Rewrite the definition simplifying the concept: “cerebellum”. Definition: Chiari malformations (CMs) are structural defects in the cerebellum. Rewrite:
explain	Rewrite the definition integrating an explanation for the concept: “cerebellum”. Definition: Chiari malformations (CMs) are structural defects in the cerebellum. Rewrite:

Table 11: Prompts used in the experimental evaluation

notators, summarized in Table 12.

Background	Percentage
4-year college degree	48%
Master’s degree	16%
2-year college degree	14%
Some college	13%
High school	3%
Professional degree (MD, JD, etc)	2%
Doctoral degree (PhD)	0.8%

Table 12: Educational background of annotators for human evaluations of LLM-rewrites. Total number of annotators was 229.

E Inference setting

For open-source models, we run inference on GPUs using the Huggingface[10] transformers implementation. To fit Falcon and Bloom models on the available GPUs, we run the models with 8-bit quantization (Dettmers et al., 2022). For the commercial models, we use the publicly available APIs to query the models and generate outputs. For all LLMs, we use top-k sampling[11].

F Results of Zero-shot Prompting

See Table 13 for the zero-shot performance results. As expected, scores are generally lower with zero-shot than few-shot. In particular, ICL examples seem to help with the meaning preservation dimension pretty consistently.

G Human evaluation agreement

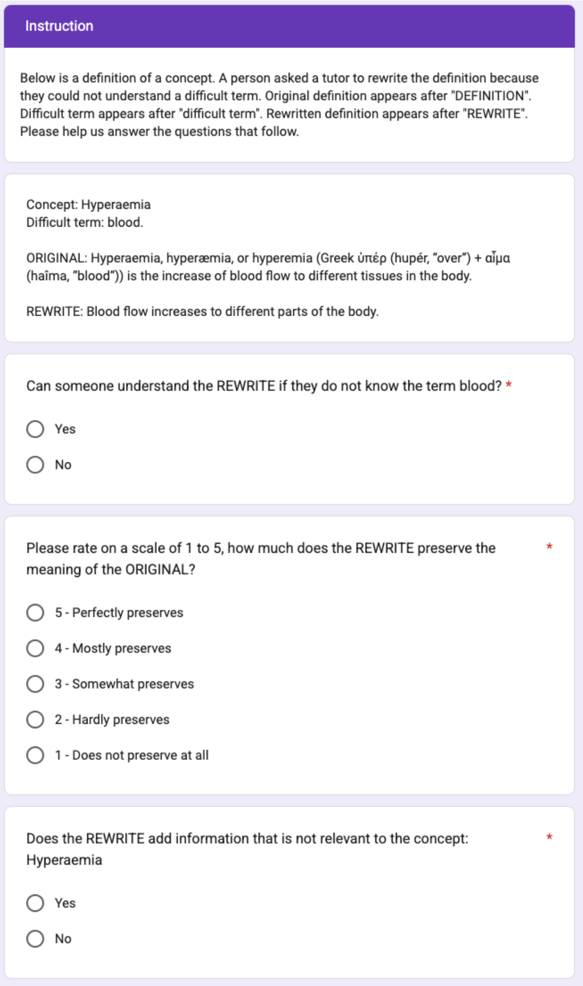
Table 14 shows the human evaluation agreement scores for our study. The inter-annotator alpha scores show weak agreement (in the range between 0.2-0.3), which is somewhat expected due to the subjective nature of some evaluations. In aggregating scores we use the majority vote between the three annotators (or the mean in the case of $\mathcal{H}_{\text{MP}}$). The Krippendorff’s alpha between individual ratings and the majority vote falls in the range of 0.6-0.7, showing that individual ratings are generally closely aligned with the majority rating.

H Automatic Metric Performances

In Table 15, we present the model performances on different automatic metrics.

[10] huggingface.co

[11] We set the value of k to 40



Instruction

Below is a definition of a concept. A person asked a tutor to rewrite the definition because they could not understand a difficult term. Original definition appears after "DEFINITION". Difficult term appears after "difficult term". Rewritten definition appears after "REWRITE". Please help us answer the questions that follow.

Concept: Hyperaemia
Difficult term: blood.

ORIGINAL: Hyperaemia, hyperæmia, or hyperemia (Greek ὑπέρ (hupér, “over”) + αἷμα (haîma, “blood”)) is the increase of blood flow to different tissues in the body.

REWRITE: Blood flow increases to different parts of the body.

Can someone understand the REWRITE if they do not know the term blood? *

○ Yes
○ No

Please rate on a scale of 1 to 5, how much does the REWRITE preserve the meaning of the ORIGINAL? *

○ 5 - Perfectly preserves
○ 4 - Mostly preserves
○ 3 - Somewhat preserves
○ 2 - Hardly preserves
○ 1 - Does not preserve at all

Does the REWRITE add information that is not relevant to the concept: Hyperaemia *

○ Yes
○ No

Figure 5: Annotation example for evaluating LLM rewritten definitions for concept simplification.

		LLM	$\mathcal{H}_{\text{MP}}$	$\mathcal{H}_{\text{RE}}$	$\mathcal{H}_{\text{RU}}$
WIKIDOMAINS	simplify	Bloomz	4.12	0.23	0.48
		Falcon	3.88	0.53	0.78
		PaLM2	4.27	0.08	0.60
		GPT4	4.02	0.67	0.70
	explain	Bloomz	4.23	0.18	0.41
		Falcon	4.00	0.47	0.65
		PaLM2	4.14	0.26	0.63
		GPT4	4.11	0.72	0.89
SCIDEF	simplify	Bloomz	4.01	0.17	0.56
		Falcon	3.87	0.49	0.56
		PaLM2	4.21	0.05	0.68
		GPT4	4.36	0.47	0.93
	explain	Bloomz	4.73	0.03	0.38
		Falcon	3.72	0.47	0.63
		PaLM2	4.53	0.37	0.64
		GPT4	4.34	0.53	0.83

Table 13: Human evaluations of LLM-generated rewrites for targeted concept simplification with zero-shot setting

	Krippendorff’s Alpha	
Metric	IAA	Ann vs. Majority
$\mathcal{H}_{\text{MP}}$	0.31	0.70
$\mathcal{H}_{\text{RU}}$	0.21	0.65
$\mathcal{H}_{\text{RE}}$	0.25	0.62

Table 14: Krippendorff’s alpha scores for the human evaluations of meaning preservation ($\mathcal{H}_{\text{MP}}$, an interval score out of 5), rewrite understanding ($\mathcal{H}_{\text{RU}}$, binary score), and rewrite easier ($\mathcal{H}_{\text{RE}}$, binary score). We report coefficients between pairs of annotators (IAA = inter-annotator agreement) and also the agreement between individual annotation and the majority vote label for that example (Ann vs Majority).

I Qualitative examples

Table 16 discusses several example outputs from the *simplify* prompt and identifies potential problems in some of the generated definitions.

J Qualitative examples from the *simplify* vs. *explain* prompts

Table 17 shows side-by-side examples of the rewritten definitions when the same model is prompted with the *simplify* prompt versus the *explain* prompt.

	LLM	Density	BLEU-4	BertSc	ΔLen	ΔAoA	ΔFlesch
WIKIDOMAINS	Bloomz	11.78	0.53	0.89	1.61	-0.24	7.30
	Falcon	6.62	0.27	0.80	12.72	4.44	**9.10**
	PaLM2	**19.57**	**0.74**	**0.92**	4.06	2.28	1.72
	GPT4	4.62	0.24	0.83	**20.28**	**5.77**	8.50
SCIDEF	Bloomz	12.81	0.62	0.92	2.35	-0.02	5.16
	Falcon	3.67	0.20	0.76	9.92	**4.55**	4.51
	PaLM2	**17.67**	**0.80**	**0.95**	2.71	0.16	2.64
	GPT4	4.24	0.24	0.83	**17.85**	3.45	**7.41**

Table 15: Automated evaluations of concept simplifications on the full test set of WIKIDOMAINS and SCIDEF

Notes	Original definition	Rewritten definition
PaLM2, Difficult concept: zenith	Tropic Of Capricorn is the southernmost latitude of latitude on the Earth at which the Sun appears zenith overhead at its culmination, which lies approximately 23.4 degrees south of the Equator.	Tropic Of Capricorn is the southernmost latitude of latitude on the Earth at which the Sun appears directly overhead at its culmination, which lies approximately 23.4 degrees south of the Equator.
GPT4, Difficult concept: zenith Note: Correctly simplifies difficult concept but changes definition significantly.	Tropic Of Capricorn is the southernmost latitude of latitude on the Earth at which the Sun appears zenith overhead at its culmination, which lies approximately 23.4 degrees south of the Equator.	The Tropic of Capricorn is the furthest point south on Earth where the Sun can be seen directly overhead at its highest point. This is roughly 23.4 degrees south of the Equator.
PaLM2, Difficult concept: vegetation	Jungle is an area covered with dense vegetation dominated by large trees, often tropical.	Jungle is an area covered with dense plants dominated by large trees, often tropical.
GPT4, Difficult concept: vegetation Note: Correctly simplifies difficult concept but changes definition significantly.	Jungle is an area covered with dense vegetation dominated by large trees, often tropical.	A jungle is a place filled with a lot of plants and big trees, usually found in hot regions.
Bloomz, Difficult concept: linguistics Note: Makes no change.	Prolog is is a logic programming language associated with artificial intelligence and computational linguistics.	Prolog is a logic programming language associated with artificial intelligence and computational linguistics.
GPT4, Difficult concept: linguistics Note: Correctly simplifies difficult concept but changes definition significantly.	Prolog is is a logic programming language associated with artificial intelligence and computational linguistics.	Prolog is a computer language used in artificial intelligence and the study of how computers can understand and use human language.
PaLM2, Difficult concept: ovoid Note: model drops information.	A coccus (plural cocci) is any bacterium that has a spherical, ovoid, or generally round shape.	A coccus (plural cocci) is any bacterium that has a spherical, or generally round shape.
GPT4, Difficult concept: ovoid	A coccus (plural cocci) is any bacterium that has a spherical, ovoid, or generally round shape.	A coccus (or cocci for many) is a type of bacterium that is shaped like a sphere, an egg, or is generally round.

Table 16: LLM rewrite examples for *simplify* prompts

model	original definition	*simplify* definition	*explain* definition
Falcon (0-shot)	Jungle is an area covered with dense **vegetation** dominated by large trees, often tropical.	The jungle is a lush and verdant forest characterized by towering trees and dense undergrowth.	The jungle is a vast expanse of land that is covered in dense vegetation, primarily consisting of tall trees and other plants. It is characterized by its humid climate and high levels of rainfall, which support the growth of a wide variety of flora and fauna.
PaLM2 (3-shot)	Chiari malformations (CMs) are structural defects in the **cerebellum**.	Chiari malformations (CMs) are structural defects in the brain.	Chiari malformations (CMs) are structural defects in the cerebellum. The cerebellum is a region of the brain that plays an important role in motor control.
GPT4 (3-shot)	Hamlet is a small human settlement, variably defined as one the size of a town, village, or **parish** or as a smaller subdivision of or satellite entity to a larger settlement.	Hamlet is a small human settlement, variably defined as one the size of a town, village, or church district or as a smaller subdivision of or satellite entity to a larger settlement.	Hamlet is a small human settlement, variably defined as one the size of a town, village, or parish or as a smaller subdivision of or satellite entity to a larger settlement. A parish is a type of administrative division used by some churches and by some civil authorities.
Bloomz (3-shot)	Arginase deficiency is an inherited disorder that causes the amino acid arginine (a building block of proteins) and **ammonia** to accumulate gradually in the blood.	Arginase deficiency is a rare disease in which the body is unable to process certain proteins properly.	Arginase deficiency is an inherited disorder that causes the amino acid arginine (a building block of proteins) and ammonia to accumulate gradually in the blood. Ammonia is the toxic byproduct of protein metabolism.

Table 17: Definitions rewritten with *simplify* vs. *explain* prompts. The difficult concept is in bold.