A Human-in-the-Loop Corpus for LLM-Based Simplification of Scientific Summaries

Kyuri Im, Michael Färber

ScaDS.AI, Technische Universität Dresden, Dresden, Germany

Abstract

Interdisciplinary research is accelerating, yet scientific papers remain difficult to read outside their home fields. We study large language model (LLM)-based simplification for scientific texts and present a human-in-the-loop workflow that turns expert summaries into more accessible versions for non-specialists. Using SciSummNet as the source corpus, we first generate baseline simplifications with GPT-4o-mini. In Phase 1, readers from STEM domains outside computer science identify difficult sentences and phrases and compare the original and GPT-simplified summaries in terms of understanding, naturalness, and simplicity. In Phase 2, computer science experts use this feedback to produce expert-edited reference simplifications. We release the resulting corpus together with human judgments and automatic evaluation results. The Phase 1 judgments show a clear preference for the GPT outputs in understanding and simplicity, while qualitative analysis of the Phase 2 edits illustrates the importance of preserving domain terminology and scientific claim strength. The resource supports the training and benchmarking of simplification systems for cross-disciplinary scientific communication.
Dataset: https://github.com/faerber-lab/scientific-text-simplification-corpus

Keywords

text simplification, scientific communication, human-in-the-loop, evaluation, corpus

1. Introduction

Interdisciplinary research is increasingly important, as many scientific advances emerge at the intersection of different fields [1, 2]. At the same time, scientific writing is typically tailored to specialists within a particular discipline rather than to researchers from neighboring fields. Domain-specific terminology, highly compressed arguments, and field-specific writing conventions can therefore substantially limit the accessibility, interpretation, and reuse of scientific findings [3].

Large language models (LLMs) provide a promising basis for simplifying scientific and other specialized texts. Their strong zero- and few-shot capabilities have been demonstrated across a broad range of language tasks [4, 5], and recent studies have explored their use for scientific and specialized text simplification [6, 7]. However, concerns remain regarding hallucinations, loss of nuance, and inconsistent writing styles. Moreover, most existing benchmarks focus on news articles or Wikipedia passages rather than scientific discourse and cross-disciplinary audiences. Consequently, suitable datasets and evaluation protocols for scientific text simplification remain limited [8].

We address this gap through a human-in-the-loop framework that operationalizes accessibility for STEM readers without expertise in the respective source domain while preserving the scientific meaning of the original text. Using SciSummNet [9] as source material, we (i) generate initial simplifications using an LLM, (ii) collect sentence- and phrase-level difficulty annotations and assessments of understanding, naturalness, and simplicity from STEM readers outside the respective domain, and (iii) create expert-edited reference simplifications informed by this feedback. The resulting design directly addresses the central trade-off between improving readability and preserving scientific precision, which remains difficult for fully automated simplification systems.

Overall, our contributions are as follows:

- A human-in-the-loop methodology for simplifying scientific summaries that centers comprehension by cross-disciplinary STEM readers while maintaining technical accuracy.

SIG Knowledge Management Workshop (FG WM) at KI 2026, 2026, Bremen, Germany
michael.faerber@tu-dresden.de (M. Färber)

- A curated corpus[1] derived from SciSummNet with (a) baseline LLM outputs, (b) sentence- and phrase-level difficulty annotations, and (c) expert-edited reference simplifications.
- An evaluation showing where LLMs improve surface readability and where expert post-editing is crucial for claim calibration and terminology fidelity.

The remainder of the paper is structured as follows. Section 2 reviews related work on scientific text simplification and human-centered evaluation. Section 3 describes the corpus construction. Section 4 presents the results of the two study phases, followed by limitations and conclusions in Sections 5 and 6.

2. Related Work

LLMs for Scientific Text Simplification. LLMs have demonstrated strong zero- and few-shot capabilities across a broad range of language tasks [4, 5]. In a zero-shot setting, the model receives only the task instructions, whereas few-shot prompting additionally provides a small number of input-output demonstrations. Recent studies have applied these capabilities to scientific and specialized text simplification [6, 7]. Their results indicate that LLMs can improve accessibility for readers outside the source domain, but risks remain regarding hallucination, loss of nuance, and changes in claim strength. Beyond these quality concerns, most widely used benchmarks are drawn from news or Wikipedia, offering limited coverage of scientific discourse and cross-disciplinary audiences [8]. This work addresses that gap by focusing on scientific summaries and collecting difficulty annotations from cross-disciplinary STEM readers.

Human-Centered Evaluation in NLP. Automated metrics such as BLEU and ROUGE [10, 11] and readability formulas like Flesch–Kincaid [12] capture surface overlap and fluency proxies but correlate imperfectly with actual understanding – especially for technical material. The simplification literature emphasizes evaluating meaning preservation, grammaticality, and fluency beyond n-gram overlap [13]. Complementing this, user-centered assessments target simplicity and reader utility for specific audiences [14, 15], as well as discourse coherence and informational adequacy [16, 17]. Empirical evidence suggests that LLM-generated simplifications can aid non-specialists, but expert oversight is often required to maintain precision and appropriate calibration [7].

Positioning. Our study operationalizes these insights by combining: (i) baseline LLM simplifications; (ii) structured feedback from STEM readers outside computer science on understanding, naturalness, and simplicity; and (iii) expert post-editing to produce reference simplifications. This design directly targets the known trade-off between readability and domain fidelity, and provides a corpus and evaluation setting aligned with real cross-disciplinary use.

3. Methodology

Our methodology combines LLM-based simplification with targeted human feedback. It consists of corpus selection, initial simplification, a two-stage user study, and automatic and qualitative evaluation of the resulting simplifications.

3.1. Dataset

We adopt SciSummNet [9], a corpus of 1,000 highly cited ACL papers with abstracts, citation contexts, and 150-word expert-written summaries. Although these summaries capture the central contributions of the respective papers, they often retain dense terminology and complex syntactic structures. Prior studies suggest that summarization alone does not guarantee readability, while unrestricted simplification can cause verbosity or information loss [18]. We therefore use the SciSummNet summaries as source texts and simplify them for readers outside computer science while aiming to preserve their scientific content.

[1]Our code and data are available at https://github.com/faerber-lab/scientific-text-simplification-corpus.

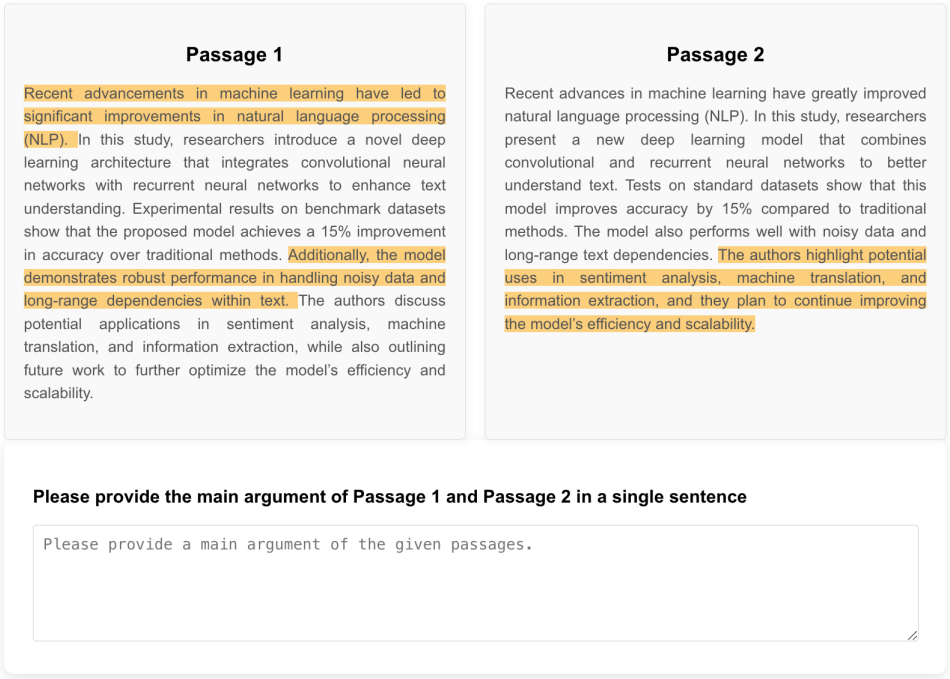

Figure 1: Interface for selecting difficult sentences in Phase 1.

3.2. Baseline Simplification

We generate initial simplified summaries using GPT-4o-mini in a zero-shot setting, i.e., without providing input-output demonstrations. The system prompt asks the model to simplify the text for readers without expertise in the scientific field and for the general public, identify complex words or key phrases, and preserve the paragraph format. The user prompt provides the paper title and source summary and asks the model to retain the title while simplifying the summary. The exact prompt is included in the released repository. These LLM outputs serve as the baseline for subsequent human evaluation and refinement.

3.3. Phase 1: Cross-Disciplinary Reader Evaluation

Participants. Participants were recruited from STEM disciplines outside computer science and were fluent in English.
Design. For each evaluation, a participant was presented with an original SciSummNet summary and its GPT-simplified version in randomized order. Participants (i) selected sentences they found difficult to understand, (ii) made a three-way comparative judgment for *Understanding*, *Naturalness*, and *Simplicity* by preferring the original version, preferring the GPT-simplified version, or indicating no difference, and (iii) highlighted specific words or phrases they considered complex. These are the labels used in the interface; elsewhere, understanding and naturalness are occasionally discussed using the closely related terms coherence and fluency, respectively. Across the completed evaluations, this resulted in 92 comparative judgments per dimension. The annotation interface and task flow are illustrated in Figures 1–3.

3.4. Phase 2: Expert-Edited Simplifications

Participants and materials. Editors were computer science graduate students or researchers with strong English proficiency. The Phase 2 task covered 92 candidate summaries, of which 47 were completed and included as expert-edited references in the current corpus and evaluation. For each item, experts received (i) the original summary, (ii) the GPT-simplified version, and (iii) user highlights from

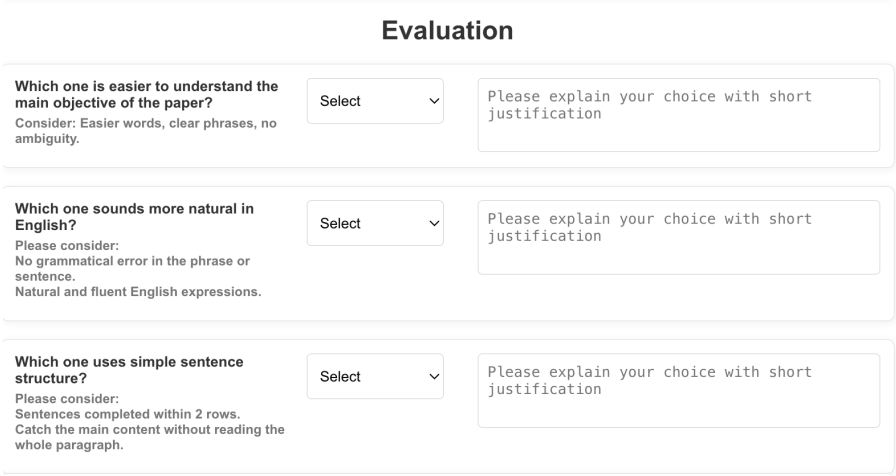

Figure 2: Interface for comparing summaries by understanding, naturalness, and simplicity in Phase 1.

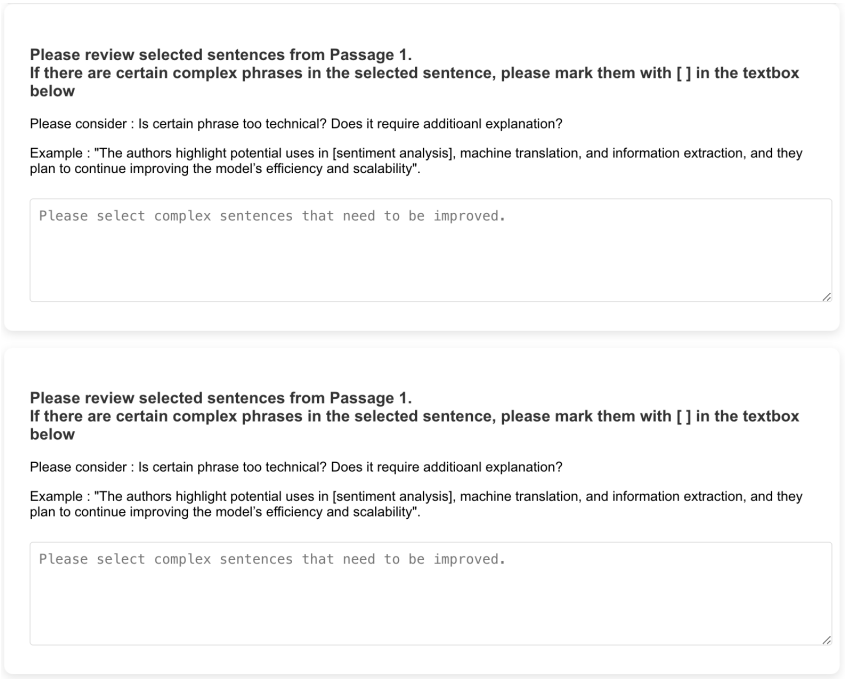

Figure 3: Interface for highlighting difficult words and phrases in Phase 1.

Phase 1 indicating difficult words or phrases.
Procedure. Experts created an expert-edited reference simplification following four guided scenarios:

1. **Both** summaries contain difficult spans: refine and merge content for clarity and fidelity.
2. **Only Original** flagged: verify and edit GPT simplifications where necessary.
3. **Only GPT** flagged: correct over-simplifications or awkward phrasing.
4. **No Flags**: polish structure and verify meaning retention.

Our editing guidelines emphasize retaining technical terminology when required for accuracy, preserving the strength and scope of scientific claims, and using shorter, well-structured sentences where possible. The user interface is shown in Figure 4.

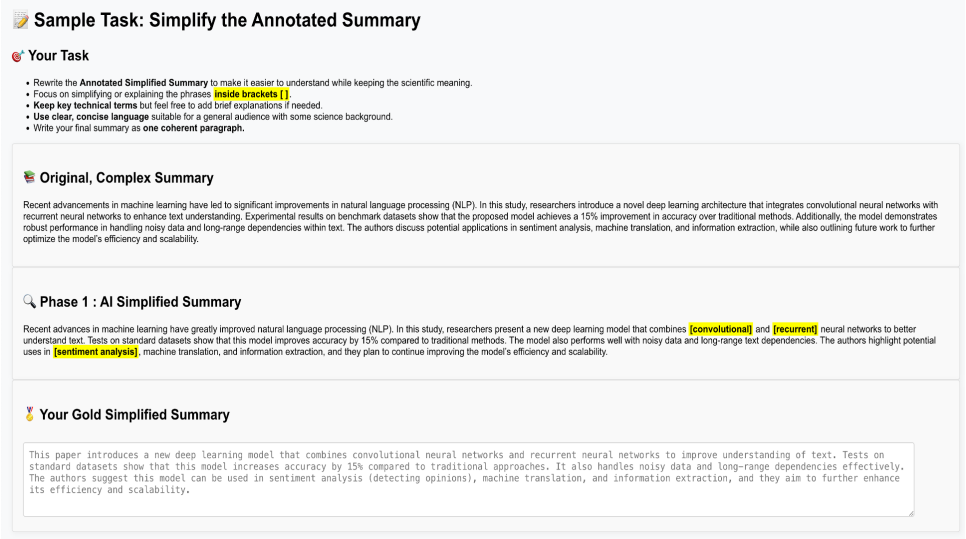

Figure 4: Interface for creating expert-edited reference simplifications in Phase 2.

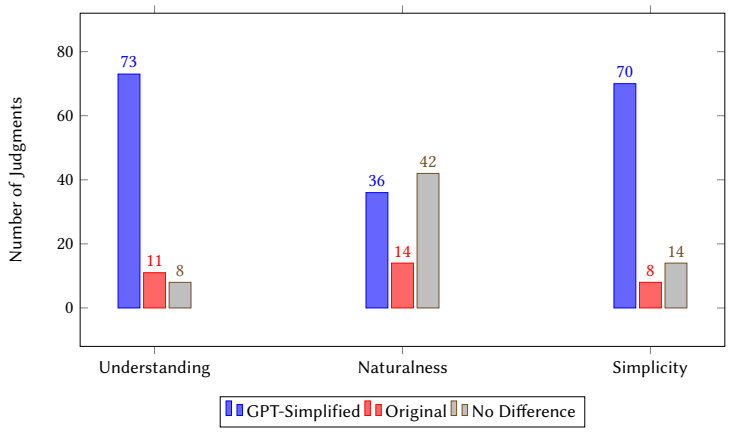

Figure 5: Comparative judgments across the three Phase 1 evaluation dimensions. Each dimension comprises 92 judgments; no inferential significance test was conducted.

4. Results

We evaluate (i) how cross-disciplinary STEM readers perceive LLM-generated simplifications of scientific summaries in Phase 1 and (ii) how expert-edited reference simplifications compare with the LLM outputs according to automatic metrics in Phase 2. We complement the quantitative results with qualitative analyses of where automated simplification is effective and where expert editing remains important.

4.1. Phase 1: Cross-Disciplinary Reader Judgments

Preferences. As shown in Figure 5, participants favored the LLM-generated summaries for *Understanding* in 73 of 92 judgments and for *Simplicity* in 70 of 92 judgments. This indicates a clear descriptive preference for the simplified versions. Judgments of *Naturalness* were more balanced: 42 indicated no difference, whereas 14 favored the original summary, suggesting that the LLM outputs occasionally contained stylistic inconsistencies or overly informal phrasing.

Difficulty Annotations. Sentence- and phrase-level annotations show that participants most frequently flagged multi-noun compounds, unexplained abbreviations, and domain-specific terminology. These annotations informed Phase 2 editing by identifying where simplification should expand compressed terminology, improve discourse flow, or clarify referents.

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 6: Example participant feedback on language complexity in scientific summaries.

Qualitative Insights. Participant comments and highlights, shown in Figures 6 and 7, reveal two recurring tensions: (1) **Precision vs. Approachability:** replacing terms such as *latent variables*, *PCFG*, or *NP-hard* with generic paraphrases can improve readability but reduce technical precision; and (2) **Style and Naturalness:** the LLM often improves local clarity but occasionally introduces awkward or overly informal formulations, whereas the original summaries generally retain a more conventional academic style.

These observations highlight a central limitation of automated simplification: LLMs can improve surface readability while compromising domain fidelity or stylistic appropriateness. Phase 2 therefore uses expert post-editing to refine the LLM outputs while preserving their scientific meaning.

4.2. Phase 2: Evaluation of Expert-Edited Simplifications

Evaluation Setup. The evaluation covers the 47 items for which an original SciSummNet summary, an LLM-generated simplification, and a finalized expert-edited simplification are available. In the *GPT-Simplified* condition, each original summary is compared with its LLM-generated version. In the *Expert-Edited* condition, the same original summary is compared with its expert-edited version.

We report four types of automatic measurements. Sentence-level BLEU is calculated for each lowercased and tokenized source-output pair and averaged across the 47 items. BERTScore F1 is calculated using `distilbert-base-uncased` and likewise averaged across items. SARI is computed without human references by supplying empty references and should therefore be interpreted as a descriptive, reference-free measure of edit operations rather than standard reference-based SARI. Finally, Flesch-Kincaid Grade Level and Reading Ease are computed over the concatenated outputs to characterize corpus-level surface readability. The metric scales and preferred directions are reported in Table 1.

Understanding: "Passage 1 explains the technical words better."

"Original": "This paper defines a generative probabilistic model of parse trees, which we call PCFG-LA. This model is an extension of PCFG in which non-terminal symbols are augmented with **latent variables**. Fine-grained **CFG rules** are automatically induced from a parsed corpus by training a PCFG-LA model using an EM-algorithm. Because exact parsing with a **PCFG-LA is NP-hard**, several approximations are described and empirically compared. In experiments using the Penn WSJ corpus, our automatically trained model gave a performance of [86.6% (F1, sentences <= 40 words)], which is comparable to that of an unlexicalized [PCFG parse]r created using extensive manual feature selection. We use a markovized grammar to get a better unannotated parse forest during decoding, but we do not markovize the training data. We right-binarize the tree bank data to construct grammars with only unary and binary productions.",

"GPT-Simplified": "This paper describes a method for creating parse trees, which are structures used in understanding sentences, called PCFG-LA. This method improves upon a basic version known as PCFG by **adding hidden elements** to the symbols used in the trees. Detailed rules for constructing these trees are automatically learned from a collection of example sentences by training the PCFG-LA model using a technique called the EM-algorithm. Since finding an exact solution with PCFG-LA is **very complex and difficult (NP-hard)**, the paper discusses several simplified methods and compares them through experiments. In tests with a well-known set of sentences (Penn WSJ corpus), our model, which was trained automatically, performed quite well with an accuracy of 86.6% for sentences up to 40 words long, similar to another model that required a lot of manual adjustment. We use a special method (markovized grammar) to get better results when interpreting sentence structures but do not apply this method to the training examples. We also change the structure of example sentence data to make it simpler, using only two types of constructions (unary and binary)."

Figure 7: Example participant feedback highlighting clarity issues in scientific summaries.

Category	Metric and scale	GPT-Simplified	Expert-Edited
Source similarity	Mean sentence BLEU (0–1, ↑)	0.949	0.349
	Mean BERTScore F1 (−1–1, ↑)	0.9907	0.9031
Simplification	Reference-free SARI (0–100, ↑)	46.76	33.72
Readability	FK Grade Level (no fixed bounds, ↓)	12.10	13.00
	FK Reading Ease (no fixed bounds, ↑)	48.54	40.89

Table 1
Automatic evaluation of 47 aligned source summaries. For GPT-Simplified, each original summary is compared with its GPT-4o-mini output; for Expert-Edited, the same original is compared with the corresponding expert-edited output. BLEU and BERTScore are averaged across aligned pairs, whereas the Flesch-Kincaid scores are computed over the concatenated output texts. SARI is calculated without human references and does not include the standard reference-based *keep* component. Arrows indicate the conventionally preferred direction. All differences are descriptive; no inferential significance tests were conducted.

Source Similarity. The GPT outputs receive higher source-similarity scores than the expert-edited outputs, including mean sentence-level BLEU (0.949 vs. 0.349) and BERTScore F1 (0.9907 vs. 0.9031). These values indicate that the GPT outputs remain substantially closer to the wording of the original summaries. They do not, by themselves, demonstrate higher semantic quality, as deliberate expert rewordings may preserve meaning while reducing lexical overlap.

Simplification Operations. The reference-free SARI variant produces a higher score for the GPT simplifications than for the expert-edited summaries (46.76 vs. 33.72). Because the calculation uses no human references and does not include the standard reference-based *keep* component, this difference should be interpreted as a descriptive indication of the performed edit operations rather than an overall quality judgment.

COMLEX Syntax: Building a Computational Lexicon

(Original)

We describe the design of Comlex Syntax, a **computational lexicon** providing detailed **syntactic information** for **approximately** 38,000 English headwords. We consider the types of errors which arise in creating such a lexicon, and how such errors can be measured and controlled. Our COMLEX syntax dictionary provides verb subcategorization information and syntactic paraphrases, but they are indexed by words thus not suitable to use in generation directly.

(GPT-Simplified)

We explain how we created Comlex Syntax, a **computer-based dictionary** that gives detailed **grammar information** for **about** 38,000 main English words. We look at the kinds of mistakes that can happen when making this dictionary, and how we can find and fix these mistakes. Our COMLEX syntax dictionary includes details about how verbs are used and different ways to say things in sentences, but because they are **organized by words**, it is not easy to use them for creating new sentences directly.

(Expert-Edited)

This paper details the design of Comlex Syntax, a **computational lexicon** offering extensive **syntactic information** for **approximately** 38,000 English headwords. The authors discuss the types of errors that can occur during lexicon creation and methods for their measurement and control. This COMLEX syntax dictionary provides verb subcategorization information and syntactic paraphrases, but because these are **indexed by** individual words, they are not directly suitable for generating new sentences.

Figure 8: Comparison of an original, GPT-simplified, and expert-edited summary from Phase 2.

Readability. The GPT-simplified summaries receive a lower FKGL value (12.10 vs. 13.00) and a higher FKRE value (48.54 vs. 40.89). The readability formulas therefore favor the GPT outputs in terms of surface-level features. However, these formulas do not assess terminological accuracy, conceptual clarity, or scientific claim calibration.

Qualitative Comparison. The expert edits follow a hybrid post-editing strategy. Editors preserve domain-critical terminology, such as *computational lexicon* and *syntactic information*, while selectively adopting clearer LLM-generated formulations for more general explanatory content. This approach aims to improve accessibility without sacrificing scientific precision. An example is shown in Figure 8.

Summary. Overall, the automatic metrics favor the GPT outputs, partly because they remain closer to the source wording and exhibit simpler surface-level characteristics. The qualitative analysis nevertheless shows how expert post-editing can preserve domain terminology and scientific claim calibration while selectively retaining useful LLM-generated phrasing. As the expert-edited summaries were not independently evaluated by the target readers, these observations should be treated as complementary evidence rather than proof that either condition is uniformly superior.

4.3. Discussion and Implications

Relation to Prior Work. The Phase 1 results are consistent with prior studies showing that LLM-generated simplifications can improve accessibility for readers outside the source domain while introducing risks related to terminology, nuance, and claim strength [6, 7]. The Phase 2 examples further illustrate how expert post-editing can retain or restore domain terminology and calibrated scientific formulations. At the same time, the automatic results reinforce previous criticism of relying on source-overlap metrics alone for text simplification [13]: Higher similarity to the source does not necessarily imply greater accessibility or better preservation of the intended scientific meaning.

Practical Implications. The results suggest that LLM-generated simplifications are most useful as initial drafts rather than as directly publishable scientific texts. Feedback from readers outside the source domain can reveal passages that remain difficult, while expert post-editing is needed to verify terminology, preserve meaning, and maintain the appropriate scope and strength of scientific claims.

Research Implications. The corpus enables joint evaluation of readability and domain fidelity by linking original summaries with LLM simplifications, reader annotations, and expert-edited references. It can therefore support the development and evaluation of models that simplify scientific texts without compromising terminology or scientific meaning.

5. Limitations and Challenges

Although our results demonstrate the potential of LLM-based and human-in-the-loop approaches for scientific text simplification, several limitations remain.

Scale and Coverage. Due to participant constraints, expert-edited reference simplifications were created for only 47 of the 1,000 source summaries. This subset may not capture the full linguistic and conceptual diversity of the dataset. Future work should extend the annotations across domains and difficulty levels.

Participant Sampling. Phase 1 involved STEM readers without a computer science background, whereas Phase 2 editors were computer science researchers. Although this design reflects the intended distinction between target readers and domain experts, it may introduce sampling bias and does not cover other relevant audiences, such as researchers from the social sciences or humanities, educators, or policy professionals.

Metric Limitations. Metrics such as BLEU, BERTScore, SARI, and Flesch-Kincaid capture only selected aspects of similarity and readability. They do not adequately measure semantic nuance, rhetorical clarity, or factual fidelity, which are central to scientific text simplification. Moreover, our analysis reports aggregate values without paired significance tests, and the SARI evaluation is reference-free. The expert-edited summaries were also not evaluated independently by target readers. The reported differences should therefore be interpreted descriptively.

Domain and Language Scope. The study is limited to English-language computer science texts. As scientific terminology and communication practices differ across disciplines and languages, the results may not generalize directly to other settings. Extending the workflow to multilingual corpora and additional scientific domains would improve its broader applicability.

Model Dependence. The initial simplifications were generated using GPT-4o-mini and a single fixed zero-shot prompt. Different models, prompts, or decoding strategies may produce different results. Systematic comparisons across models and prompting approaches remain future work.

Annotation Reliability. Phase 1 annotations reflect perceived difficulty and may therefore contain noise. In particular, participants may conflate unfamiliar terminology with linguistic complexity. Expert post-editing partly mitigates this limitation, but the reliability of the annotations should be examined more systematically.

Ethical Considerations. Text simplification may unintentionally distort scientific claims, for example by overstating findings or omitting methodological qualifications. Expert validation reduces this risk but cannot eliminate it. Simplified texts should therefore be used with particular care in educational, public-facing, and policy-related contexts.

6. Conclusion

This work introduced a human-in-the-loop framework for simplifying scientific summaries with large language models (LLMs), addressing accessibility barriers in interdisciplinary research communication. Starting from expert-written summaries in SciSummNet, we generated baseline simplifications using GPT-4o-mini and engaged cross-disciplinary STEM readers to identify linguistic and conceptual difficulties. Their annotations informed a second phase in which domain experts revised the LLM outputs to produce expert-edited reference simplifications.

Phase 1 shows that readers generally preferred the GPT-generated versions for understanding and simplicity, while also identifying cases of awkward phrasing and reduced technical precision. In Phase 2, the automatic metrics generally favor the GPT outputs because they remain closer to the source, whereas qualitative inspection illustrates how expert post-editing can preserve terminology and scientific claim calibration. These findings motivate treating LLM simplifications as drafts that benefit from feedback by cross-disciplinary readers and domain-informed review.

The resulting corpus includes original summaries, LLM-generated simplifications, reader annotations, and expert-edited references, providing a reproducible resource and evaluation setting for scientific text simplification.

Future work will extend the number and disciplinary coverage of the expert-edited summaries, evaluate these summaries with independent target readers, and compare alternative models and prompting strategies. Further work should also develop evaluation methods that jointly capture accessibility, factual fidelity, terminology preservation, and scientific claim calibration.

Data and Code Availability

The corpus, annotation artifacts, and code are publicly available at: https://github.com/faerber-lab/scientific-text-simplification-corpus

Ethics Statement

The study involved cross-disciplinary STEM readers and domain experts. All participants provided informed consent and could withdraw at any time. We collected no personally identifying information. As text simplification may alter scientific claims, our guidelines and expert review explicitly focused on preserving meaning and terminology.

Acknowledgements

This research was funded by the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) and the German Federal Ministry of Research, Technology and Space (BMFTR) through the Software Campus project (01IS23070). We thank the reviewers and Shuzhou Yuan for valuable feedback.

References

[1] M. Bonaventura, V. Latora, V. Nicosia, P. Panzarasa, The advantages of interdisciplinarity in modern science, arXiv preprint (2017). URL: https://doi.org/10.48550/arXiv.1712.07910.

[2] E. Cunningham, B. Smyth, D. Greene, Collaboration in the time of COVID: A scientometric analysis of multidisciplinary SARS-CoV-2 research, Humanities and Social Sciences Communications 8 (2021) 240. URL: https://doi.org/10.1057/s41599-021-00922-7.

[3] S. S. Al-Thanyyan, A. M. Azmi, Automated text simplification: A survey, ACM Computing Surveys 54 (2022). URL: https://doi.org/10.1145/3442695.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in Neural Information Processing Systems 33 (2020) 1877–1901.

[5] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, arXiv preprint (2021). URL: https://doi.org/10.48550/arXiv.2109.01652.

[6] B. Engelmann, F. Haak, C. K. Kreutz, N. Nikzad-Khasmakhi, P. Schaer, Text simplification of scientific texts for non-expert readers, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), 2023, pp. 2987–2998. URL: https://ceur-ws.org/Vol-3497/paper-250.pdf.

[7] T. Guidroz, D. Ardila, J. Li, A. Mansour, P. Jhun, N. Gonzalez, X. Ji, M. Sanchez, S. Kakarmath, M. Bellaiche, M. A. Garrido, F. Ahmed, D. Choudhary, J. Hartford, C. Xu, H. J. Serrano Echeverria, Y. Wang, J. Shaffer, E. Y. Cao, Y. Matias, A. Hassidim, D. R. Webster, Y. Liu, S. Fujiwara, P. Bui, Q. Duong, LLM-based text simplification and its effect on user comprehension and cognitive load, arXiv preprint (2025). URL: https://doi.org/10.48550/arXiv.2505.01980.

[8] J. Qiang, M. Huang, Y. Zhu, Y. Yuan, C. Zhang, K. Yu, Redefining simplicity: Benchmarking large language models from lexical to document simplification, arXiv preprint (2025). URL: https://doi.org/10.48550/arXiv.2502.08281.

[9] M. Yasunaga, J. Kasai, R. Zhang, A. R. Fabbri, I. Li, D. Friedman, D. R. Radev, ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 7386–7393. URL: https://doi.org/10.1609/aaai.v33i01.33017386.

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

[11] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[12] J. P. Kincaid, J. Fishburne, Robert P., R. L. Rogers, B. S. Chissom, Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel, Technical Report Research Branch Report 8-75, Chief of Naval Technical Training, 1975. URL: https://stars.library.ucf.edu/istlibrary/56/.

[13] E. Sulem, O. Abend, A. Rappoport, BLEU is not suitable for the evaluation of text simplification, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 738–744. URL: https://doi.org/10.18653/v1/D18-1081.

[14] A. Siddharthan, Syntactic simplification and text cohesion, Research on Language and Computation 4 (2006) 77–109. URL: https://doi.org/10.1007/s11168-006-9011-1.

[15] S. Vajjala, D. Meurers, Readability assessment for text simplification: From analysing documents to identifying sentential simplifications, ITL - International Journal of Applied Linguistics 165 (2014) 194–222. URL: https://doi.org/10.1075/itl.165.2.04vaj.

[16] W. Xu, C. Callison-Burch, C. Napoles, Problems in current text simplification research: New data can help, Transactions of the Association for Computational Linguistics 3 (2015) 283–297. URL: https://doi.org/10.1162/tacl_a_00139.

[17] D. Kauchak, Improving text simplification language modeling using unsimplified text data, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013, pp. 1537–1546. URL: https://aclanthology.org/P13-1151/.

[18] F. Zaman, F. Kamiran, M. Shardlow, S.-U. Hassan, A. Karim, N. R. Aljohani, SATS: Simplification-aware text summarization of scientific documents, Frontiers in Artificial Intelligence 7 (2024) 1375419. URL: https://doi.org/10.3389/frai.2024.1375419.