Numeric Information in Elementary School Texts Generated by LLMs vs Human Experts

Anastasia Smirnova*
San Francisco State University
smirnov@sfsu.edu

Erin S. Lee
University of California Berkeley
leoz0113@berkeley.edu

Shiying Li
San Francisco State University
sli63@sfsu.edu

Abstract

LLMs can address long-standing problems in education, such as the lack of instructional materials, by generating grade-appropriate content. We evaluate GPT-4o's ability to generate informational texts for elementary school children. We specifically focus on the model's ability to represent numeric information in text, such as fractions, ratios, and percentages, and assess it with respect to the human baseline. The analysis shows that both humans and GPT-4o reduce numeric information as texts get simplified but do so to a different degree and in a different manner: GPT-4o retains more percentages, while humans use more fractions and ratios. We suggest that these strategies provide different learning opportunities for students.

1 Introduction

Large Language Models (LLMs) have great potential to improve the quality of education (Abdelghani et al., 2024; Han et al., 2024; Yan et al., 2024). They can be used to address long-standing issues in schools, such as shortage of teachers (Edwards et al., 2024) or lack of good instructional materials (Oakes and Saunders, 2004), by generating grade and age-appropriate educational content for students (Scaria et al., 2024; Tan et al., 2025). Diliberti et al. (2024) report that among teachers who employ AI in the classroom, 48% use it to adapt content to the appropriate grade level.

In this paper, we focus on LLMs' ability to adapt informational texts for elementary school children. Informational texts contain quantitative information and expose students to mathematical concepts outside of the math curriculum. The introduction of informational texts in schools in the US was motivated by the demands of the technologically advanced society and the need to develop quantitative literacy (numeracy) in general population (Agnello and Agnello, 2019; Bookman et al., 2008; Steen, 1997, 1999).

Informational texts contain different types of numeric information, as the following passage about paleontological research demonstrates.

Reumer and two colleagues looked in the collections of the Natural History Museum and the Naturalis Biodiversity Center in Leiden, both in the Netherlands, and found 16 samples of mammoth vertebrae from the base of the neck. Seven of the samples were missing the part that would have clued the researchers in on whether a cervical rib had been attached. Of the remaining nine, six were normal and three once had a cervical rib. That worked out to an incidence of 33.3 percent.

Of particular interest here are the last two sentences that allow students to understand how proportions and percentages work.

LLMs' ability to adapt informational text for a specific grade level depends on their mathematical proficiency, their ability to understand quantitative information in text, and to represent it in the form that is appropriate for elementary school children. Mathematical proficiency can be considered an emergent ability in LLMs. McCoy et al. (2024) argue that as a consequence of their design – LLMs were trained to predict next word in text – their performance on tasks that require quantitative skills is sensitive to input probabilities. Thus, GPT-4 performs well on a standard, high-frequency task, such as Celsius-to-Fahrenheit conversion: *multiply by 9/5 and add 32*, but is likely to underperform on a task that has similar complexity but lower input probability: *multiply by 7/5 and add 31*.

Previous work on LLMs' mathematical proficiency returned mixed results. Patel et al. (2023) assessed the ability of GPT-3 model to simplify math word problems for elementary school children. They showed that GPT-3-generated texts are simpler, but noted problems with accuracy. In one instance, GPT-3 inaccurately simplified *she*

*Corresponding author

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 1: Full Papers, pages 183–191

gives each student an eighth of a foot of ribbon as *she gives each student 1 inch of ribbon*. More advanced models perform better – GPT-4 shows 35% improvement in accuracy on math problems compared to GPT-3 (Mishra et al., 2024) – but not at the domain expert level. Mishra et al. (2024) demonstrated that GPT-4o tends to overrely on decimal approximation when working with fractions. Moreover, while GPT-4o showed 90% accuracy on fraction addition tasks, its performance dropped to 61% when instructed to recompute the task with one of the original fractions changed. These results suggest that LLMs' numeric competence is different from human competence (Lee et al., 2024; Lucy et al., 2024).

In what follows, we evaluate LLMs' ability to adapt numeric information in texts for elementary school children. While LLMs' numeric competence is usually assessed on benchmark math tests, we focus on LLMs' ability to convey numeric information in the context of text simplification. Text simplification involves the reduction of structural and lexical complexity of a text, while maintaining its meaning (Shardlow, 2014; Siddharthan, 2014). It is a promising technique for generating age- and grade-appropriate materials with LLMs (Patel et al., 2023).

Previous work on LLMs' ability to generate grade-appropriate content by means of text simplification mostly focuses on the overall readability metrics (Murgia et al., 2023; Patel et al., 2023) or lexical features (Valentini et al., 2023). In these studies, the complexity of a text is operationalized in terms of average word and sentence lengths (shallow textual features), as well as lexical and syntactic features. Simplified texts have shorter words and sentences, more concrete, age-appropriate vocabulary, and simple clauses. These measures, however, do not evaluate how numeric information is conveyed. Similarly to linguistic information, numeric information can be conveyed at different levels of complexity. Proportions, for example, can be represented as fractions, ratios, and percentages (Power and Williams, 2011), and the choice of representation has implications for comprehension and understanding (Bautista et al., 2011). Since numeric information is not included in the standard text complexity measures, little is known about how well LLMs can simplify numeric information in texts.

Our study addresses this gap by evaluating LLMs' ability to convey numeric information in texts and assessing their performance vis-à-vis human experts. We chose to evaluate a particular LLM, GPT-4o by OpenAI, one of the most advanced models at the time of writing. This choice is motivated by GPT-4o's superior performance on math tasks in comparison to other models (Lucy et al., 2024; Mishra et al., 2024) and its widespread use in education.[1]

Our evaluation of GPT-4o and human experts focuses on two questions:

1. How does the amount of numeric information change as texts get simplified?

2. How is difficult mathematical information (proportions) represented in simplified texts?

We report two main findings. First, as texts get simplified, the amount of numeric information is reduced in both human-simplified and GPT-4o-simplified texts. Crucially, GPT-4o reduces numeric information to a greater extent than humans do. Second, we find that humans and GPT-4o use different strategies when simplifying complex information (proportions) with GPT-4o preserving more complex numeric representation (percentages).

2 Study 1: Amount of Numeric Information

Numeric information imposes additional cognitive demands on readers, and thus increases text complexity (Agnello, 2021). We expect that the amount of numeric information will decrease as texts are adapted for lower grade levels.

2.1 Data

Our texts come from Newsela, a provider of educational materials for K-12 curriculum. The Newsela corpus (Xu et al., 2015) is a parallel corpus of informational texts, consisting of the original texts and the corresponding human-simplified texts. There are 5 levels of text complexity within the corpus, from the most complex (level 0) to the least complex texts (level 4). As Figure 1 shows, the distribution of texts by grade and complexity levels is not uniform, with the two largest subsets being texts for grade 12, level 0 (the most complex texts) and texts for grade 4, level 4 (the most simplified texts). These are the two grade levels that we choose to analyze in our study.

[1] https://openai.com/index/introducing-chatgpt-edu/

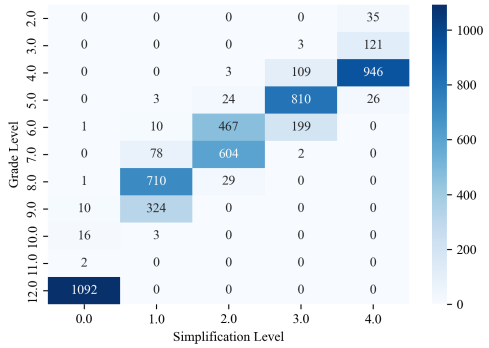

Figure 1: Distribution of texts by grade and simplification level in Newsela corpus.

2.2 Constructing Three Corpora

Our analysis is based on a subset of the Newsela corpus (original and human-simplified texts) and the corresponding GPT-4o-generated texts. For the original corpus, we randomly selected 100 Newsela texts for grade 12, level 0. To construct the human-simplified corpus, we matched these original texts with the corresponding 100 simplified texts for grade 4, level 4. We generated the corpus of GPT-4o-simplified texts by submitting 100 original texts as input to GPT-4o model with instructions to simplify.

We used OpenAI API (model = GPT-4o, temperature = 1) with zero-shot prompting strategy. The prompt was designed to match the style (informational texts), grade level (grade 4), and the average length of texts in human-simplified corpus. Thus, since the average number of words in human-simplified texts for grade 4, level 4 was 680, this length requirement was specified as part of the prompt. The prompt was formulated as follows: "In approximately 680 words, simplify the text below for a fourth-grade reading level written in the newspaper genre."

We did not specifically instruct the model to simplify numeric information. This choice is motivated by the consideration to keep instructions for LLMs and humans as similar as possible (Lampinen, 2024). Since human experts in Newsela use readability scores (Lexile) to guide their simplification process (Agnello, 2021), and since these scores do not take numeric complexity into account, numeric complexity was not referenced in the prompt to GPT-4o. Thus, neither human experts nor the model are specifically instructed to simplify numeric information.

We manually examined GPT-4o-simplified texts for hallucinations and found none. Moreover, our analyses showed that GPT-4o can adequately reduce textual complexity for a specific grade level (Smirnova et al., 2025).

2.3 Extracting Numeric Expressions

Our definition of numeric expressions is based on Agnello (2021). Numeric expressions include counts and measures, arithmetic operations, fractions, percentages, ranges, and others.[2] To extract numeric expressions from texts, we designed a regular expression-based pipeline. Texts were lightly preprocessed to normalize special characters and whitespace, while pattern matching was performed case-insensitively to maximize coverage. Regular expressions were then applied to the preprocessed texts, and sentences containing numeric matches were extracted using rule-based splitting. The results were recorded in three output files for original, human-simplified, and GPT-4o-simplified texts.

2.4 Results

We computed the average number of numeric expressions in the three corpora. The original texts (M=23.55, SD=14.42) contain more numeric expressions compared to both human-simplified (M=12.36, SD=7.15) and GPT-4o-simplified texts (M=9.45, SD=7.10) (see Figure 2). The difference in the distribution of numeric expressions in original and human-simplified texts was statistically significant on a paired t-test ($t(99)=9.042$, $p<0.00001$), and so was the difference between original and GPT-4o-simplified texts ($t(99)=11.698$, $p < 0.00001$). Importantly, the difference between GPT-4o-simplified and human-simplified texts was also statistically significant ($t(99)=4.097$, $p=0.0001$).

Fewer numeric expressions in GPT-4o-simplified texts might suggest that these texts are simpler compared to the corresponding human-simplified texts. However, the number of numeric expressions alone is not sufficient to address this question. In Study 2 we analyze how different numeric expression types are distributed in simplified texts.

[2] See https://github.com/sub-mit/numeracy for the full list of numeric expression types, the corresponding regular expressions and the examples of sentences they match.

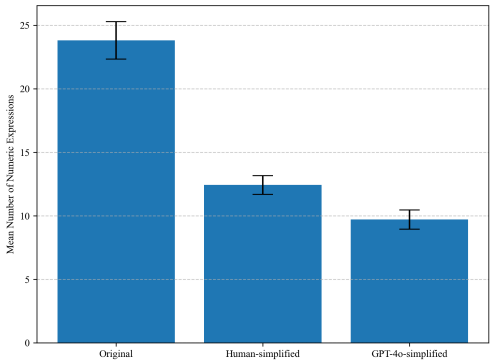

Figure 2: Numeric expression means for original, human-simplified, and GPT-4o-simplified texts. Error bars are +/-1 standard error.

3 Study 2: Complexity of Numeric Expressions

In this study we analyze how proportions are expressed in texts. Proportions can be represented as percentages (*25 percent*), fractions (*one-fourth*), and ratios (*one in four*) (Power and Williams, 2011). The analyses of educational materials and studies with human participants showed that percentages are the most complex numeric expression type (Bautista et al., 2011; Power and Williams, 2011; Wu, 2011). As texts become simplified, we expect that the percentages in the original text will be replaced with other expressions that can convey proportions.

3.1 Percentages in Original Texts

From the list of sentences with numeric expressions in original texts (Study 1), we extracted all sentences that mention "percent". There was a total of 120 unique sentences (types) from 47 texts. Several sentences contained multiple numeric expressions with "percent" in them. Each such expression was treated as an independent token. As a result, we ended up with a total of 147 token sentences referencing percentages.

3.2 Passage Alignment

In order to analyze how percentages from the original texts were represented in the corresponding human-simplified and GPT-4o-simplified texts, we implemented fine-tuned Neural Conditional Random Field (CRF) passage alignment algorithm by Jiang et al. (2020). The alignment is based on the similarity score between passages.

The Neural CRF model is specifically designed for sentence alignment tasks in text simplification (Jiang et al., 2020). It employs a linear-chain CRF integrated with neural network components to align complex and simplified text pairs. The alignment sequence is determined by combining semantic similarity scores derived from fine-tuned BERT embeddings with transition features that reflect sentence order within parallel documents. We select this model for aligning our Newsela corpora because it was trained on Newsela-style datasets, and the authors have provided a version specifically fine-tuned for Newsela sentence alignment.

Our alignment process consisted of two steps described below, data preprocessing and the identification of the most similar passage.

3.2.1 Data Preprocessing

Newsela passages are smaller than paragraphs; they contain one or more sentences. To each Newsela passage we assigned a unique passage_id, formatted as follows: `{corpus_type}__{slug}__{language}__{n_passage}`.

- corpus_type:
  - source_files_grade12: the corpus of original source texts;
  - human_simplified_grade4: human-simplified texts for grade 4;
  - llm_simplified_grade4: the corpus of GPT-4o-simplified texts;
- slug: the slug of the article from which the passage is extracted, e.g. afghan-taxidriver;
- language: the language of the passage;
- n_passage: if the passage_a is the nth passage in the article, then n_passage of passage_a is n.

3.2.2 Getting The Most Similar Passage

The program compares each passage in the original text corpus with every candidate passage in the simplified corpora using the passage-to-passage_id map which provides information about the article. We followed the notebook provided by Jiang et al. (2020) to build our own passage alignment pipeline and used their pre-trained fine-tuned Newsela sentence alignment model for our tasks. For each article within a simplified corpus, we identified and

Text type	Text passage	Numeric Type
Original text	Halloween is crucial to the company, accounting for **25 percent** of Party City's $1.6 billion in annual retail sales.	Percentages
Human-simplified	Party City is a big retailer. It has many stores. Halloween is very important to the company. **One-quarter** of its sales each year come from Halloween.	Fraction
GPT-4o-simplified	Party City is another big store. It sells Halloween items in regular stores and special Halloween City stores. Halloween makes up **25%** of Party City's sales.	Percentages

Table 1: Types of numeric information in aligned passages from original, human-simplified, and GPT-4o-simplified texts.

selected the passage that had the highest similarity score as the aligned passage. The results were recorded as an aligned triplet of original – GPT-4o-simplified – human-simplified passages with similarity scores for subsequent analysis. See Appendix A for an example of aligned triplet with similarity scores.

This implementation resulted in 147 aligned triplets across three conditions (total of 441 passages, i.e. 147 x 3).

3.3 Qualitative Analysis and Coding

We manually examined all aligned triplets. This allowed us to assess how accurately GPT-4o conveys numeric information as well as alignment accuracy. We did not find any numerical inaccuracies in GPT-4o-generated texts, but we did find mismatches in alignment. In cases of content mismatch within a triplet, we consulted full texts side-by-side and searched them for a better candidate to replace the mismatched passage in either human-simplified or GPT-4o-simplified texts. We ended up replacing 53 passages (31 passage replacements in human-simplified and 22 passage replacements in GPT-4o-simplified texts).

We manually coded how numeric expressions referencing percentages were represented in simplified texts. Based on the previous literature (Agnello, 2021; Bautista et al., 2011), we developed a coding system consisting of 5 categories: Percentages, Ratio, Fraction, Non-numeric word, and Dropped. Dropped means that the information was absent in simplified texts. Non-numeric words, such as quantifiers "few" and "many" convey information non-numerically. Of the remaining categories, fraction is the least difficult numeric expression. Ratio can be viewed as a complex fraction (Wu, 2011) but it is less complex than percentages.

Num Type	Humans	GPT-4o
Percentages	4	30
Ratio	12	9
Fraction	18	1
Non-numeric	30	32
Dropped	83	75
Total	147	147

Table 2: Representation of percentages in human-simplified and GPT-4o-simplified texts.

Percentages are the most sophisticated way to represent proportions (Bautista et al., 2011). Table 1 presents an example of aligned passages and the codes for numeric expressions.

3.4 Results: Complex Numeric Types

Chi-square test shows that there are statistically significant differences between human-simplified and GPT-4o-simplified texts in terms of the types of numeric expressions used to represent percentages ($p < 0.00001$, $\chi^2(16)=75.5$). The agreement in the choice of numeric expressions between human-simplified and GPT-4o-simplified texts is 53%. Table 2 shows the distribution by type. Both GPT-4o and humans drop a substantial number of numeric expressions with percentages. When this information is preserved, GPT-4o tends to retain the same numeric type, percentages, while humans tend to use fractions and ratios.

4 Conclusion

In this study we evaluated GPT-4o's ability to convey numeric information in texts simplified for elementary school children and compared its performance vis-à-vis human experts. Study 1 showed that both humans and GPT-4o reduce the number

of numeric expressions as they simplify texts, but GPT-4o does so to a greater extent. Since numeric expressions increase text complexity, these results might suggest that GPT-4o-simplified texts are less complex. Study 2 showed that GPT-4o-simplified texts retain percentages, the most difficult numeric type, to a greater extent than human-simplified texts do.

Is GPT-4o's strategy less effective? While linguistically and numerically difficult texts can present a challenge for the reader, they can also provide a unique learning opportunity.The analysis of GPT-4o-generated passages shows that percentages are presented in a way that is easy for a child to understand. Specifically, these texts make the relationship between numbers and percentages transparent: *Some samples were missing parts, but of the nine they could study, three had a cervical rib. This means around 33% of the mammoths had these extra ribs.* From this perspective, retention of complex numeric expressions in GPT-4o-generated texts can be viewed as a learning opportunity, fostering the development of numeracy in elementary school children. At the same time, simplifications that avoid difficult content might ultimately slow down learners' progress (Crossley et al., 2014).

5 Limitations

There are several limitations that arise from the novelty and complexity of the phenomenon under consideration. First, while our choice of GPT-4o model is motivated by its capabilities and widespread application in educational context, it is not clear whether these results will generalize to other LLMs.

Second, we analyzed representation of numeric information in a context of a general text simplification task, but we did not discuss how numeric complexity is related to linguistic complexity. Just as numeric and linguistic features interact in word problem tasks (Daroczy et al., 2015, 2025), linguistic factors can affect representation of numeric information in educational texts.

Finally, we operationalized numeric complexity as (i) the amount of numeric information and (ii) the type of numeric information in texts. While these are standard measures of numeric complexity (Agnello, 2021; Bautista et al., 2011), there are limitations to the approach that is based solely on the distributional frequency of numeric expressions in text. Text comprehension by the intended end users, elementary school children, can serve as additional evaluation metrics for assessing the quality of informational texts generated by LLMs. A comprehension study can directly compare different numeric simplification strategies, contrasting texts that retain complex numeric types (percentages) with texts that represent the same information as fractions or non-numeric words.

References

Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2024. Gpt-3-driven pedagogical agents to train children's curious question-asking skills. *International Journal of Artificial Intelligence in Education*, 34(2):483–518.

Ellen C Agnello. 2021. Simplified but not the same: Tracing numeracy events through manually simplified newsela articles. *Numeracy*, 14(2):1–20.

Ellen C Agnello and Kevin M Agnello. 2019. Crossing the final frontier: Exploring the numeracy demands of texts read in english language arts. *Numeracy*, 12(2):7.

Susana Bautista, Raquel Hervás, Pablo Gervás, Richard Power, and Sandra Williams. 2011. How to make numerical information accessible: Experimental identification of simplification strategies. In *Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13*, pages 57–64. Springer.

Jack Bookman, Susan L Ganter, and Rick Morgan. 2008. Developing assessment methodologies for quantitative literacy: A formative study. *The American Mathematical Monthly*, 115(10):911–929.

Scott A Crossley, Hae Sung Yang, and Danielle S McNamara. 2014. What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.

Gabriella Daroczy, Christina Artemenko, Magdalena Wolska, Detmar Meurers, and Hans-Christoph Nuerk. 2025. Are text comprehension and calculation processes in word problem solving sequential or interactive? an eye-tracking study in children. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*.

Gabriella Daroczy, Magdalena Wolska, Walt Detmar Meurers, and Hans-Christoph Nuerk. 2015. Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology*, 6:348.

Melissa Diliberti, Heather L Schwartz, Sy Doan, Anna K Shapiro, Lydia Rainey, and Robin J Lake.

2024. *Using Artificial Intelligence Tools in K-12 Classrooms*. RAND.

Danielle Sanderson Edwards, Matthew A Kraft, Alvin Christian, and Christopher A Candelaria. 2024. Teacher shortages: A framework for understanding and predicting vacancies. *Educational Evaluation and Policy Analysis*.

Ariel Han, Xiaofei Zhou, Zhenyao Cai, Shenshen Han, Richard Ko, Seth Corrigan, and Kylie A Peppler. 2024. Teachers, parents, and students' perspectives on integrating generative ai into elementary literacy education. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics.

Andrew Lampinen. 2024. Can language models handle recursively nested grammatical structures? a case study on comparing models and humans. *Computational Linguistics*, 50(4):1441–1476.

Unggi Lee, Youngin Kim, Sangyun Lee, Jaehyeon Park, Jin Mun, Eunseo Lee, Hyeoncheol Kim, Cheolil Lim, and Yun Joo Yoo. 2024. Can we use gpt-4 as a mathematics evaluator in education?: Exploring the efficacy and limitation of llm-based automatic assessment system for open-ended mathematics question. *International Journal of Artificial Intelligence in Education*, pages 1–37.

Li Lucy, Tal August, Rose E Wang, Luca Soldaini, Courtney Allison, and Kyle Lo. 2024. MathFish: Evaluating language model math reasoning via grounding in educational curricula. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5644–5673, Miami, Florida, USA. Association for Computational Linguistics.

Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Shubhra Mishra, Gabriel Poesia, Belinda Mo, and Noah D. Goodman. 2024. Mathcamps: Fine-grained synthesis of mathematical problems from human curricula. *Preprint*, arXiv:2407.00900.

Emiliana Murgia, Zahra Abbasiantaeb, Mohammad Aliannejadi, Theo Huibers, Monica Landoni, and Maria Soledad Pera. 2023. Chatgpt in the classroom: A preliminary exploration on the feasibility of adapting chatgpt to support children's information discovery. In *UMAP '23 Adjunct: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '23 Adjunct, page 22–27, New York, NY, USA. Association for Computing Machinery.

Jeannie Oakes and Marisa Saunders. 2004. Education's most basic tools: Access to textbooks and instructional materials in california's public schools. *Teachers College Record*, 106(10):1967–1988.

Nirmal Patel, Pooja Nagpal, Tirth Shah, Aditya Sharma, Shrey Malvi, and Derek Lomas. 2023. Improving mathematics assessment readability: Do large language models help? *Journal of Computer Assisted Learning*, 39(3):804–822.

Richard Power and Sandra Williams. 2011. Generating numerical approximations. *Computational Linguistics*, 38(1):113–134.

Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. How good are Modern LLMs in generating relevant and high-quality questions at different bloom's skill levels for Indian high school social science curriculum? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 1–10, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Anastasia Smirnova, Kyu beom Chun, Wil Louis Rothman, and Siyona Sarma. 2025. Text simplification for children: Evaluating llms vis-à-vis human experts. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Lynn Arthur Steen. 1997. *Why numbers count: Quantitative literacy for tomorrow's America*. New York: College Entrance Examination Board.

Lynn Arthur Steen. 1999. Numeracy: The new literacy for a data-drenched society. *Educational Leadership*, 57:8–13.

Kehui Tan, Jiayang Yao, tianqi pang, Chenyou Fan, and Yu Song. 2025. Elf: Educational llm framework of improving and evaluating ai generated content for classroom teaching. *J. Data and Information Quality*.

Maria Valentini, Jennifer Weber, Jesus Salcido, Téa Wright, Eliana Colunga, and Katharina von der Wense. 2023. On the automatic generation and simplification of children's stories. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3598, Singapore. Association for Computational Linguistics.

Hongxi Wu. 2011. *Understanding Numbers in Elementary School Mathematics*. American Mathematical Society.

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

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

A Appendix: Passage Alignment

The program compares each passage in the original text corpus with every candidate passage in the simplified corpora based on the similarity scores. Table 3 presents an example of an aligned triplet with similarity scores.

Slug	Original passage	Human-simplified passage	Human-simplified: Similarity Score	GPT-4o-simplified passage	GPT-4o-simplified: Similarity Score
predatoryfish-decline	The removal of top predators has been called "humankind's most pervasive influence on nature," and it is as detrimental in the sea as it is on land. Consumers prefer predatory fish like grouper, tuna, swordfish and sharks to species lower on the food chain such as anchovies and sardines, providing strong incentives for fishermen to catch the bigger fish. Going after the more valuable predators first, fishing them until there aren't enough left to support a fishery and then moving on to species lower in the food chain, a pattern sometimes observed in global fisheries, has been called "fishing down the food web."	The result is something called "fishing down the food web." Fishermen go after the more valuable predators first. They fish them until there aren't enough left. Then they move on to smaller fish that are lower on the food chain. The bigger fish start to disappear	0.9997756	Overfishing big, important fish in the sea is causing trouble. People like to eat big fish like tuna, swordfish, and sharks. These are called predatory fish because they eat smaller fish. Because people want to eat these fish, fishermen catch a lot of them. Once there aren't many big fish left, they move on to catching smaller fish like anchovies and sardines. This is known as "fishing down the food web."	0.9971335

Table 3: Example of passage alignment and similarity scores.