Improving Textbook Accessibility through AI Simplification: Readability Improvements and Meaning Preservation

Benny G. Johnson[1,*], Bill Jerome[1], Jeffrey S. Dittel[1], and Rachel Van Campenhout[1]

[1] *VitalSource Technologies, Raleigh, NC 27601, USA*

Abstract

Generative artificial intelligence has the potential to tackle many longstanding educational challenges, including helping students comprehend difficult textbook material. Textbooks are considered the gold-standard for rigorous and expert-developed educational content, yet still pose challenges to students who struggle with the complexity of textbook language. A strength of large language models (LLMs) is their ability to manipulate text according to specific requirements. An LLM was harnessed to create a “simplifier” tool in a higher education ereader platform, allowing students to select a textbook passage and receive a simplified version of that content. In this study, we analyzed 54,371 simplifier interactions to compare the original textbook content and simplified versions according to estimated readability, lexical and syntactic simplification, and semantic fidelity. Results indicate that the simplifier tool was able to reduce complexity of the original text while maintaining meaning, laying groundwork for future studies involving student perception and comprehension outcomes. The practical implications of this tool for enhancing textbook accessibility and supporting student comprehension are discussed.

Keywords

artificial intelligence, large language models, simplification, readability, semantic fidelity

1. Introduction

Textbooks are considered the gold-standard for content, as they are developed by reputable subject matter experts and subjected to thorough accuracy reviews. Textbooks are often assigned by faculty for students to read as part of their coursework; however, it is known through decades of research that students do not read as expected. A longitudinal study between 1981 and 1997 found student textbook reading declined in that period [1]. Several research studies across disciplines found that only a small percentage of students (16–27%) reported reading before class [2–5]. Russell et al. [6] gained further insights into student reading using ebook data, finding that when faculty did not employ a reading strategy, students read only 14% of the textbook. When asked why they did not read, in addition to common factors such as time limitations or level of perceived importance, some students noted they needed scaffolding for readings and were unsure of how to approach the textbook [6].

Difficulty and readability were factors for struggling students when interviewed as part of an internal user experience study. College students, who had previously used learning features as part of their traditional university courses, volunteered to discuss new learning feature prototypes as well as their own experiences, motivations, and struggles. Students reported that they often found the textbook content intimidating, struggled with the language, or felt the material was too complex to comprehend. This led students to give up on their reading assignments prematurely or avoid reading altogether. Their experiences provided additional tangible, student-centered motivation to tackle the central challenge of textbook readability. Although the current study does not assess students’

---

iTextbooks'25: Sixth Workshop on Intelligent Textbooks, July 26, 2025, Palermo, Italy

[*] Corresponding author.

benny.johnson@vitalsource.com (B. G. Johnson); bill.jerome@vitalsource.com (B. Jerome); jeff.dittel@vitalsource.com (J. S. Dittel); rachel.vancampenhout@vitalsource.com (R. Van Campenhout)

0000-0003-4267-9608 (B. G. Johnson); 0000-0002-4200-155X (B. Jerome); 0000-0002-4913-4427 (J. S. Dittel); 0000-0001-8404-6513 (R. Van Campenhout)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

perceptions of simplified text, it represents an essential first step in validating whether AI-based simplifications are linguistically simpler and semantically faithful in authentic use.

While student interviews provide valuable contextualization of the ways in which readability can deter some students, Sheridan-Thomas [7] frames textbook comprehension as an issue of access and equity:

> Students who struggle with extracting important information and making meaning from textbook reading do not have the same access to course material as competent textbook readers. Helping all students comprehend textbook reading is an equity issue. For courses in which textbooks are used, whether as the main source of information or as a secondary reference, all students need to be able to use the textbook with as much competence and independence as possible. (p. 267)

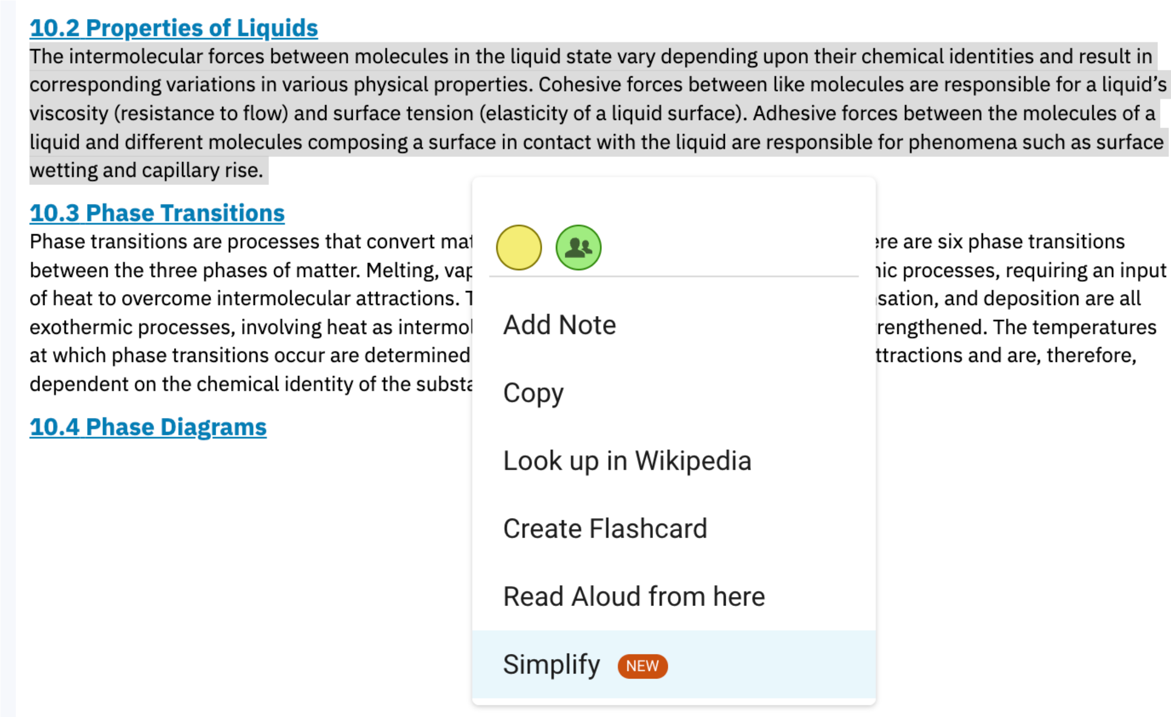
Strategies for teachers to support students are also discussed; however, although instructors can employ various pedagogical strategies to mitigate comprehension difficulties, providing personalized reading assistance at scale often exceeds available instructional resources. Consequently, scalable technological solutions are increasingly appealing.

The significant advancement of large language models (LLMs) has made it possible to address this challenge in a personalized manner. A strength of LLMs is their ability to manipulate language as directed, making them promising tools for producing accessible, simplified versions of complex academic texts. Recent studies have begun to explore LLM-based simplification. For example, Guidroz et al. [8] conducted a randomized controlled study involving over 4,500 participants, demonstrating that LLM-generated simplifications significantly improved reading comprehension and reduced perceived cognitive load, particularly in complex domains such as biomedical articles and financial texts. Similarly, recent LLM-based tools such as SimplifyMyText have been specifically designed to create plain-language adaptations aimed at enhancing inclusivity and accessibility [9]. Additionally, progressive approaches proposed by Fang et al. [10] illustrate that LLMs can handle complex document-level simplifications by systematically decomposing tasks from discourse-level down to lexical-level adjustments. However, although these studies collectively demonstrate the potential of LLMs for simplifying complex texts, empirical evidence from authentic educational settings remains limited.

To address this gap, the current study investigates student-initiated LLM-based simplifications generated within real-world textbook environments using an embedded ereader interface. In fall 2024, the VitalSource Bookshelf platform introduced an LLM-powered text simplification tool as a free enhancement within the ereader interface (available in textbooks from publishers granting permission for generative AI features). Students can highlight a passage of text and select "Simplify" (Figure 1) to receive a simplified version displayed in an interactive side panel chat window next to the textbook content (Figure 2). The primary goal of the simplifier is to reduce lexical and syntactic complexity to improve readability and comprehension for students. It does not explicitly aim to summarize or deeply elaborate content beyond clarifying complex sentences.

After the initial simplification, the student is prompted to attempt to restate the content in their own words to check their understanding or ask for another simplification; however, the current study focuses specifically on analyzing the initial simplification event.

The theoretical underpinning for the simplification approach employed here aligns with cognitive load theory [11], which posits that learning is optimized when extraneous cognitive load is minimized. Complex lexical and syntactic structures in textbooks can represent extraneous load, hindering students' ability to engage deeply with instructional content, particularly when combined with the inherently high intrinsic cognitive load of challenging academic material. By simplifying these linguistic structures, we aim to reduce unnecessary cognitive effort, enabling students to better construct coherent mental representations as described by Kintsch's Construction-Integration Model [12]. This theoretically informed approach highlights the potential for simplified text to support not just immediate readability, but deeper comprehension and retention of complex academic content.


10.2 Properties of Liquids

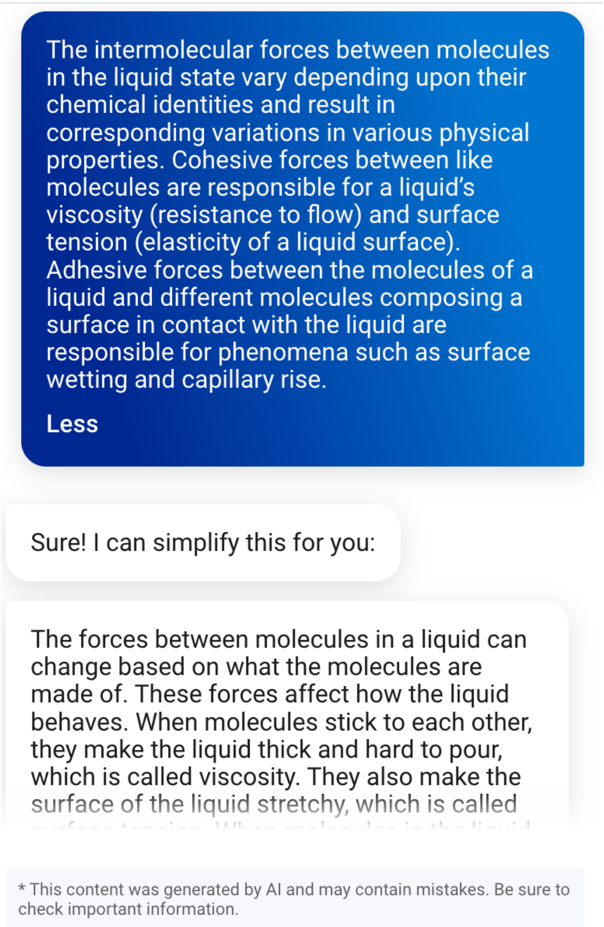
The intermolecular forces between molecules in the liquid state vary depending upon their chemical identities and result in corresponding variations in various physical properties. Cohesive forces between like molecules are responsible for a liquid’s viscosity (resistance to flow) and surface tension (elasticity of a liquid surface). Adhesive forces between the molecules of a liquid and different molecules composing a surface in contact with the liquid are responsible for phenomena such as surface wetting and capillary rise.

10.3 Phase Transitions

Phase transitions are processes that convert mat … ere are six phase transitions between the three phases of matter. Melting, vap … ic processes, requiring an input of heat to overcome intermolecular attractions. T … sation, and deposition are all exothermic processes, involving heat as intermol … rengthened. The temperatures at which phase transitions occur are determined … ttractions and are, therefore, dependent on the chemical identity of the substa

10.4 Phase Diagrams

Figure 1: Activating the Simplify tool in the Bookshelf ereader interface by highlighting textbook content and selecting "Simplify."


The intermolecular forces between molecules in the liquid state vary depending upon their chemical identities and result in corresponding variations in various physical properties. Cohesive forces between like molecules are responsible for a liquid’s viscosity (resistance to flow) and surface tension (elasticity of a liquid surface). Adhesive forces between the molecules of a liquid and different molecules composing a surface in contact with the liquid are responsible for phenomena such as surface wetting and capillary rise.

Sure! I can simplify this for you:

The forces between molecules in a liquid can change based on what the molecules are made of. These forces affect how the liquid behaves. When molecules stick to each other, they make the liquid thick and hard to pour, which is called viscosity. They also make the surface of the liquid stretchy, which is called

* This content was generated by AI and may contain mistakes. Be sure to check important information.

Figure 2: Simplified text displayed in an interactive side chat panel within the ereader.

This study examines the practical effectiveness of simplifications across a substantial dataset of more than fifty thousand requests, assessing their impacts along key dimensions of readability, lexical and syntactic simplification, and semantic fidelity. Specifically, the study addresses two core research questions:

- **RQ1:** To what extent does LLM-based simplification improve the readability of textbook content?
- **RQ2:** Does the simplification process reliably preserve original meanings?

By grounding analysis in spontaneous student engagement and systematically assessing simplifications in authentic educational settings, this study contributes to understanding how LLM-generated simplifications function in practice and lays the groundwork for evaluating their potential to support student comprehension and educational equity. Although cognitive load theory motivates the design of the simplifier, the current study does not directly assess cognitive load reduction.

2. Method

2.1. Simplification Procedure

The textbook simplifications were generated using OpenAI GPT-4o [13]. The process was carried out on demand in real time upon student request, using an interface embedded in the ereader platform (Figures 1 and 2). Parameter settings were `temperature = 0`, `top_p = 1`, and `max_tokens = 4095` to ensure consistency and determinism in the simplification outputs.

The simplifier was prompted to act as a helpful college professor, assisting a student who reported difficulty understanding highlighted sentences from a textbook. The prompt explicitly instructed the simplifier to reduce sentence complexity, decrease reading level by approximately four grade levels, substitute specialized terms with simpler and more general vocabulary, and maintain a conversational, positive tone. However, no explicit instructions were given regarding the target length of simplified text, nor was guidance provided against summarizing beyond simplification.

The LLM was given the student-selected text requiring simplification along with additional surrounding content from the textbook. Context was determined by including the immediate paragraph containing the selection and the larger section or subsection enclosing that paragraph. This approach aimed to provide sufficient relevant context without incorporating excessive portions of text. However, due to variation in textbook formatting, the extracted context may not always align precisely with clearly defined chapter subsections. While the LLM may have been pretrained on similar domain material, providing local textbook context helps ensure the simplification is tailored to the student's selected passage. No additional post-simplification filters or fidelity checks were applied during real-time student interactions.

2.2. Data Collection and Analysis

The dataset consists of student-initiated simplification events recorded between September 1, 2024, and April 30, 2025. The dataset contains 54,371 events generated by 11,689 students across 2,082 distinct textbooks. This dataset is publicly available in our open-source data repository [14], facilitating replication and further research. Almost 95% of usage was from higher education institutions in the United States and Canada. The ereader platform did not collect any student demographic characteristics. Using the BISAC major subject heading classification for the textbooks [15], the top subject domains as a percentage of the data were Social Science (29.7%), Political Science (16.1%), and Psychology (13.8%). The impact of simplification was quantified along four dimensions:

- Readability: the ease with which a text can be read and processed, related to amount of effort required by the reader
- Lexical simplification: the replacement of complex words or phrases with simpler alternatives
- Syntactic simplification: the reduction of structural complexity of sentences
- Semantic fidelity: the degree to which the original meaning is maintained after simplification

Each metric was computed on both the original student-selected passage and its LLM-simplified version. Differences between simplified and original texts (Δ = simplified − selected) were analyzed. Because the simplification feature addresses both lexical and syntactic aspects of the student-selected text, additional analyses were performed to gain insight into the relative contribution of each type of simplification.

Readability improvements (RQ1) were measured using two widely recognized metrics. The primary measure was the Flesch–Kincaid Grade Level (FKGL), chosen for its direct interpretability in terms of U.S. educational grade levels [16]. Complementing FKGL, the Flesch Reading Ease (FRE) scale was used [17]. Notably, both FKGL and FRE rely on two core linguistic variables: average syllables per word, which primarily reflects lexical complexity, and average words per sentence, which captures syntactic complexity. However, these features are weighted differently by each metric; in particular, FRE places greater emphasis on word length compared to FKGL and thus serves as an additional robustness check. Both metrics were computed using the readability module in the NLTK library [18], which determines sentence boundaries, word counts, and syllable counts internally.

While FKGL and FRE have known limitations, they remain standard proxies for readability in educational research due to their transparency and alignment with grade-level norms. Their use here provides a practical and interpretable means of assessing changes in estimated readability across a large dataset. Future work may explore more advanced metrics, including those based on language models, to complement these analyses.

Lexical simplification was assessed by examining replacements of less common or more morphologically complex words with simpler alternatives. The primary lexical measure was the change in mean corpus log probability (Δ log *p*, natural log units) of content words (nouns, verbs, adjectives, and adverbs), calculated using the precomputed probabilities available from spaCy's `en_core_web_lg` model (version 3.5.0) [19]. This model's word frequency estimates are based on large-scale web and news corpora, making it suitable for general-purpose lexical analysis. Positive Δ values indicate substitutions of less frequent words with higher-frequency (more common) words, generally corresponding to simpler vocabulary [20]. As a complementary lexical measure, the change in average word length in characters was computed, reflecting morphological simplification.

Syntactic simplification was assessed by measuring reduction in sentence structural complexity. The primary syntactic measure was the average change in dependency tree depth (Δ dependency depth), calculated as the mean distance (in ancestor links) from each word to its sentence root using spaCy's dependency parser (`en_core_web_lg`). A well-established syntactic complexity measure is dependency length (linear distance between syntactically linked words), which has been shown to impact processing difficulty. While dependency depth is distinct from dependency length, it similarly reflects hierarchical structure and has been proposed as a proxy for syntactic complexity. Futrell et al. [21] provide large-scale evidence that minimizing syntactic dependencies supports processing efficiency, motivating the use of structural measures like depth in simplification analysis. To complement this, the change in average sentence length (Δ words / sentence) was computed, reflecting the degree to which simplification involved clause splitting.

Preserving semantic fidelity during simplification (RQ2) was evaluated primarily using cosine similarity between the selected and simplified text, computed on embeddings obtained via the `all-mpnet-base-v2` model from the `sentence-transformers` library [22], which has demonstrated strong performance across various semantic textual similarity tasks. Prior research has demonstrated that cosine similarities derived from Sentence-BERT embeddings correlate strongly with human judgments of semantic similarity [22]. Other metrics commonly used in text simplification research (e.g., BLEU, ROUGE, SARI) rely on reference-based comparisons and are oriented primarily toward evaluating lexical overlap or n-gram similarity with human-authored simplifications. Because the current study involves spontaneous, student-initiated simplifications without curated reference texts and prioritizes semantic fidelity and readability in authentic educational contexts, these metrics were not directly applicable.

To identify a cosine similarity threshold for acceptable semantic fidelity, an empirical validation procedure was conducted using an established semantic similarity rating scale [23], which ranges from 0 (different topics) to 5 (completely equivalent). The scale characterizes a rating of 4 as "mostly equivalent, but some unimportant details differ," which serves as a conservative minimum standard for acceptability in the context of assisting college students struggling with textbook readability. Ratings of 3 or lower indicate more substantial alterations, such as extensive summarization, which exceed the intended scope of the simplification approach.

Cosine similarity scores were partitioned into bands of width .1 (.5–.6, ..., .9–1.0). From each band, 40 original-simplified pairs were randomly selected. Each pair was independently rated for semantic similarity by OpenAI's o3 LLM [24], blinded to cosine similarity values. While the similarity ratings were generated by an LLM, the authors independently reviewed samples of these ratings and the rationale provided for each case and found them to be reasonable and well-aligned with human judgments. Prior work has also demonstrated strong correlation between LLM-based semantic similarity judgments and human ratings [22]. The lowest similarity band for which all randomly selected pairs received a rating of 4 or better was used to define the threshold. The sample size of 40 was chosen based on a power analysis using the Wilson method for estimating binomial proportion confidence intervals (CIs). Observing 40 consecutive acceptable ratings provides a 95% CI of 95.6% ± 4.4%, ensuring a lower bound of over 90% for the true proportion of acceptable cases. Adjacent bands were also evaluated, confirming those below the threshold failed to consistently achieve ratings of 4 or higher, whereas higher bands passed, reinforcing the threshold's stability.

Although alternative approaches could arrive at different threshold values, this approach is justified on several grounds: it is empirically derived, use-case-specific, reproducible, and leverages the LLM's broad semantic knowledge, making it likely better suited than a single human rater for evaluating pairs across numerous diverse textbook domains.

Because the `all-mpnet-base-v2` embedding model has an input limit of 384 tokens (i.e., subword units used by language models to process text, approximately 300 words), longer texts required truncation to meet this constraint. Such longer texts comprised 17.5% of student selections. Cosine similarities were examined separately for shorter (≤ 300 words) and longer (> 300 words) selected passages, finding the distributions to be closely aligned. This suggests that embeddings computed by truncating these longer passages still robustly represented their semantic content. Therefore, additional chunking or embedding aggregation methods were deemed unnecessary, as they would likely not substantially alter semantic similarity assessments when passages maintain a consistent focus throughout.

As a complementary semantic metric, the compression ratio (CR) was computed as the ratio of the simplified passage length to the original passage length in words. An automated metric in text simplification evaluation [25], CR serves as an additional diagnostic tool when interpreted jointly with cosine similarity. Because simplification strategies can vary, it is treated qualitatively rather than used as a strict threshold. For instance, a low cosine similarity coupled with a CR significantly below 1 may indicate excessive reduction or summarization, potentially leading to the omission of critical information. A study by Schwarzer [26] found that lower CRs (indicating greater length reduction) correlate positively with perceived simplicity but negatively with adequacy. Conversely, a low cosine similarity with a CR significantly above 1 could suggest elaboration or introduction of new content not present in the original text. By jointly analyzing cosine similarity and CR, different types of semantic divergence can be better identified and categorized, facilitating more targeted qualitative assessments.

A scatterplot of compression ratio versus cosine similarity revealed no visually discernible boundaries or clusters, indicating a continuous rather than categorical relationship between these metrics and semantic fidelity. Consequently, cosine similarity and CR thresholds were used as diagnostic guidelines rather than definitive indicators, highlighting the necessity of qualitative analysis to accurately assess potential meaning loss or elaboration.

Readability formulas such as FKGL and FRE assume continuous prose. When the selected text deviates significantly from typical prose, such as glossary entries, answer-key lists, or structured outlines, these formulas can yield extreme, uninformative values. For example, an extended run-on "sentence" created by a bulleted list of phrases lacking punctuation can artificially inflate FKGL or sharply decrease FRE scores, resulting in misleading values unrelated to the simplification tool's actual performance. Manual inspection revealed that virtually all passages assessed in the top 1% in reading difficulty by either metric (FKGL > 44.0 or FRE < -60.9) represented these non-prose formats. These outliers (*n* = 634, 1.2% of the dataset) were therefore excluded from analysis. Very low FKGL/high FRE values, indicating already-readable prose, were not removed because such passages could not exaggerate estimated readability improvements. Re-running the full dataset without trimming altered the mean FKGL improvement by ~0.5 grade levels and the mean FRE improvement by less than 2 points (but considerably reduced standard deviations), with no change to the overall statistical conclusions.

3. Results and Discussion

Table 1 presents descriptive statistics (mean, standard deviation, first quartile, median, third quartile) summarizing the extent of readability improvements, lexical and syntactic simplifications, and semantic fidelity across all simplification events analyzed. These results address RQ1 by quantifying improvements in readability. To address RQ2, we then examine semantic fidelity metrics in more detail, reporting the proportion of simplification events falling below the empirically determined cosine similarity threshold and analyzing illustrative examples to explore potential sources of divergence, such as shifts in structure, vocabulary, or emphasis that may affect perceived meaning.

Table 1
Descriptive statistics for readability, lexical, syntactic, and semantic fidelity metrics for simplification events.

Category	Metric	Mean	SD	Q_1	Median	Q_3
Readability	Δ FKGL	-7.37	4.61	-9.20	-6.53	-4.51
	Δ FRE	31.34	16.95	20.18	29.33	40.07
Lexical	Δ log p	1.02	0.52	0.66	0.98	1.33
	Δ chars / word	-0.41	0.37	-0.62	-0.38	-0.17
Syntactic	Δ dependency depth	-0.98	0.81	-1.33	-0.87	-0.48
	Δ words / sentence	-14.62	13.28	-18.65	-11.45	-6.62
Semantic	Cosine similarity	.85	.08	.81	.87	.91
	Compression ratio	0.80	0.22	0.66	0.79	0.92

3.1. Readability

Prior to simplification, the mean FKGL of selected textbook passages was 16.65, indicating content typically written at a level substantially above typical undergraduate reading expectations. The average simplification lowered FKGL by 7.37 grade levels to 9.28, bringing the text into a more accessible range for college-level readers. The interquartile range (Q_1 = -9.20, Q_3 = -4.51) shows that simplifications consistently resulted in meaningful readability improvements, with even the least-improved examples achieving several grade levels of improvement. The mean FRE increase was approximately 31 points (Q_1 = 20.18, Q_3 = 40.07), reinforcing that texts were easier to read post-simplification.

3.2. Lexical Simplification

Lexical simplification was assessed through changes in word familiarity and length. The mean increase of 1.02 in the $\Delta \log p$ metric corresponds roughly to a 2.8-fold increase in average content word frequency. This indicates words were replaced with more common synonyms, increasing lexical familiarity for readers. The interquartile range (0.66 to 1.33) indicates consistent lexical simplifications. Simplified texts also showed an average reduction of 0.41 characters per word, suggesting a preference for shorter, simpler words.

3.3. Syntactic Simplification

Simplified texts also showed clear evidence of structural simplification. Dependency depth, reflecting syntactic complexity, was reduced on average by 0.98 levels, signaling less complex sentence structures. With an interquartile range from -1.33 to -0.48, the results indicate readers encounter fewer deeply embedded modifiers, which lowers working memory load and improves clarity. The average simplification reduced sentence length by about 14.6 words. These reductions make sentences easier to parse and understand, supporting the syntactic effectiveness of the simplification process.

3.4. Semantic Fidelity

It is important to confirm that semantic fidelity (preservation of original meaning) is maintained alongside reductions in complexity. Mean cosine similarity was .85, suggesting that while the simplified texts differed structurally and lexically from their originals, the core meanings remained well-preserved. The narrow interquartile range (.81 to .91) indicates stable semantic fidelity across the majority of simplifications. Simplified texts on average retained 80% of the original length. The interquartile range (66% to 92%) shows variability, but consistently high ratios are consistent with simplifications that reduce extraneous complexity without losing critical information.

Collectively, these metrics indicate that substantial readability improvements can be achieved while maintaining high semantic fidelity. This alignment suggests high potential educational value for college students, as simplified texts effectively reduce cognitive load without compromising essential meaning.

To further assess semantic fidelity, an empirical threshold for acceptable cosine similarity was established at .7 using the procedure described in the Method section. Applying this threshold, 94.5% of original-simplified pairs demonstrated acceptable fidelity. The remaining 5.5% warrant further investigation, as potential loss of meaning is indicated. These lower-similarity cases likely reflect more substantial semantic shifts, such as unintended summarization, elaboration beyond simplification, or possible inaccuracies. Further analysis can characterize the nature of semantic divergences and inform subsequent refinement of the simplification process.

First, however, it is important to recognize that a cosine similarity below .7 does not automatically indicate an unacceptable simplification. Of the 2,980 pairs with low cosine similarity, the majority (76.8%) scored at least .6, i.e., only slightly below the threshold. The low-similarity cases were therefore divided into moderately low ($\geq .6$, $n = 2{,}289$, 4.3% of the dataset) and very low ($< .6$, $n = 691$, 1.3% of the dataset) similarity groups for further investigation.

Applying the previously established sampling method, 95% confidence intervals were calculated for the proportion of acceptable simplifications in each group. In the moderately low similarity group, 37 of 40 pairs were rated acceptable, resulting in a 95% CI of 88.8% ± 8.6%. In the very low group, 31 of 40 pairs were rated acceptable, giving a 95% CI of 75.1% ± 12.6%. These findings suggest that even at substantially lower cosine similarities, most simplifications remain acceptable in preserving semantic fidelity.

3.4.1. Example 1

The following pair from a Social Science textbook (cosine similarity = .55) illustrates how simplifications with relatively low cosine similarity can remain semantically acceptable.

> Insofar as corrections remains at the heart of our social policy—rather than as a supplemental or marginal support as it was throughout most of United States history—it is the Iron State stealing from the future of the Golden State.
>
> When we focus too much on prisons as a main part of our social policy, it takes away from other important areas. In the past, prisons were just a small part of our approach. Now, they take up a lot of attention and resources. This focus on prisons is like taking away from our future growth and success.

The simplification demonstrates substantial readability improvement, reducing FKGL from 19.0 to 7.2. Several factors appear to be contributing to the lower cosine similarity:

- Metaphorical language and named entities were removed. Terms like "Iron State" and "Golden State" carry substantial semantic weight through metaphor and allusions to the prison system and an idealized California, respectively. Omitting these terms in the simplified version removed a dense semantic anchor, thus lowering the cosine similarity despite the preservation of the core meaning.
- Key domain terms experienced semantic shifts. The term "corrections" was simplified to "prisons," "supplemental or marginal support" became "small part of our approach," and "stealing from the future" transitioned to "taking away from our future growth." While these paraphrases effectively retained the intended meaning, the phrase-level embeddings for each substitution may occupy different positions in semantic space.
- Rhetorical intensity and evaluative tone were softened in the simplified text. The original vivid and critical expression "stealing from the future" was rendered in more neutral economic language.

The combination of such factors illustrates how low similarity values may occur despite preservation of the original's essential meaning as perceived by human readers.

3.4.2. Example 2

Next, cases with potential substantive meaning loss are considered, specifically those exhibiting very low compression ratios. We consider cases where the simplified text contains fewer than half the original number of words (CR < 0.5). Among pairs with low cosine similarity, a strong negative correlation was observed between the length of original selections and their CR ($r = -.72$, $p < .001$). This highlights that longer original passages were substantially more likely to undergo extensive summarization in simplification.

Detailed examination of these pairs revealed notable trends. The primary strategy identified was summarization and condensation, going further than lexico-syntactic simplification for readability improvement. This approach led to significant reductions in supporting details, examples, historical context, nuanced definitions, and sometimes important qualifications originally present. Despite these substantial reductions, simplified versions generally retained accurate representations of the original core ideas. Semantic drift, although possible, was typically minimal, and hallucinations were not observed in these cases.

The following simplification from a Psychology textbook (CR = 0.36, cosine similarity = .54) illustrates this pattern.

> Reaction chains are similar to FAPs, but with one major difference—each set of responses in a reaction chain requires an appropriate stimulus to set it off. Recall that once a fixed-action pattern (FAP) begins, the animal usually continues the sequence even when the stimuli that set off the behavior are removed. In the previous squirrel and nuts example, the animal continues to dig a hole and bury the non-existent nut, even if the nut is removed. In contrast, a reaction chain requires the presence of a specific stimulus to activate each link in the sequence of behavior. An organism's performance produces stimuli that set off the next series of responses in the chain; these behaviors in turn produce the stimuli followed by another set of responses. Presenting a stimulus that ordinarily occurs in the middle part of the sequence activates the chain at that point rather than at the beginning. Also, unlike FAPs, if the stimuli that activate behavior are removed, the sequence is disrupted. The courtship ritual of the male and female stickleback fish (Gasterosteus aculeatus) is a reaction chain (Figure 3.2). Reaction chains often show behavioral flexibility similar to MAPs, but here we describe an idealized behavioral sequence. During the mating season, the reaction chain is initiated when a male stickleback sees a female and begins a zigzag dance, displaying his red underbelly. For a receptive female with eggs, the male's display functions as a stimulus that causes the female to arch her back and reveal her swollen abdomen. This initiates swimming by the male to a previously constructed nest. The female follows and enters the nest with her tail protruding, causing the male to nibble at the base of her tail and initiate the egg-laying sequence. Once the female has deposited her eggs, the male ousts her from the nest, enters the nest, and fertilizes the eggs, thereby completing the reaction chain.

> Reaction chains and fixed-action patterns (FAPs) are alike, but they have a key difference. In a reaction chain, each step needs a specific trigger to start. Once a FAP starts, the animal keeps going even if the trigger is gone. For example, a squirrel will keep burying a nut even if the nut is taken away. But in a reaction chain, each step needs its own trigger to happen. The animal's actions create new triggers for the next steps. If you give a trigger that usually happens in the middle, the chain starts there instead of at the beginning. Unlike FAPs, if you take away the triggers in a reaction chain, the sequence stops.

In this case, a detailed biological description was substantially condensed, yet the core conceptual contrast was preserved. The simplification omits technical elaborations, such as the named species and the full courtship sequence, in favor of a generalized explanation. While contextual richness is reduced, the central distinction between FAPs and reaction chains is retained, and no hallucinated content is introduced. The occurrence of extensive summarization was not an explicitly intended outcome of the simplification method; however, it represents a manageable side effect compared to problematic issues like hallucination or substantive meaning distortion. Should summarization be deemed undesirable, adjustments through targeted prompt engineering could minimize this behavior.

Pedagogically, extensive summarization may offer advantages, particularly for simplifying very long or complex passages. By distilling lengthy selections to core ideas, simplified texts might improve clarity and comprehension, reduce cognitive load, and make complex topics more accessible to learners. The educational value of this approach depends on instructional goals, context, and specific learner needs and warrants further investigation.

3.4.3. Example 3

Very high compression ratios, where the simplified text contains over 50% more words than the original (CR > 1.5), were also analyzed. These cases, while less common than those involving summarization, typically reflect elaboration and explanation, i.e., the expansion of dense or technical content into more accessible language. This process often includes defining key terms, adding context, restating implicit ideas explicitly, or providing illustrative examples.

Although the expanded output introduces additional wording, semantic drift is rare and generally mild when it occurs. The increased length is usually the result of purposeful elaboration rather than hallucination or distortion. In these cases, the simplification trades conciseness for clarity, aiming to enhance comprehension rather than merely reduce complexity.

The following simplification from a Psychology textbook (CR = 2.03, cosine similarity = .64) illustrates this pattern.

> Publication bias is a fundamental problem with our method of creating the official record of scientific inquiry such that studies that report statistically significant findings are more likely to be published than studies that report no statistically significant findings.
>
> When scientists do research, they often look for results that show a clear effect or change. These are called statistically significant findings. Studies with these kinds of results are more likely to get published in scientific journals. On the other hand, studies that don't find a clear effect, known as "negative studies," are less likely to be published. This creates a problem because it means that the official record of scientific research might not include all the important information.

In this example, the simplified text is substantially longer primarily because it elaborates on key concepts to improve clarity, rather than because it introduces new or unrelated information. While the original text concisely employs technical terms such as “publication bias” and “statistically significant findings,” the simplified version uses more accessible language and explicitly defines these concepts. Notably, it introduces the term "negative studies," a concept implied but not explicitly labeled in the original passage. The expanded length results from explaining technical terms in simpler vocabulary and restating ideas clearly to ensure reader comprehension. This illustrates that simplification does not necessarily involve shortening text; rather, it can involve detailed unpacking and explicit elaboration to enhance reader understanding.

While not an intended behavior of the system, such elaboration may serve pedagogical goals when precision and clarity outweigh brevity. Interestingly, although summarization and elaboration may seem like opposing strategies, both can plausibly emerge in response to different forms of complexity. Dense, highly technical language may induce the LLM to elaborate and clarify, while verbose or example-laden passages may lead to condensation. Although neither behavior was explicitly instructed in the prompt, both reflect the LLM’s responsiveness to the local context and illustrate the nuanced, context-sensitive nature of simplification. Whether such behavior is beneficial depends on instructional context and the needs of the student requesting the simplification, a question that invites future empirical investigation.

4. Conclusion

This study demonstrated that LLM-generated simplifications of textbook content can substantially improve readability without sacrificing semantic fidelity. These findings also highlight that textbook content often contains unnecessary lexical and syntactic complexity, as evidenced by students actively selecting these passages and requesting simplification, suggesting they posed challenges to comprehension. Cosine similarity analysis provided empirical evidence that simplified passages reliably retain essential meanings, even when similarity scores are relatively low. Further analysis of cases with low cosine similarity and extreme compression ratios revealed additional simplification strategies, such as summarization and elaborative explanation. Collectively, these results underscore the practical value of targeted simplification—making textbook content more accessible and equitable for struggling students. While this suggests potential value for students, we stop short of claiming comprehension gains, which would require further study.

While the study does not include direct evaluations from students, its purpose was to establish foundational evidence that LLM-based simplifications reduce linguistic complexity while preserving meaning. We view this as a necessary precursor to studies that assess perceived helpfulness or effects on comprehension. Although our analysis relies on automated metrics, these offer scalable and objective indicators of meaning preservation and linguistic change. Nevertheless, automated metrics may not fully capture all nuances of textual understanding, and future work should incorporate complementary human-centered evaluations. Despite these limitations, the current findings provide a rigorous foundation for the continued development of educationally aligned simplification methods.

Several directions warrant further exploration. One priority is investigating how students perceive the quality and helpfulness of simplifications, perhaps through lightweight feedback mechanisms such as the "thumbs up/down" method used in AI-generated question evaluation [27]. These perceptions could provide insights into preferences and trade-offs between simplification and exact meaning retention. Additional work should also explore whether simplification leads to improvements in comprehension or downstream learning outcomes—ultimately the most important goal for educational applications. More broadly, methodological refinements may include identifying optimal thresholds for semantic fidelity and readability improvement linked to enhanced learning. In addition, extending analyses to a broader range of textbooks and subject areas would strengthen generalizability. Future iterations may also explore adaptive supports, including optional inline glosses or tunable elaboration levels, to better serve diverse learner populations, particularly those who struggle with reading complex academic texts.

Acknowledgments

We thank the participating publishers for granting permission to enable generative AI features in their textbook content and permitting the release of the resulting data for open research. We are grateful to Lainey Murdock, Margaret Thompson-Schulz, and Mike Tapply for their support in preparing the open dataset. Finally, we thank the reviewers for their thoughtful and constructive feedback on this paper.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI o3 and GPT-4.5 for: refining draft content; paraphrasing and rewording; grammar and spelling checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Burchfield, C. M., & Sappington, J. (2000). Compliance with required reading assignments. *Teaching of Psychology*, *27*(1), 58–60.

[2] Berry, T., Cook, L., Hill, N., & Stevens, K. (2010). An exploratory analysis of textbook usage and study habits: Misperceptions and barriers to success. *College Teaching, 59*(1), 31–39. https://doi.org/10.1080/87567555.2010.509376

[3] Clump, M. A., Bauer, H., & Bradley, C. (2004). The extent to which psychology students read textbooks: A multiple class analysis of reading across the psychology curriculum. *Journal of Instructional Psychology, 31*(3), 227–232. https://psycnet.apa.org/record/2004-19597-007

[4] Connor-Greene, P. A. (2000). Assessing and promoting student learning: Blurring the line between teaching and testing. *Teaching of Psychology, 27*(2), 84–88. https://doi.org/10.1207/S15328023TOP2702_01

[5] Schneider, A. (2001, May 4). Can plot improve pedagogy? Novel textbooks give it a try. *The Chronicle of Higher Education, 47*(35), A12.

[6] Russell, J.-E., Smith, A. M., George, S., & Damman, B. (2023). Instructional strategies and student eTextbook reading. In *Proceedings of the ACM International Conference* (pp. 613–618). https://doi.org/10.1145/3576050.3576086

[7] Sheridan-Thomas, H. K. (2008). Assisting struggling readers with textbook comprehension. In K. A. Hinchman & H. K. Sheridan-Thomas (Eds.), *Best practices in adolescent literacy instruction* (pp. 164–184). Guilford Press.

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

[9] Färber, M., Aghdam, P., Im, K., Tawfelis, M., & Ghoshal, H. (2025). SimplifyMyText: An LLM-based system for inclusive plain language text simplification. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part IV* (pp. 418–424). Springer. https://doi.org/10.1007/978-3-031-88717-8_32

[10] Fang, D., Qiang, J., Zhu, Y., Yuan, Y., Li, W., & Liu, Y. (2025). *Progressive document-level text simplification via large language models*. arXiv. https://doi.org/10.48550/arXiv.2501.03857

[11] Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

[12] Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge University Press.

[13] OpenAI. (2024, August 8). *GPT-4o system card*. https://openai.com/index/gpt-4o-system-card/

[14] VitalSource Supplemental Data Repository. (2025). https://github.com/vitalsource/data

[15] Book Industry Study Group (2022). *Complete BISAC subject headings list*. https://www.bisg.org/complete-bisac-subject-headings-list

[16] Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy enlisted personnel* (Research Branch Report 8-75). Naval Air Station Memphis: Chief of Naval Technical Training.

[17] Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233. https://doi.org/10.1037/h0057532

[18] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit.* O'Reilly Media.

[19] Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in Python. https://doi.org/10.5281/zenodo.1212303

[20] Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, *23*(1), 84–101.

[21] Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. In *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

[22] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

[23] Agirre, E., Diab, M., Cer, D., & Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM 2012)* (pp. 385–393). Association for Computational Linguistics. https://aclanthology.org/S12-1051/

[24] OpenAI. (2025, April 16). *OpenAI o3 and o4-mini system card*. https://openai.com/index/o3-o4-mini-system-card/

[25] Alva-Manchego, F., Martin, L., Scarton, C., & Specia, L. (2019). EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations (pp. 49–54). Association for Computational Linguistics. https://aclanthology.org/D19-3009.pdf

[26] Schwarzer, M. (2018). *Crowdsourcing text simplification with sentence fusion* [Bachelor's thesis, Pomona College]. https://cs.pomona.edu/classes/cs190/thesis_examples/Schwarzer.18.pdf

[27] Johnson, B. G., Dittel, J., & Van Campenhout, R. (2024). Investigating student ratings with features of automatically generated questions: A large-scale analysis using data from natural learning contexts. In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 194–202). https://doi.org/10.5281/zenodo.12729796