Conducting a Large-Scale User Study on Scientific Text Simplification

Kyuri Im
kyuri.im@mailbox.tu-dresden.de

August 2025

Abstract

As interdisciplinary collaboration drives scientific progress, researchers increasingly face challenges in understanding literature outside their domains. This is particularly evident in computer science, where technical jargon and complex writing hinder accessibility for non-experts. This study investigates LLM-based methods for scientific text simplification to address this barrier. Using SciSummNet as the base corpus, summaries are simplified with GPT-4o-mini, followed by clarity and readability feedback collected from non-expert participants. Informed by this feedback, domain experts produce gold-standard simplified summaries. The resulting dataset offers a valuable resource for developing AI systems that enhance cross-disciplinary understanding in STEM research.
Research dataset is available here :`https://github.com/kim33/Scientific-Text-Simplification.git`

1 Introduction

The rapid growth of interdisciplinary research has transformed the landscape of modern science, enabling breakthroughs at the intersection of fields such as biology, physics, engineering, and computer science. [2, 4]However, as collaboration across domains becomes more common, a significant barrier persists: the accessibility of scientific literature. Academic papers are often written in highly technical language, making them difficult to understand for researchers outside the originating discipline. [1] This is especially true in fields like computer science, where dense terminology and abstract concepts can limit engagement from non-expert readers.

Efforts to address this challenge have led to increased interest in automated text simplification, particularly using large language models (LLMs). These models have demonstrated the potential to generate readable and faithful simplifications of complex texts. [12] Yet, there remains a lack of high-quality datasets and evaluation frameworks tailored specifically to the needs of interdisciplinary audiences.

This study aims to develop and evaluate LLM-based methods for scientific text simplification, with a focus on making academic writing more accessible to non-experts across STEM fields. Central to our approach is a large-scale user study designed to assess how well simplified texts support understanding for readers without a background in the target domain. The insights and data gathered from this study will be used to create a curated, gold-standard dataset that can serve as both a benchmark and a training resource for future research in AI-driven text simplification.

2 Related Work

2.1 LLMs and Scientific Text Simplification

More recently, models such as GPT-3, GPT-4, and ChatGPT have shown state-of-the-art performance in zero-shot and few-shot simplification of highly technical content, including scientific abstracts and research articles. [3, 17] Recent studies show that LLM can simplify academic texts for non-experts with moderate success, although hallucination and loss of detail remain concerns. [5] Liu et al. (2025) conducted a large-scale study evaluating LLM-generated simplifications across 31 scientific texts, showing measurable gains in reader comprehension for non-specialists. [7] Despite promising results, most existing benchmarks focus on general news or Wikipedia content. Few resources exist for high-quality, annotated datasets targeting scientific or interdisciplinary readers, which this project aims to address.

2.2 Human-Centered Evaluation in NLP

As NLP systems increasingly shape public communication and education, human-centered evaluation has become essential alongside automated metrics. Traditional tools like BLEU [11], ROUGE [9], and FK-Grade [6] assess surface-level text features but often overlook actual user comprehension, trust, and utility—especially in technical domains like computer science.

Recent work emphasizes the need for multi-dimensional human evaluation in text simplification. Sulem et al. [15] highlight the importance of assessing meaning preservation and fluency beyond grammatical correctness. Similarly, Liu et al. [7] show that LLM-generated simplifications of scientific texts can improve non-expert comprehension.

In interdisciplinary communication, user-centered evaluation is especially valuable. Readers often lack domain knowledge, making it critical to assess whether simplified texts enhance their confidence and understanding. This has driven increased use of user studies, surveys, and task-based evaluations in NLP—particularly for educational and accessibility-focused applications.

Accordingly, this study conducts a large-scale user study with non-experts, gathering structured feedback on clarity, usefulness, and comprehension. This informs the creation of a high-quality dataset that reflects both algorithmic needs and real-world accessibility.

3 Methodology

3.1 Dataset

For text simplification and evaluation, this study adopts SciSummNet [20], a benchmark dataset for scientific summarization. It comprises 1,000 highly cited ACL papers, each paired with its abstract, citation contexts from other papers, and a 150-word expert-written gold summary. While SciSummNet captures the core content of scientific papers, its summaries often retain technical terminology, limiting accessibility for non-experts. Prior work suggests that summarization alone is insufficient, as it can preserve complex terms, whereas simplification without focus may lead to verbosity. [21] Hence, further simplification is crucial for generating concise and accessible scientific content for higher comprehension in interdisciplinary domains.

3.2 Baseline Simplification Generation

The study first takes a set of gold summaries from SciSummNet and generates simplified scientific summaries using GPT-4o-mini. These initial simplifications serve as baseline texts for subsequent evaluation.

3.2.1 Baseline Evaluation

To assess the necessity and effectiveness of text simplification for scientific paper summaries in the computer science domain, we evaluated both the original and GPT-simplified summaries using a combination of automatic readability and simplification metrics. The evaluation includes the Flesch-Kincaid Grade Level, Flesch Reading Ease[6], SARI[19], and LENS-SALSA[10] scores. Since reference texts from human annotators were not available, the SARI score is computed in a reference-less mode, comparing the original summaries (as source) to the GPT-simplified versions (as system outputs) only.

The results are as follows:

- **Flesch-Kincaid Grade Level: 12.40** The simplified summaries still require at least a high school or early college-level reading ability, indicating that the simplification process did not significantly lower the complexity to general public readability levels.
- **Flesch Reading Ease: 47.83** This score corresponds to "difficult" text, suggesting that while simplification had some impact, the resulting summaries remain challenging for non-expert readers.
- **Average SARI Score: 50.04** The SARI score—designed specifically for evaluating text simplification—suggests moderate effectiveness. A score around 50 implies that the system makes some beneficial edits, such as appropriate additions, deletions, or retentions, but still has room for improvement.
- **LENS-SALSA Score (Average): 54.46** This score reflects semantic alignment between simplified and original text while accounting for linguistic simplicity. A mid-range score like this indicates

a reasonable balance between preserving meaning and reducing complexity, but also highlights potential gains from more targeted simplification.

Overall, the automatic evaluation suggests that while the GPT-simplified summaries exhibit minor improvements in readability, the text still remains complex and academic in tone. The relatively high FK grade level underscores that these simplifications are often subtle and do not fully bridge the gap between technical and general-audience comprehension. These findings motivate the need for further simplification—either through model fine-tuning or human-in-the-loop revision—to produce summaries that are genuinely more accessible, especially for non-experts or interdisciplinary readers. The moderate SARI and LENS scores also reinforce the opportunity for enhanced simplification strategies that maintain fidelity while increasing clarity.

3.3 User Study

3.3.1 Phase 1: Identification of Complex Language and Evaluation of GPT-Simplified Summaries

Phase 1 is designed to collect user feedback on the complexity of language used in scientific summaries. As described in Figure 1, it involves comparing the original and GPT-simplified versions of summaries and asking participants—non-experts in computer science but from other STEM fields and fluent in English—to identify complex words or phrases that require further simplification for improved understanding. Participants are presented with both the original and the AI-simplified summaries and are asked to click on sentences they find difficult or in need of simplification. To mitigate potential bias and prevent participants from inferring which version is original or simplified, the presentation order of the two summaries is randomized for each survey instance. Following this, participants are asked to evaluate the simplified summary based on three key dimensions commonly used in the text simplification and natural language processing (NLP) literature:

1. **Simplicity** Measures the extent to which the language has been made more accessible, particularly through reductions in syntactic and lexical complexity [14, 16].

2. **Information Coherence** – Assesses whether the logical flow and conceptual integrity of the original content are preserved in the simplified version, which is especially important for technical and scientific texts [18, 8].

3. **English Fluency** – Evaluates the grammatical correctness, naturalness, and stylistic quality of the simplified summary to ensure it remains readable and appropriate for a general audience [13].

These three dimensions collectively form a robust and user-centered framework for evaluating the quality of scientific text simplification. In the final part of the survey, participants are asked to highlight complex words or phrases from the sentences they previously flagged, using an annotation interface. Figure 1 illustrates the evaluation criteria and the annotation interface provided during the study.

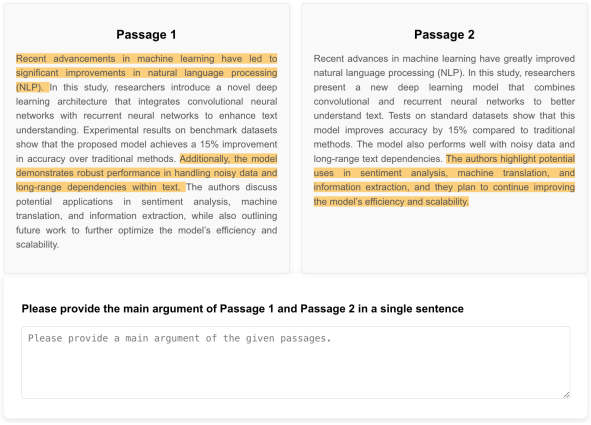
3.3.2 Phase 2: Generation of Gold-Standard Simplified Summaries

Building on the annotations collected in Phase 1, Phase 2 focuses on producing gold-standard simplified summaries of scientific texts in the field of computer science. Participants in this phase are either currently enrolled in or have completed a Master's degree or higher in computer science, and are fluent in English. As described in Figure 2 each participant is presented with the original summary and its corresponding GPT-generated simplified version. Words or phrases identified as complex during Phase 1 are highlighted using brackets. Participants are asked to generate high-quality simplified summaries based on these inputs, following specific guidelines depending on the annotation scenario. There are four distinct cases:

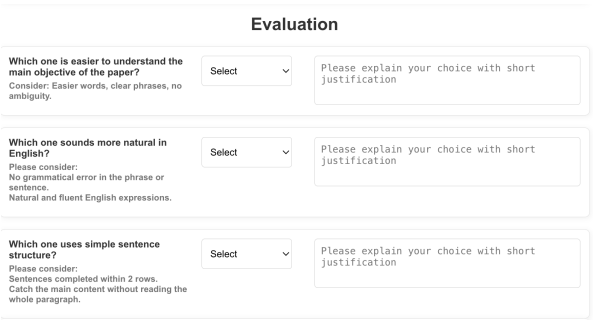
Case 1. Both summaries contain annotated complex phrases.

Participants are instructed to examine the bracketed phrases in both the original and GPT-simplified summaries. They should assess whether the simplified version accurately and clearly rephrases the complex terms while preserving the original meaning and grammatical correctness. The final gold-standard summary should further improve clarity, ensuring that all bracketed content is effectively simplified.

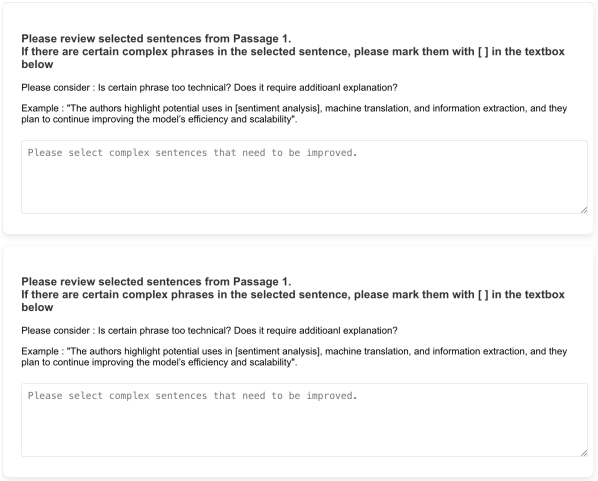
Case 2. Only the original summary contains annotated complex phrases.



(a) Select Complex Sentences



(b) Evaluation Question



(c) Annotation Task

Figure 1: Survey interface used in Phase 1

Note: Participants rated each summary on simplicity, coherence, and fluency.To reduce bias, the positions of the original and simplified summaries were randomized for each participant. This ensured users could not easily infer which version was machine-generated.

In this case, participants should verify that the complex terms identified in the original summary are appropriately simplified in the GPT version. If necessary, they must revise the simplified version to ensure it accurately reflects the original content while enhancing clarity and accessibility.

Case 3. Only the GPT-simplified summary contains annotated complex phrases.

Participants should rephrase the bracketed content in the GPT-simplified summary to ensure that it clearly and accurately conveys the intended meaning of the original summary.

Case 4. Neither summary contains annotations.

This scenario indicates that no complex terms were flagged in Phase 1. Participants are asked to carefully review the GPT-simplified summary to confirm that it accurately captures the core meaning of the original text. They should revise any part of the simplified summary as needed to improve readability, coherence, and flow.

Through this structured annotation and rewriting process, Phase 2 aims to create a high-quality benchmark dataset for evaluating scientific text simplification systems.

4 Evaluation

4.1 Analysis of Phase 1: Assessing GPT-Simplified Scientific Summaries

To assess the effectiveness of GPT-based simplification for scientific summaries, we analyzed participants' feedback across three key dimensions: **simplicity**, **understanding**, and **English fluency**. These

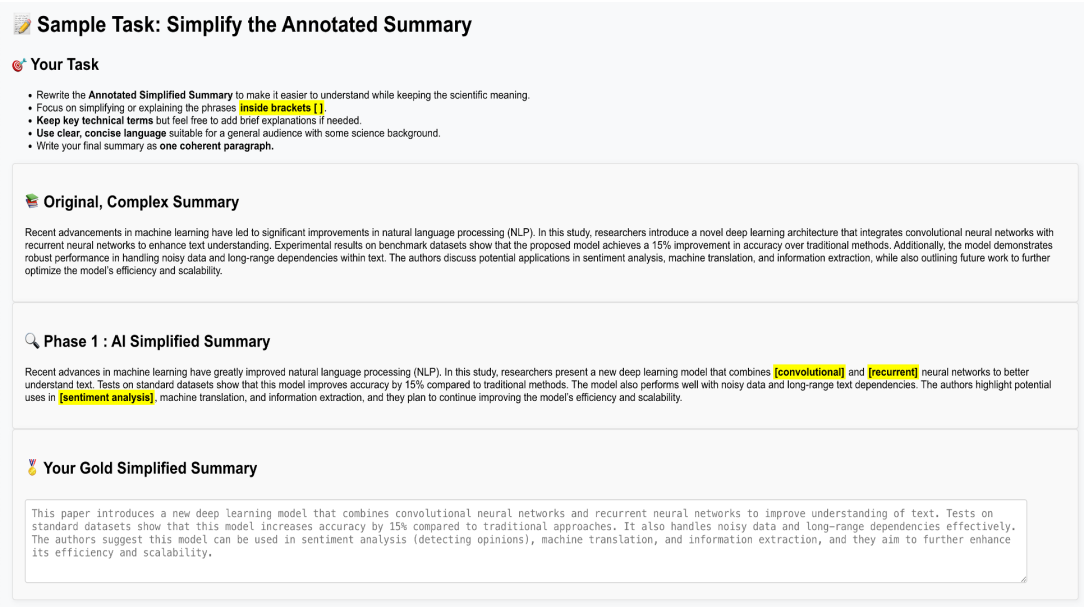


Figure 2: Survey interface used in Phase 2

Note: Each participant for phase 2 is presented with the original summary and GPT-simplified summary with annotations from phase 1. Then they are asked to generate gold simplified summary of the scientific paper.

criteria are commonly used in the text simplification and NLP literature to evaluate both surface-level readability and deeper semantic fidelity. Rather than relying on automatic metrics alone, this evaluation drew on **direct human judgments** from participants with STEM backgrounds (outside of computer science), who engaged with pairs of original and simplified summaries.

As described in Figure 3, the evaluation results indicate that participants generally perceived the GPT-simplified summaries as clearer and more understandable than the original versions. In particular, for the Understanding and Simplicity dimensions, the GPT-simplified summaries received overwhelmingly higher ratings, with over 70 out of 92 participants selecting the simplified version. This supports the notion that even automated simplification tools like GPT can significantly enhance accessibility in technical writing for a broader academic audience. In the Naturalness dimension, results were more balanced. A substantial number of participants (42 out of 92) reported no perceptible difference between the two summaries, suggesting that GPT-simplified texts preserved fluency reasonably well. However, a non-negligible number of users still preferred the original version's naturalness (14 responses), highlighting occasional awkwardness or stylistic artifacts in the AI-generated output.

Together, these findings support the utility of automated simplification in enhancing accessibility while also revealing areas—particularly in stylistic fluency—where further refinement may be needed. The interactive feedback from participants, especially the highlighted complex words and phrases, provides actionable insights for improving simplification strategies in subsequent phases.

While quantitative evaluation from Phase 1 indicated that GPT-simplified summaries were generally preferred—particularly for simplicity (70 votes) and information coherence (73 votes)—user comments reveal nuanced insights that highlight the limits of automated simplification. As shown in Figure 4, a participant noted that the original summary sounded more natural, emphasizing that its word choices such as "arbitrary threshold," "notion of correctness," and "sentence-level QE system" adhered more closely to conventional academic English. In contrast, the GPT-simplified version included phrases like "chosen limit" and "idea of correctness is not easy to understand," which, while potentially more accessible, were described as slightly informal or awkward in a scholarly context. Another participant remarked that the original passage conveyed ideas in a more structured and digestible manner due to shorter sentences and fewer stacked phrases—suggesting that syntactic simplicity is not always achieved through lexical substitution alone.

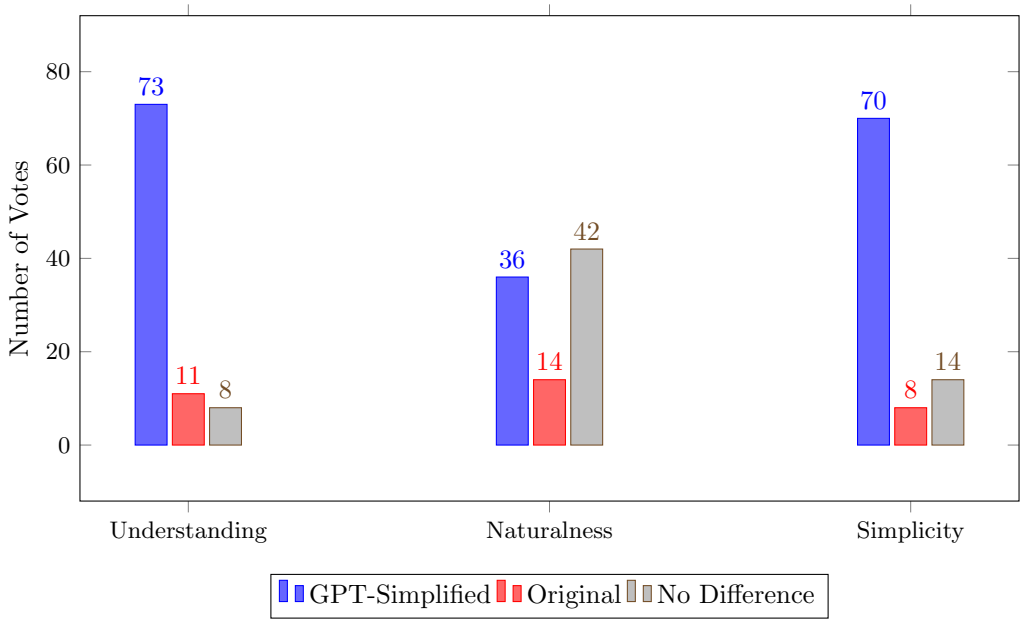


Figure 3: Participant preferences across dimensions in Phase 1 evaluation

Naturalness: "Passage 1 sounds a little more natural overall.
Its word choices ("arbitrary threshold," "notion of correctness," "sentence-level QE system") match standard academic English and read smoothly, whereas some substitutions in Passage 2—such as "idea of correctness is not easy to understand" and "chosen limit"—feel slightly informal or awkward. Both passages are grammatically correct, but Passage 1's vocabulary and phrasing **align better with conventional, fluent scholarly style**.."

Simplicity : Passage 1—its sentences are a little shorter and have **fewer parenthetical insertions or stacked phrases**, so each idea is delivered in a cleaner, more easily digestible structure.

"Original": "We present a detailed study of confidence estimation for machine translation. Various methods for determining whether [MT] output is correct are investigated, for both [whole sentences] and [words]. Since the [notion of correctness] is not intuitively clear in this context, different ways of defining it are proposed. We introduce a [sentence-level QE system] where an **[arbitrary threshold]** is used to classify the [MT output] as good or bad. Since the **notion of correctness** is not intuitively clear in this context, different ways of defining it are proposed. We present results on data from the NIST 2003 Chinese-to-English MT evaluation. We introduce **a sentence level QE system** where an arbitrary threshold is used to classify the MT output as good or bad. We study sentence and word level features for translation error prediction.",

"GPT-Simplified": "We present a detailed study of confidence estimation for machine translation. Various methods for determining whether MT (Machine Translation) output is correct are investigated, for both whole sentences and words. Since the **[idea of correctness] is not easy to understand** in this context, different ways of defining it are proposed. We introduce a [sentence-level QE (Quality Estimation) system] where a chosen limit is used to classify the [MT output] as good or bad. We present results on data from the NIST 2003 Chinese-to-English MT evaluation. We introduce a sentence level QE (Quality Estimation) system where **a chosen limit** is used to classify the MT output as good or bad. We study sentence and word level features for translation error prediction."

Figure 4: Example of Participant Feedback on Language Complexity in Scientific Summaries

In addition to this, while the GPT-simplified version is designed to enhance accessibility, some participants reported that the original summary offered better conceptual clarity, especially for technically dense content. For instance, as shown in Figure 5one participant commented, "Passage 1 explains the technical words better," indicating that the original formulation provided more precise or appropriately technical descriptions. In this case, the simplified version replaced key terms such as "latent variables", "PCFG", and "NP-hard" with more general or loosely defined language (e.g., "hidden elements" or "very complex and difficult"), which may have reduced the perceived rigor or informativeness of the explanation for a scientifically literate audience. These findings underline the trade-offs inherent in automated text simplification: while GPT-based models can significantly enhance surface-level readability and accessibility for general STEM audiences, they may also introduce stylistic inconsistencies or oversimplifications that obscure critical domain-specific meaning. In several cases, participants preferred the original summaries, noting that key technical terms or explanations were clearer and more precise in the unsimplified version. This highlights the limitations of fully automated approaches, especially in disciplines like computational linguistics where conceptual precision and terminological accuracy are essential. As such, these results reinforce the necessity of human-in-the-loop evaluation and expert-guided annotation—both for assessing simplification quality and for guiding model improvement. Although tools like GPT can expedite the simplification process, they cannot fully replace the nuanced judgment and domain expertise required to produce high-quality, gold-standard simplified summaries that maintain both clarity and technical fidelity.

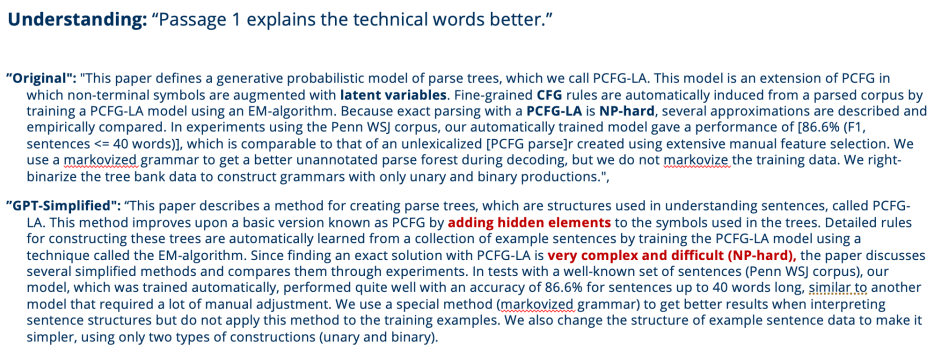


Figure 5: Example of Participant Feedback on Clarity in Scientific Summaries

4.2 Gold Summary Evaluation

To evaluate improvement in the quality of text simplification of gold simplified summaries, a comparative analysis between GPT-generated simplifications and human-annotated gold-standard simplifications is conducted using a diverse set of metrics across three categories: Semantics, Simplicity, and Readability.

4.2.1 Semantics

Semantic preservation was assessed using BLEU, LENS, and BERTScore (Precision, Recall, F1) and Paraphrase similarity. The GPT-simplified summaries achieved significantly higher scores across all semantic metrics. Specifically, the BERTScore F1 for GPT outputs reached 0.9907, surpassing the human simplifications at 0.9031, indicating a higher degree of semantic alignment with the original text. BLEU and LENS scores also favored the GPT simplifications (0.949 vs. 0.349, and 53.77 vs. 50.56, respectively), although it is worth noting that BLEU is limited in its ability to capture meaningfully different but valid paraphrases — a case where human simplifications may intentionally diverge lexically while remaining faithful in meaning.

4.2.2 Simplicity

Simplicity was measured using SARI, SAMSA, compression ratio, average word length, and average sentence length. Human simplifications showed slightly lower SARI score (33.72 vs. 46.76), suggesting that GPT performed simplification with better balance between deletion, addition, and retention of tokens. While GPT tended to retain more of the original content (as reflected by its slightly higher compression ratio), this often came at the cost of reduced structural simplification. Compared to average word length 5.38 and average sentence length 164.42 of the original corpus, human-simplification contained lower average sentence length, indicating improved structural simplification. Higher SAMSA score of GPT-simplification (0.3094 vs. 0.2530) indicates higher quality of sentence splitting but the difference in improvement is not significant.

4.2.3 Readability

Readability was assessed with Flesch–Kincaid Grade Level (FKGL) and Reading Ease (FKRE). GPT-simplified summaries scored FKGL = 12.10 (vs. 13.00 for human) and FKRE = 48.54 (vs. 40.89), indicating a modest advantage in surface accessibility. However, both versions fall in the advanced/college range—expected for technical abstracts—so the numerical gap is small and unlikely to matter for the intended expert or semi-expert audience. Crucially, FK metrics reflect sentence length and syllable counts, not terminological accuracy or scientific nuance; therefore the slightly lower FKRE for human edits should not be read as deterioration.

4.2.4 Gold Summary Example

Overall, automated simplification metrics in Table 1 tend to rate GPT outputs as "better" because they favor shorter, more frequent words. Yet for scientific texts, these scores conflate surface simplicity with quality. As indicated in Figure 6, GPT often loosens precise terms—computational lexicon → computer-based dictionary, syntactic information → grammar information—and softens claims ("not

suitable" → "not easy"). By contrast, the human gold edits use a hybrid, post-editing strategy: they preserve domain-critical terminology and key sentences from the original (e.g., computational lexicon, syntactic information, indexed by), while selectively borrowing GPT's clearer phrasing for non-technical scaffolding (e.g., "This paper details...," "The authors discuss..."). This produces texts that remain scientifically precise yet easier to read, and it is more efficient in practice because annotators reuse only the helpful GPT paraphrases instead of rewriting from scratch. Consequently, standard metrics overvalue GPT's surface readability and undervalue human edits that optimize terminological fidelity, claim strength, audience fit, and expert-in-the-loop efficiency.

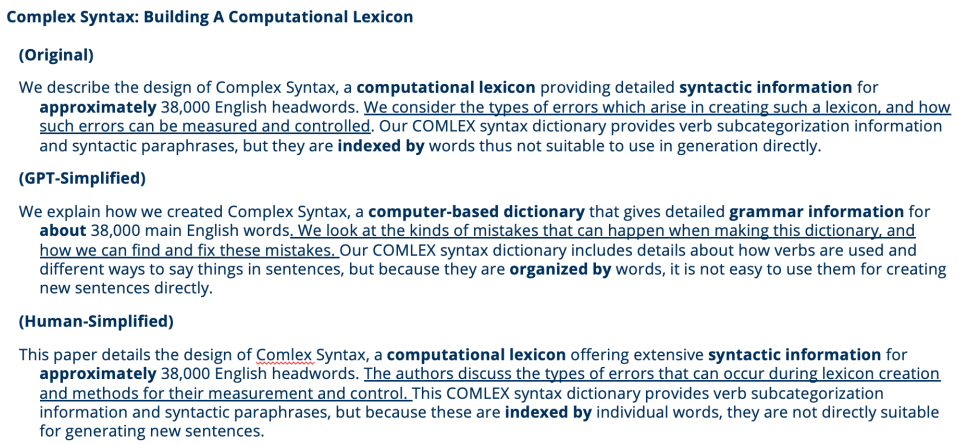


Figure 6: Example of Gold Simplification in Phase 2

4.2.5 Summary of Evaluation

Overall, automatic metrics favor GPT. It scores higher on semantic alignment and shows advantages on operation-aware simplicity metrics and FK readability. However, these measures emphasize lexical overlap, token-level edits, and sentence splitting rather than domain adequacy. Human gold edits make more consequential structural choices with shorter sentences relative to the original and preserve technical terminology and calibrated claims, while selectively borrowing clearer non-technical phrasing from GPT. For a scientific audience, the small readability gap is unlikely to be consequential, and GPT's higher scores partly reflect greater content retention rather than deeper restructuring. Taken together, expert-in-the-loop post-editing better balances precision, professionalism, and accessibility.

Category	Metric	GPT-Simplified	Gold-Simplified
Semantics	BLEU	0.949	0.349
	LENS	53.7663	50.5648
	BERTScore Precision	0.9880	0.8999
	BERTScore Recall	0.9936	0.9066
	BERTScore F1	0.9907	0.9031
	Paraphrase Similarity	0.9880	0.9046
Simplicity	SARI	46.7589	33.7167
	SAMSA	0.3094	0.2530
	Compression	1.0522	1.0286
	Avg Word Length	5.4163	5.7087
	Avg Sentence Length	172.0851	167.7447
Readability	FK Grade Level	12.10	13.00
	FK Reading Ease	48.54	40.89

Table 1: Evaluation comparison between GPT-Simplified and Human-Simplified scientific summaries.

5 Limitations and Challenges

Despite the insights gained in the evaluation, some limitations must be acknowledged:

First, the evaluation was conducted on only 47 finalized human-simplified summaries out of a total of 1,010, due to constraints in participant recruitment and follow-through. This small subset may not fully represent the diversity of the full dataset in terms of domain complexity, linguistic variation, or summarization difficulty, limiting the generalization of the findings.

Another limitation lies in the evaluation of the gold-standard simplified summaries, which could be strengthened through an additional human annotation phase. While automatic metrics such as BLEU, LENS, Flesch-Kincaid Grade Level, and Reading Ease offer a general assessment of textual simplicity, they may not fully capture the nuances of simplification in the context of scientific content. Given that the dataset originates from computer science research papers and is intended for interdisciplinary audiences, domain-specific considerations—such as maintaining technical accuracy while enhancing accessibility—are critical. Incorporating an additional evaluation stage involving human annotators from the target user group (e.g., interdisciplinary researchers) would likely yield more accurate and contextually meaningful assessments of summary quality, especially in capturing conceptual clarity and relevance.

6 Conclusion

This study explored a human-in-the-loop framework for scientific text simplification, targeting the accessibility challenges that arise in interdisciplinary research. Starting from GPT-4o-mini-generated simplified summaries of computer science papers, non-expert participants were engaged to identify remaining complexity and evaluate the clarity and fluency of the outputs. These annotations informed a second phase in which domain experts revised the GPT outputs to produce high-quality, gold-standard simplified summaries.

The findings demonstrate that while GPT-generated simplifications offer strong surface-level readability and semantic fidelity, human refinement is essential for improving structural simplicity, preserving domain-specific meaning, and enhancing overall accessibility for interdisciplinary audiences. Expert annotators, guided by real user feedback, were able to make informed decisions about which technical terms to retain or rephrase, resulting in summaries that strike a more effective balance between clarity and scientific rigor.

By combining the scalability of LLMs with targeted human intervention, this research contributes a curated dataset and methodological workflow that can support the development and evaluation of future AI systems for scientific communication. Ultimately, this work underscores the importance of collaborative, user-informed approaches in building simplification tools that are not only accurate, but also practically useful for a diverse range of readers in STEM fields.

References

[1] Suha S. Al-Thanyyan and Aqil M. Azmi. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2), March 2021.

[2] Moreno Bonaventura, Vito Latora, Vincenzo Nicosia, and Pietro Panzarasa. The advantages of interdisciplinarity in modern science, 2017.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Eoghan Cunningham, Barry Smyth, and Derek Greene. Collaboration in the time of covid: a scientometric analysis of multidisciplinary sars-cov-2 research. *Humanities and Social Sciences Communications*, 8(1), October 2021.

[5] Björn Engelmann, Fabian Haak, Christin Katharina Kreutz, Narjes Nikzad Khasmakhi, and Philipp Schaer. Text simplification of scientific texts for non-expert readers, 2023.

[6] Rudolf Flesch. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007, 2007.

[7] Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric, Cao, Yossi Matias, Avinatan Hassidim, Dale R Webster, Yun Liu, Sho Fujiwara, Peggy Bui, and Quang Duong. Llm-based text simplification and its effect on user comprehension and cognitive load, 2025.

[8] David Kauchak. 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, 2013.

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

[10] Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. Lens: A learnable evaluation metric for text simplification. *arXiv preprint arXiv:2212.09739*, 2022.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[12] Jipeng Qiang, Minjiang Huang, Yi Zhu, Yunhao Yuan, Chaowei Zhang, and Kui Yu. Redefining simplicity: Benchmarking large language models from lexical to document simplification, 2025.

[13] Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36, 2015.

[14] Advaith Siddharthan. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109, 2006.

[15] Elior Sulem, Omri Abend, and Ari Rappoport. BLEU is not suitable for the evaluation of text simplification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[16] Sowmya Vajjala and Detmar Meurers. Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, 165(2):194–222, 2014.

[17] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

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

[19] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016.

[20] Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks, 2019.

[21] Farooq Zaman, Faisal Kamiran, Matthew Shardlow, Saeed-Ul Hassan, Asim Karim, and Naif Radi Aljohani. Sats: simplification aware text summarization of scientific documents. *Frontiers in Artificial Intelligence*, Volume 7 - 2024, 2024.