International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-024-00420-2

ARTICLE

Rewriting Content with GPT-4 to Support Emerging Readers in Adaptive Mathematics Software

Kole A. Norberg[1] · **Husni Almoubayyed**[1] · **Logan De Ley**[1] · **April Murphy**[1] · **Kyle Weldon**[1] · **Steve Ritter**[1]

Accepted: 7 July 2024
© International Artificial Intelligence in Education Society 2024

Abstract

Large language models (LLMs) offer an opportunity to make large-scale changes to educational content that would otherwise be too costly to implement. The work here highlights how LLMs (in particular GPT-4) can be prompted to revise educational math content ready for large scale deployment in real-world learning environments. We tested the ability of LLMs to improve the readability of math word problems and then looked at how these readability improvements impacted learners, especially those identified as emerging readers. Working with math word problems in the context of an intelligent tutoring system (i.e., MATHia by Carnegie Learning, Inc), we developed an automated process that can rewrite thousands of problems in a fraction of the time required for manual revision. GPT-4 was able to produce revisions with improved scores on common readability metrics. However, when we examined student learning outcomes, the problems revised by GPT-4 showed mixed results. In general, students were more likely to achieve mastery of the concepts when working with problems revised by GPT-4 as compared to the original, non-revised problems, but this benefit was not consistent across all content areas. Further complicating this finding, students had higher error rates on GPT-4 revised problems in some content areas and lower error rates in others. These findings highlight the potential of LLMs for making large-scale improvements to math word problems but also the importance of additional nuanced study to understand how the readability of math word problems affects learning.

Keywords Large Language Model · Artificial Intelligence · Intelligent Tutoring System · Personalized Learning · Readability

✉ Kole A. Norberg
knorberg1@carnegielearning.com

[1] Carnegie Learning, Inc, 4 Smithfield Street, Floor 8, Pittsburgh, PA 15222, USA

Published online: 19 July 2024

Rewriting Content with GPT-4 to Support Emerging Readers in Adaptive Mathematics Software

Recent advancements in Large Language Models (LLMs) have demonstrated impressive performance across a variety of writing tasks (Ali et al., 2023, Anthropic, 2023, Chen et al., 2021, Gomez-Rodriquez & Williams 2023, Mugaanyi et al., 2024, OpenAI, 2023). The current work applied this technology to adjusting the reading level of educational materials to meet the needs of a key student demographic–emerging readers (i.e., students whose reading comprehension skills lag behind their peers; see Participants section for a description of how emerging readers were identified for the current study). In a recent study, human authors followed a style guide focused on clarity and specificity, to revise a set of math word problems from Carnegie Learning's MATHia adaptive learning software (Almoubayyed et al., 2023a). A randomized experiment showed that emerging readers who received the revised content spent less time completing the problems and achieved higher rates of mastery compared to emerging readers who received the original content. LLMs offer a scalable solution to expanding the revision process but given the tendency for LLMs to introduce inaccurate information when revising text (Arbel & Becher, 2023; Butler et al., 2024; Huang et al., 2024), it is important to test the extent to which LLM revisions drive improvements in readability and performance across a range of content.

As a first step in this process, we previously prompted ChatGPT-4 (March 2023 version) to revise the same set of word problems using the same style guide used during the human revision process (Norberg et al., 2023). ChatGPT-4 successfully produced revised problems with significantly improved readability – sometimes outperforming the human-revised problems on common readability metrics (e.g., Flesch-Kincaid Grade Level [FKGL], Kincaid et al., 1975), as well as newer metrics (e.g., the modified Crowdsourced Algorithm of Reading Comprehension [CAREC-M], Crossley et al., 2019, and Sentence-BERT [SBERT], Crossley et al., 2023, Reimers & Gurevych, 2019). This was a promising start, but improvements to readability scores do not always translate to improvements to reading comprehension (e.g., McNamara et al., 1996; O'reilly & McNamara, 2007; Ozuru et al., 2009). Thus, the current study extended prior work by (a) assessing the effect of texts revised by the LLM on student performance in MATHia and (b) expanding the prompt to produce revisions for a broader range of problems to understand how effects seen in one content domain generalize to others. We expected the LLM revised problems to have a similar effect on student performance in MATHia as the human rewrites in Almoubayyed et al. (2023a). That is to say we expect the revised problems to improve performance specifically for emerging readers.

Relationship Between Reading and Math Learning

A central motivation to revise math word problems is the well-established relationship between reading skill and performance in math (e.g., Almoubayyed et al., 2023c; Daroczy et al., 2015; Greisen et al., 2021; Helwig et al., 1999;

Koedinger & Nathan, 2004). The relationship may in part be causal. Struggling with or failing to accurately decode the text of the problem can interfere with numerical processing and impair a student's ability to integrate numerical information with the larger text (Daroczy et al., 2015; Fuchs et al., 2006; Fuchs et al., 2018). This can lead to a kind of Matthew effect (Merton, 1968) wherein typical readers also advance in mathematics at a faster rate than emerging readers.

Math word problems can pose specific challenges for reading comprehension. For example, (1) is the introduction to a math problem meant to be solved by middle school students.

> (1) Michelle is walking her two immense dogs. At the same time the dogs see two squirrels on opposite sides of the street. Michelle sits down to try to hold the dogs, but it does not work. First the larger dog pulls Michelle so hard that she is thrown 7 feet backwards. She lets go of the leash of the larger dog, and then the smaller dog drags her forwards at a rate of 2.4 feet per second. Let the distance backward from where Michelle sits down be negative and the distance forward be positive.

At a basic level, this problem features some vocabulary that may be difficult for emerging readers at this grade level (e.g., *immense*) and some mathematical phrasing (e.g., *Let*) that may be unfamiliar to a middle school math learner. As with many math word problems, it also contains temporal and spatial adverbs (*First*, *Then*, *backwards*, *forwards*). Although these phrases are critical to understanding the relationship among the sentences, they can also slow down processing (Bestgen & Vonk, 2000; Hoeks et al., 2004; Zwaan, 1996) especially for younger readers (Cain & Nash, 2011), and generally require the reader to hold more information active for longer during the text integration process (Millis & Just, 1994). Thus, even before adding in the requirement that students track mathematical components of the text or perform computations, word problems can present challenges to student comprehension.

Unfortunately, reading difficulties are widespread. In 2022, 69% of 8th graders were categorized as non-proficient readers (National Center for Educational Statistics, 2022). Although students need exposure to challenging text in order to grow their reading abilities (Betts, 1946; Keene & Zimmerman, 1997; Miller, 2002; Morris et al., 2019; Mounla et al., 2011), when a task is too complex, rates of learning can decline and frustration grow (Metcalfe, 2011). Improving the readability of math word problems is one way to ensure that students can focus on learning math without increasing cognitive load related to reading.

The effect of reading comprehension on math learning can be quantified by looking at the correlations between math problem performance and end-of-year English Language Arts (ELA) test scores. Almoubayyed et al. (2023c) looked at the ratio of correlations between outcomes in MATHia and ELA test scores and outcomes in MATHia and math state scores. They found that the ratio was greater than average for some content areas, suggesting that success here was more dependent on reading skills. To support readers in these spaces, a style guide for creating readable, grade-appropriate word problems for middle school students

was developed (Almoubayyed et al., 2023a). The guidance included using short sentences, reducing anaphors and lexical diversity, and simplifying vocabulary. The researchers then revised MATHia word problems to fit the new style guide. Students who received the revised problems had higher rates of mastery over the content and finished the problems up to 30% faster than students who received the original problems. Importantly, the benefits of the revisions were stronger for students identified as emerging readers. These results indicate that math learning can be enhanced by improving the readability of the word problems.

Although improvements to readability generally aid reading comprehension and reduce cognitive load (e.g., Just & Carpenter, 1980), there are circumstances where performance can actually decline as readability increases (e.g., McNamara et al., 1996; O'reilly & McNamara, 2007; Ozuru et al., 2009). This decline in comprehension with improved readability specifically affects students who have high domain knowledge but lower reading ability (O'reilly & McNamara, 2007). It may be that for emerging readers who are less likely to engage in active reading, increasing the ease of processing of a text that already covers a topic familiar to the reader leads to overconfidence in comprehension and further reduces close reading. Indeed, readers are less likely to attend to nonsense words embedded in texts with higher readability and are more likely to display over confidence in their ability to comprehend these texts suggesting that a tendency to engage in active reading is related to how difficult the text *feels* to the reader (Norberg, 2022). Thus, improving the readability of word problems needs to be balanced with analysis of student performance to ensure the revisions are not having a negative effect for any subgroup.

Improving Readability with LLMs

Intelligent tutoring systems (ITS), like MATHia, have thousands of word problems grouped into content areas which we term workspaces. The findings from Almoubayyed et al. (2023c) suggested additional workspaces beyond those revised in Almoubayyed et al. (2023a) would benefit from revisions to improve readability. The human revision process took ~ 45 h to revise 200 problems (30 scenarios).[1] This time included completing the revisions, placing the revisions into the code base, and evaluating them for quality assurance both before and after the problems were added to the code base. To make readability improvements more scalable, we tested the ability of LLMs to complete the revisions.

LLMs are being tested as an option to reduce the human labor required for improving text readability in multiple domains, including health care, law, and education (Arbel & Becher, 2023; Butler et al., 2024; Huang et al., 2024). In each study,

[1] MATHia word problems are created from scenarios which provide the bulk of the text for the problem. A problem generator then swaps out substitutable components of the problem (e.g., pronouns, names, objects, and numbers) to create multiple versions of the scenario. 30 scenarios in this case created 200 problems, 100 in each of two workspaces. Students do not see more than one version of each scenario in a single workspace.

LLMs were consistently able to improve the readability of the texts according to common readability benchmarks (e.g., Flesch-Kincaid Grade Level [FKGL]; Kincaid et al., 1975) which can assess vocabulary difficulty, sentence complexity, cohesion, or their combinations. Despite the improvements in readability metrics, each study also cited concerns about the LLMs injecting false information into the narrative and the importance of keeping a human-in-the-loop to evaluate revisions.

Results from these prior studies were in line with our prior findings. Norberg et al. (2023) prompted ChatGPT-4 (March 2023 version) to revise the same word problems revised by humans in Almoubayyed et al. (2023a). The ChatGPT-4 revisions were significantly improved over the original problems on a metric assessing syntactic and lexical diversity as well as cohesion. As in other studies, not all of ChatGPT-4's revisions were viable for use in MATHia. In side-by-side comparisons with the original content, human reviewers rejected 13% of the revisions as missing critical information or having other errors which they perceived as reducing their ability to comprehend the problem. These problems were revised again by the LLM until they passed human inspection. Even with this rate of rejection, the LLM revisions reduced time to revise the word problems by more than half. However, the process still took 19 h with the bulk of that time spent incorporating the problems into the MATHia code base though human evaluations of the problem and time spent prompting ChatGPT-4 through its online interface were also a factor. In the current study we overview steps we took to further reduce the time in human labor.

Critically, to our knowledge, studies using LLMs to improve text readability (e.g., Arbel & Becher, 2023; Butler et al., 2024; Huang et al., 2024) have only evaluated the text itself, not how well it enhances comprehension or performance. In the current study, we go beyond readability metrics and human qualitative evaluations to assess the impact of LLM-based content revisions on student performance. We further test the generalizability of such findings by performing additional revisions on new sets of word problems and evaluating student performance in these new workspaces.

Current Study

The current study had two primary objectives. The first was to test the ability of the LLM to perform revisions with greater automaticity across a wider variety of problems while still maintaining the same improvements in readability. Generalizing to new workspaces and increasing automaticity required the formation of a new prompt, a process we detail in the *Materials* section. Broadly, the new prompting process required the LLM to revise the problems in the same format as they appeared in the code. This greatly reduced human labor related to encoding the problems and was critical to creating a scalable solution.

The second objective of the study was to assess how the LLM revisions affected student performance in MATHia and whether the effects were similar to those found for the human revisions in Almoubayyed et al. (2023a). We expected similar results for all revised workspaces. However, improvements in readability may not translate to improvements in comprehension or math performance, and if they do, they may

not demonstrate effects equally across all types of problems or for all reading levels. Thus, we performed the tests across four workspaces to assess how well improvements to readability generalized across different content domains.

Method

The ultimate goal of the LLM generated problems was to improve student math performance. Thus, the primary experimental task in the study assessed student learning by evaluating student error and mastery rates as well as their time to completion and how many problems the student had to solve before reaching completion. Before placing LLM generated problems in front of students, we also ensured the validity of the rewrites by assessing their readability using readability metrics and having human evaluators accept or reject the revisions. We report the reliability of the LLM for improving readability metrics, the human evaluators' rejection rate, and the effect of the LLM revisions on student outcomes.

Participants

MATHia is currently used by over 600,000 students in the United States, mostly students in grades 6–12 who use it as part of Carnegie Learning's blended math curriculum. Students who started any of the target learning domains (i.e., MATHia workspaces) between August 15th and December 22nd, 2023 and completed the workspace prior to the time of analysis (January 30, 2024) were included in the study ($n = 83{,}082$).[2] Study participants were randomly assigned to receive either the control (pre-existing MATHia problems) or problems rewritten by GPT-4.

This work is motivated by helping emerging readers to access math content, so it was important to evaluate how this subpopulation responded to the revisions. Participant's reading ability was determined using a deep learning model trained on student performance (i.e., errors and hint usage) in an introductory MATHia workspace which does not include math (Almoubayyed et al., 2023c). The model consists of a neural network that has been shown to accurately predict end-of-year ELA exam scores, with an area under the receiver operator characteristic curve (AUC) of 0.80. Almoubayyed et al. (2023b) found that the model generalized well to a district in a different state with a different exam (AUC = 0.76). It did not show systemic bias in comparisons based on race or gender, even when tested on a different district than the one on which it was trained. We used this model to evaluate students' predicted reading ability. Following the procedure used in Almoubayyed et al. (2023a), we

[2] Students from two districts participating in a separate study (which used the same materials) or from districts who have opted out of field trials were excluded from enrollment. 12.99% of students had not completed the workspace prior to analysis. There were no systematic differences in completion based on condition (13.06% of students from the control condition and 12.91% of students from the LLM rewrites condition).

categorized students as emerging readers if their predicted reading ability placed them in the bottom quartile of students within a workspace. Within each condition (LLM rewrite or control) the percentage of students categorized as emerging and non-emerging readers was similar, rounding to between 24–25% of students categorized as emerging readers.

Teachers can permit students to bypass the introductory lesson from which predictions about reading ability are made. When this happens, we are not able to predict their reading ability. Thus, students who did not complete the introductory lesson (12.25%) were dropped from analysis. There were no systematic differences in this exclusion across conditions (<0.01% difference). This left 72,905 students available for analysis.

It was possible, depending on the sequence the teacher was using and the students rate of progression through the sequence, that the same student could complete more than one of the target workspaces over the course of the semester. 27.61% of students completed more than one of the target workspaces (18.26% completed 2 workspaces, 9.84% completed 3 workspaces, and fewer than 0.001% completed all 4 workspaces). Due to typical patterns related to when and how often workspaces in MATHia are assigned by teachers, student enrollment across workspaces was unequal (11,362 in *Analyzing Models of Two-Step Linear Equations Integers*, 13,307 in *Analyzing Models of Two-Step Linear Equations Rationals*, 42,471 in *Modeling the Constant of Proportionality*, and 33,565 in *Modeling Linear Relationships Using Multiple Representations*). See Tables 1 and 2 for sample sizes within each cell.

Given the observed effect sizes found by Almoubayyed et al. (2023a) (Cohen's $d=0.15$) and a significance threshold of $\alpha=0.05$, we estimated that a minimum of 972 observations would provide sufficient power ($\beta=0.80$) to detect a difference between the two groups. Thus, all models detailed in the Analytic Plan were well powered.

Materials

MATHia

The revised problems were drawn from and tested in MATHia (formerly Cognitive Tutor, (Ritter et al., 2007)), an ITS developed by Carnegie Learning, Inc. MATHia is typically used for around 40% of the in-classroom instruction and practice time. Students are assigned specific content in MATHia by their teachers to fit with the learning objectives for their math course. Math lessons in MATHia, called "workspaces," can be either "Concept Builders" that teach a math concept, or "Mastery" workspaces that allow the students to practice problems towards mastery of a set of skills or knowledge components (KCs). MATHia uses the student's work on different steps within a problem (correct answers, errors, hint requests) to determine whether the student has mastered the associated KC using Bayesian Knowledge Tracing (Corbett & Anderson, 1994). KCs typically require multiple demonstrations of correct performance (without errors or hint requests) to reach mastery, and students will typically complete between 3 and 25 problems within a workspace to

Table 1 Student Outcomes for Simpler Workspaces

Workspace	Reading Ability	Condition	*n*	Promotion Rate	Errors per Problem	Completed Problems	Time (mins)
Analyzing Models of Two-Step Equations—Integers	Emerging (Quartile 1)	Original Problems	967	0.24 ± 0.43	5.01 ± 2.17	15.9 ± 7.03	74.92
		Human Rewrites	1428	0.18 ± 0.38	4.56 ± 2.27	14.66 ± 6.94	28.48
		GPT-4 Rewrites	1426	0.20 ± 0.40	4.69 ± 2.23	14.95 ± 7.04	30.53
	Typical (Quartiles 2–4)	Original Problems	2920	0.08 ± 0.27	3.58 ± 1.95	11.72 ± 6.06	18.08
		Human Rewrites	4296	0.06 ± 0.24	3.24 ± 1.90	10.90 ± 5.87	18.38
		GPT-4 Rewrites	4212	0.07 ± 0.25	3.20 ± 1.93	10.98 ± 6.04	17.72
Analyzing Models of Two-Step Equations—Rationals	Emerging (Quartile 1)	Original Problems	513	0.19 ± 0.40	4.14 ± 2.20	14.38 ± 7.05	18.97
		Human Rewrites	1692	0.12 ± 0.32	3.54 ± 2.03	12.23 ± 6.69	19.33
		GPT-4 Rewrites	1622	0.12 ± 0.32	3.44 ± 2.00	12.15 ± 6.70	18.31
	Typical (Quartiles 2–4)	Original Problems	1608	0.04 ± 0.21	3.06 ± 1.68	10.37 ± 5.49	11.04
		Human Rewrites	5002	0.03 ± 0.18	2.55 ± 1.58	9.22 ± 5.02	13.17
		GPT-4 Rewrites	4991	0.03 ± 0.18	2.42 ± 1.61	9.26 ± 5.13	13.38

Students labeled emerging readers are in the bottom quartile of predicted reading ability. ± values represent standard deviations from the mean. Median time is reported due to time to completion being right skewed. Statistics from the original problems are based on data that was originally reported on in Almoubayyed et al. (2023a). Means are those before separating based on promotion status.

Table 2 Student Outcomes for Complex Workspaces

Workspace	Reading Ability	Condition	*n*	Promotion Rate	Errors per Problem	Completed Problems	Time (min)
Modeling the Constant of Proportionality	Emerging (Quartile 1)	Originals	5284	0.05 ± 0.22	7.14 ± 4.24	11.77 ± 5.23	79.49
		GPT-4 Rewrites	5260	0.06 ± 0.23	7.33 ± 4.51	11.86 ± 5.37	77.95
	Typical (Quartiles 2–4)	Originals	16,903	0.01 ± 0.11	5.30 ± 3.27	9.77 ± 4.09	54.77
		GPT-4 Rewrites	15,834	0.01 ± 0.12	5.39 ± 3.15	9.93 ± 4.13	54.77
Modeling Linear Relationships Using Multiple Representations	Emerging (Quartile 1)	Originals	4031	0.19 ± 0.47	5.22 ± 3.10	17.53 ± 7.11	100.08
		GPT-4 Rewrites	4080	0.19 ± 0.47	5.40 ± 3.15	17.51 ± 7.11	99.03
	Typical (Quartiles 2–4)	Originals	12,700	0.08 ± 0.38	4.35 ± 2.28	15.08 ± 6.75	71.89
		GPT-4 Rewrites	12,754	0.06 ± 0.37	4.43 ± 2.31	14.75 ± 6.63	69.98

Students labeled emerging readers are in the bottom quartile of predicted reading ability. ± values represent standard deviations from the mean. Median time is reported. Means are those before separating based on promotion status.

reach mastery, depending on their performance. When MATHia determines that the student has mastered all KCs associated with the workspace, it progresses the student to the next workspace. Occasionally, students complete a predetermined number of problems (typically 25) without mastering every skill in a workspace. In those cases, students are also moved on to the next workspace. We call the latter case a "promotion."

Workspaces

Workspaces in this study were selected to meet two criteria: (a) They had high historic usage during fall semesters and (b) the ratio of correlations between the workspace's outcomes and ELA test scores to that of the correlation between the workspace's outcomes and math state scores was greater than average, which is interpreted as more highly related to reading than average (Almoubayyed et al., 2023c). Four workspaces were selected. Two of the workspaces we label *Simpler* and two *Complex* based on differences in their computational complexity and the number of steps required to complete the problem. The two sets of workspaces were similar in readability (see Table 3 in the Results) though the *Complex* workspaces had more words in the primary text of the problem ($m = 75.96$, $sd = 23.62$) compared to the *Simpler* workspaces ($m = 46.5$ words, $sd = 6.55$), $t(231.36) = -15.54$, $p < 0.001$.

Simpler Workspaces The two *Simpler* workspaces, *Analyzing Models of Two-Step Equations Integers* (targets 7th grade) and *Analyzing Models of Two-Step Equations Rationals* (targets 8th grade), required students to recognize how components of the text were reflected in an equation. The only difference between the two workspaces was the type of numbers they employed: The *rationals* workspace included decimal numbers instead of using only integers. To solve the problem, students dragged and dropped the appropriate part of an equation onto its description (see Fig. 1).

The *Simpler* workspaces were selected for this study because prior revisions by humans had successfully improved their readability and reduced the time emerging readers spent mastering the workspace (Almoubayyed et al., 2023a). Following the successful outcomes resulting from the human revisions, they formally replaced the original problems within MATHia prior to the start of the experiment.[3] Thus, for the *Simpler* workspaces, the control problems represent the human rewrites and allow us to directly test the efficacy of LLM rewritten content to that of human rewritten content within the same workspace.

Complex Workspaces The two *Complex* workspaces involve graphing. One of the workspaces, *Modeling the Constant of Proportionality* (part of the 7th & 8th grade sequences), requires students to create tables, construct an

[3] After finding strong effects of improved student performance when receiving the human re-written problems as compared to the original problems (Almoubayyed et al., 2023a), it would have been unethical to continue to present students with the original problems in these workspaces.

Table 3 Readability Metrics

	CAREC-M			SBERT*			NDC			FKGL		
Prompt Length	Original	Human	LLM	Original	Human	LLM	Original	Human	LLM	Original	Human	LLM
ChatGPT-4 Prompt Simpler Work-spaces	0.25 (0.01)	0.24 (0.01)	0.19 (0.01)	-0.21 (0.06)	-0.13 (0.07)	-0.13 (0.08)	8.61 (0.17)	8.41 (0.21)	8.70 (0.17)	5.85 (0.28)	5.51 (0.28)	5.47 (0.33)
GPT-4 Prompt Complex Work-spaces	0.23 (0.01)	NA	0.16 (0.02)	-0.12 (0.03)	NA	-0.03 (0.02)	7.54 (0.16)	NA	6.71 (0.16)	6.10 (0.30)	NA	4.29 (0.25)

Mean and std. errors for readability metric. Lower values indicate greater readability *except for SBERT where higher (here less negative) scores are better.

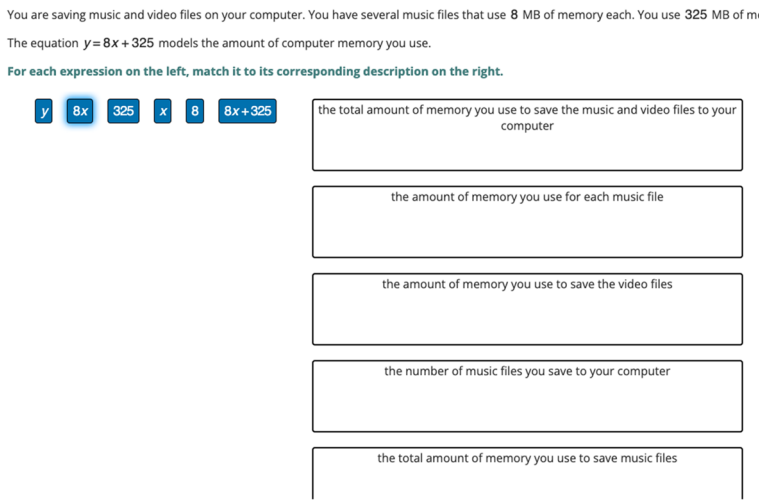

Fig. 1 Example problem from one of the *Simpler* workspaces**.** *Note.* The word problem is from *Analyzing Models of Two-Step Equations Integers* and asks students to recognize how the components of the equation relate to the text. *Analyzing Models of Two-Step Equations Rationals* uses the same word problems and structure but replaces integers with decimal numbers. The problem presented was revised by humans as part of Almoubayyed et al. (2023a)

equation, and plot the values from the table onto a graph (see Fig. 2). The second workspace, *Modeling Linear Relationships Using Multiple Representations* (part of the 8th grade sequence), requires students to use the number properties to evaluate and solve one- and two-step equations (see Fig. 3). The

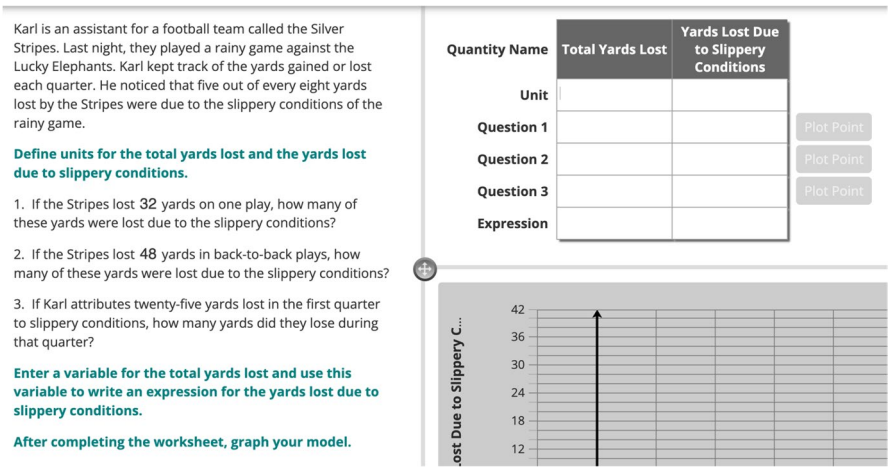

Fig. 2 Example problem from *Complex* workspace *Modeling the Constant of Proportionality***.** *Note.* The word problem asked students to recognize units in the text, construct an equation to answer the questions, and plot the points on the graph. The problem presented is the non-revised version

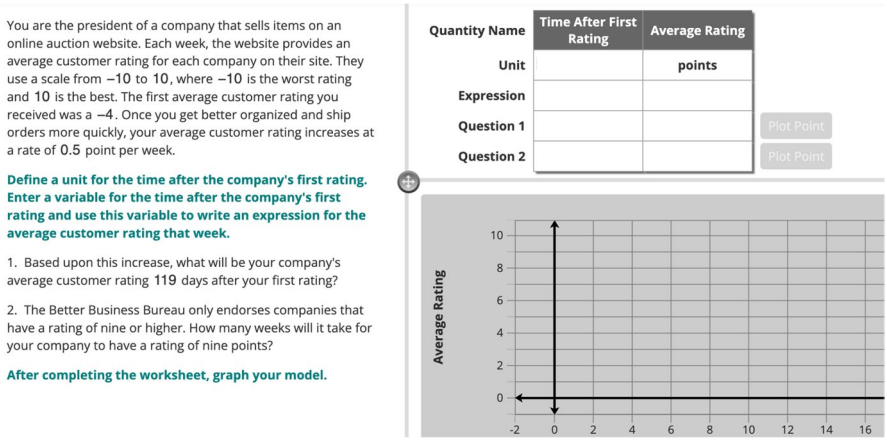

Fig. 3 Example problem from *Complex* workspace *Modeling Linear Relationships Using Multiple Representations*. *Note.* The word problem asked students to recognize units in the text, construct an equation to answer the questions, and plot the points on the graph. The problem presented is the non-revised version

Complex workspaces were chosen because of the ways they differed from the *Simpler* workspaces in the types of steps students were required to complete, allowing us to test the generalizability of effects on student performance. The *Complex* workspaces still had greater correlations to reading than average, but the correlations were weaker than for the *Simpler* workspaces. Indeed, while all steps in the *Simpler* workspaces required students to read for comprehension, half of the steps in the *Complex* workspaces required extracting information from the text while the remaining steps were more strictly mathematical (e.g., plotting points on a graph). Supporting the increased complexity of the new workspaces are measures of time to completion and error rate. Students took on average 54.31 min longer to complete *Complex* as compared to *Simpler* workspaces, $t(39{,}741) = 140.9$, $p < 0.001$. Students also made 1.87 more errors per problem in *Complex* as compared to *Simpler* workspaces, $t(41{,}350) = 109.8$, $p < 0.001$.[4]

LLM Prompts

We used OpenAI's ChatGPT-4 online interface (March 14 version) for revising the *Simpler* workspaces or GPT-4 API (June 13 version) for revisions to the *Complex*

[4] As the same student may have completed more than one workspace, hierarchical linear models of time to completion and error rate included students as a random intercept.

workspace (OpenAI, 2023).[5] The shift in interfaces followed practical considerations related to scaling the workflow and emerging accessibility to the API. The switch to using the API necessitated changes to the prompting style. Below we overview the original prompting technique and then how we modified this technique to create a shorter prompt which was capable of revising a broader range of word problems. Ultimately both techniques resulted in similar improvements in readability of the word problems (see Table 3).

Simpler Workspaces Prompt In Norberg et al. (2023) we distilled the style guide used by human authors to revise word problems in Almoubayyed et al. (2023a) into a chain-of-thought prompt to ChatGPT-4 (accessible on OSF, https://osf.io/xwz3h/). Chain-of-thought prompting provides examples of expected output and explains the reasoning behind the examples (Saravia, 2022; Wei et al., 2022). The prompt led to revisions with significantly improved readability over the original problems as measured by the CAREC-M which assess syntax, lexical diversity, and cohesion, $p < 0.001$.

Despite its success in improving readability, the ChatGPT-4 prompt had a number of practical drawbacks. First, the resulting prompt was long (1,572 words, 2,143 tokens, and requiring approximately 995 output tokens). As a result, ChatGPT-4's response was often truncated requiring a follow up prompt to *continue* before it would finish. At that time, a shorter prompt was required for compatibility with the API. Second, the prompt was workspace specific. In creating a prompt for a new set of workspaces, we sought to create one which would generalize to support revisions to a broader range of problems. Finally, the word problems which were revised by ChatGPT-4 were not embedded within code. Encoding the problems following revision still required significant human labor.

Complex Workspaces Prompt In creating a new prompt which would be compatible with the API for GPT-4, we leveraged a finding from Norberg et al. (2023) that ChatGPT-4 performed better when it was asked to explain a readability concept (e.g., passive versus active voice) and then identify it within the text before attempting a revision. In the new prompt, we asked the LLM to first explain what readability issues existed in the problem, then to explain how it could revise the problem to address these issues, and finally to revise the problem. This method was not as consistently effective as the chain-of-thought approach, but it was more generalizable to new word problems with novel structures. To compensate for the reduced consistency, we implemented a recursive workflow in which we leveraged Python libraries to assess readability and asked GPT-4 to revise again when its revision did not meet the threshold for improvement (one standard deviation improvement on measures of FKGL and SBERT; defined later in Measures and Analytic Plan).

[5] We tested other LLMs but at the time of revisions only GPT-4 was able to retain the correct structure of the problem in revision.

We also changed the prompt to direct GPT-4 to do some of the encoding work, thus reducing the need for a human to encode the problems following revision. Instead of providing the text of the problem as seen by the user, we provided GPT-4 with the text as it looks in the MATHia code base, see (1). This includes tags to identify components of the problem (e.g., < slope >) and placeholder text which ensures variation between problems (e.g., ~*heShe* instead of *she*).

(1) At the beginning of the summer, there are < interceptPhrase > living in a field. < slope > Each < ind > week < /ind > for the rest of the summer, another rabbit moves in < /slope >. 'How many < {dep} > will be living in the field < indep > < indOtherUnitVal > < Other > < /indep > after summer starts?'.

The new prompt presented the LLM with a more challenging task. GPT-4 had to identify issues with readability in a sparse text where key elements were alluded to but not stated directly. Improving readability while still retaining the placeholders and tags would create a truly scalable solution. We supported GPT-4 processes with regex functions to clean up errors and identify anomalies. In 2–6 we outline this workflow.

(2) *Provide initial prompt*: This comprised the problem (with placeholders) and broad instructions for revision. This initial prompt was 483 words (651 tokens) and is provided in Appendix A.
(3) *Calculate readability scores for GPT-4's output*: This comprised Flesch-Kincaid Grade Level (FKGL; Kincaid et al., 1975) and SBERT score (Crossley et al., 2023; Reimers & Gurevych, 2019). These metrics reflect different aspects of readability, with FKGL assessing the complexity of the sentences and vocabulary (in terms of length) and SBERT assessing the semantic similarity of the text to other texts.
(4) *Additional Revisions with GPT-4, as needed*: If the SBERT and FKGL scores were not improved by more than one standard deviation, steps 2 and 3 could be repeated up to three times. In these cases, GPT-4 was provided the initial prompt, its prior output, and a new prompt giving more specific advice on elements of the text to revise, such as reducing vocabulary complexity (187 words, 257 tokens; see Appendix A for the full prompt). Even after three revision attempts, 29% of revisions did not reach the required threshold. In these cases, the revision with the best combined FKGL and SBERT score was accepted.
(5) *Correct GPT-4 errors, as needed*: GPT-4's output often contained minor errors in the structure of the code tags (e.g., < slope > without < /slope >) or duplicate words where the meaning of placeholder text had not been understood (e.g., ~*heShe* followed by *he*). We relied on Python regex functions to identify these issues, fixing them directly where possible or alternatively flagging them for human intervention.
(6) Following the automated revision process, manual human inspection for quality assurance was completed. Reviewers looked at the original problems side-by-side with the revisions and were asked to accept the problem if they believed it did not decrease readability. During human review, problems were sent back for

further LLM revisions at a rate of 21% which was greater than the 13% revision rate in Norberg et al. (2023). Human review of the problems was implemented again after the problems had been uploaded into MATHia. This time, humans answered each question in the revised workspaces to ensure all problems were solvable and could be appropriately tutored with the information provided in the LLM rewrite. This process did not identify any additional errors related to the LLM rewrite. Even with the increased revisions, the new workflow which automated the encoding process reduced human labor to ~ 6 h compared to the 19 h required Simpler Workspace Prompt procedure.

Procedure

Experimental design, condition assignment, and outcome monitoring were conducted using UpGrade, an open-source platform for managing field tests in EdTech software that has been integrated into MATHia (Ritter et al., 2020). Teachers are responsible for assigning students a workspace to complete in MATHia and choose workspaces which match their educational objectives. Teachers are aware of the possibility of ongoing experimental testing in MATHia but do not know which workspaces may currently be part of a study. When a student started one of the target workspaces in this study, UpGrade assigned them to one of the two conditions (control or LLM rewrite) with equal probability. Because the analyses on the *Simpler* workspaces use a different control (i.e., human rewrites as opposed to the original problems) and that content was generated with a different prompting procedure, we separate our comparisons into Experiment A and B.

Measures and Analytic Plan

Readability

We first assessed the readability of the LLM revised problems as compared to the original problems and, in the case of the *Simpler* workspaces, as compared to the human revised problems. The measures we selected each assess unique text characteristics including word frequency, word and sentence complexity, semantic similarity within a text, and the interaction of several of these features. The New Dale-Chall (NDC; Chall & Dale, 1995) is calculated based on sentence length and the percentage of words in a text that may be considered unfamiliar (i.e., are not part of a set of 5,000 pre-selected common words). The NDC was chosen as it evaluates vocabulary difficulty, an area where LLMs have previously struggled (Norberg et al., 2023). FKGL (Kincaid et al., 1975) compares word counts to sentence and syllable counts to determine the appropriate grade level for a text. It is one of the most widely used open-source methods of calculating readability, including for estimating the readability of LLM generated texts (e.g., Arbel & Becher, 2023; Butler et al., 2024). Despite its widespread use, the FKGL has had mixed results in terms of predicting reading comprehension (cf. Crossley et al.,

2017; Duffy, 1985; Zainurrahman et al., 2024). This has led to the development of metrics which consider features related to syntax, lexical diversity, and cohesion of which the Modified Crowdsourced Algorithm of Reading Comprehension (CAREC-M; Crossley et al., 2019) was selected because it was modified to evaluate shorter texts like the ones used in this study and correlates more strongly with reading outcomes than the FKGL (Choi & Crossley, 2022). SBERT (Crossley et al., 2023; Reimers & Gurevych, 2019) is a newer transformer-based deep learning model which assesses the semantic similarity within a text and performs well relative to earlier metrics. Higher SBERT values indicate greater readability. All metrics were evaluated using the Automatic Readability Tool for English (ARTE, Choi & Crossley, 2022). Significance comparisons were conducted using t-tests to compare LLM rewrites to the original problems for all four outcomes.

Performance Outcomes

All models included contrast coded fixed effects of condition (LLM rewrites compared to control), student reading level (emerging compared to non-emerging), and their interaction. When a significant interaction was present, pairwise contrasts were performed to assess simple effects with a Tukey correction applied to correct for multiple comparisons. Because students were nested within schools, we also included a random effect of school. We used the maximal random effects structure supported by the data (Matuschek et al., 2017). The structure used for each model is reported in the Tables in Appendix B. Emerging readers were identified based on the model described above in *Participants*.

We considered four outcome measures also assessed by Almoubayyed et al. (2023a): *Promotion Rate*, *Errors per Problem*, *Total Problems Completed*, and *Time to Completion*. Promotion rate and time to master a workspace in MATHia have reliably predicted end-of-year state math test scores (Zheng et al., 2019). Thus, these measures offer a well-rounded assessment of student performance in a workspace.

Promotion Rate Promotion rate refers to the percentage of students who failed to master at least one skill within a workspace after completing a pre-defined number of problems (25 for the workspaces in this experiment). Interventions which decrease promotion rates suggest that more students are able to master the material as a result of the intervention. As this is a binary variable, we used a generalized linear mixed effects model.

For the remaining comparisons, we only consider students who mastered the workspace. As students who do not master the material also have greater error rates, complete the maximum number of problems, and spend more time in the workspace, segmenting this population out from further analyses ensures we are not double-counting the effect captured in the model on promotion rate.

Time to Complete a Workspace The time in minutes that a student spends to complete a workspace. Time to complete a workspace can be affected by variables

external to student learning. We use this measure in combination with other measures to provide additional insight into student performance across conditions. Specifically, time here is a measure of how long it took students to master all skills within a workspace. Students spending less time while still mastering the same skills as their peers in the control group would be a positive learning outcome.

Errors Per Problem The average number of errors a student made per problem.

Total Problems Completed MATHia updates students' mastery of relevant skills as they work through problems in a workspace. Once students have mastered all skills in a workspace, they graduate to the next workspace. Thus, the number of problems completed reflects the rate of the students' progression towards mastery.

Statistical tests were performed in R Project for Statistical Computing using lme4 (Bates et al., 2016) and emmeans (Lenth, 2022).

Additional Comparisons

A comparison between human revised word problems and the original problems was conducted as part of Almoubayyed et al. (2023a). The success of these rewrites led to the replacement of the original problems with the human revised problems in MATHia. Thus, our primary comparison for the *Simpler* workspaces is between the LLM rewritten and human revised problems. If the LLM revisions result in similar improvements to performance, we expect a statistically null difference in the outcomes based on which revision (LLM or human) that the student received. Because a null effect only suggests that a difference between the groups was not detected, we add an additional comparison to prior data collected for the original word problems. As an additional set of analyses, we thus compare the LLM revisions (data collected in Fall 2024) to the results from Almoubayyed, Bastoni et al. (data collected in January 2023, $n = 6008$ students who received the original problems).

Results

Readability Outcomes

Both prompting methods yielded significant improvements in readability. As reported in Norberg et al. (2023), readability for problems revised by the ChatGPT-4 prompt showed significantly improved CAREC-M scores (measuring syntax, lexical diversity, and cohesion) over the original problems, $t(58) = 4.04$, $p < 0.001$, and human rewrites, $t(58) = 4.09$, $p < 0.001$. All other differences were non-significant, all *p-values* > 0.05. For problems revised by the GPT-4 prompt, the LLM rewrites were significantly improved over the original problems across all metrics,

including SBERT scores, $t(74) = -2.07$, $p = 0.04$, CAREC-M scores, $t(74) = 2.94$, $p = 0.004$, New Dale-Chall, $t(72) = 3.67$, $p < 0.001$, and Flesch-Kincaid Grade Level, $t(74) = 4.69$, $p < 0.001$. See Table 3 for means and standard errors.

Performance Results

We separate student outcomes based first on their workspace type (Simpler or Complex) and then based on the outcome being measured. We focus in these sections on the results related to the manipulation. For all outcomes there was a significant effect of reading ability. Emerging readers were more likely to be promoted (i.e., not master a workspace). Those who did master the workspaces also spent longer in the workspace, needed to answer more questions before achieving mastery, and had higher error rates than non-emerging readers. Main effects related to reading ability are reported in the tables in Appendix B alongside the random effects structure.

Simpler Workspace Results (Experiment A)

Experiment A compared outcomes for students who receive the GPT-4 revised problems to students who received human revised problems. Almoubayyed et al. (2023a) found the human rewrites had improved outcomes compared to the original problems (i.e., lower promotions rates, fewer errors per problem, fewer problems needed before completing the workspace, and decreased time to complete the workspace). Thus, successful GPT-4 revisions, when compared to the human rewrites, should reflect improved or null performance. Critically, a null result does not mean there is no difference between the two groups, thus we complement the results here with a comparison to the outcome data for the non-revised problems from Almoubayyed, Bastoni, et al. See Table 1 for descriptive statistics.

Promotion Rate

For *Analyzing Models of 2-Step Equations—Integers* there was a marginal effect of revision type (i.e., human or ChatGPT-4 revised). Students who received the ChatGPT-4 revised problems had 1.18 times (95% CI:[1.00, 1.18]) greater odds of being promoted as compared to students who received the human revised problems, $z = 1.90$, $p = 0.06$.[6] The interaction between revision type and reading ability was not significant, $z = 1.39$, $p = 0.16$.

For *Analyzing Models of 2-Step Equations—Rationals,* there were no significant effects of the type of rewrite overall, $z = 0.46$, $p = 0.65$, nor its interaction with reading ability, $z = -0.46$, $p = 0.65$.

Although there was some evidence that emerging readers were more likely to master the *Analyzing Models of 2-Step Equations—Integers* when they received

[6] Outcomes for a generalized linear model are in log odds. We back transform log odds to odds for ease of interpretation within the text.

the human rewritten content. It was still possible that the ChatGPT-4 revised problems represented an improvement over the original problems. This was the case. For *Analyzing Models of 2-Step Equations—Integers,* students who received the ChatGPT-4 revised problems had 0.82 times lower odds (95% CI:[0.72, 0.94]) of being promoted compared to students who received the original problems, $z = -2.88$, $p = 0.004$. The interaction between revision type and reading ability was not significant, $z = -0.18$, $p = 0.86$.

Results were similar for *Analyzing Models of 2-Step Equations—Rationals.* Students who received the ChatGPT-4 revised problems had 0.64 times lower odds (95% CI:[0.52, 0.77]) of being promoted compared to students who received the original problems, $z = -4.57$, $p < 0.001$. The interaction between revision type and reading ability was not significant, $z = -1.26$, $p = 0.21$.

Time to Completion

For *Analyzing Models of 2-Step Equations—Integers* and *Analyzing Models of 2-Step Equations—Rationals* there were no significant differences in time to completion based on rewrite condition nor its interaction with reading ability, all p-values were > 0.10.

Almoubayyed et al. (2023a) reported medians for time to completion following the discovery of outliers in their data. Indeed, 8% of data from the prior study exceeded the maximum time to completion found in the current study. Because there is a systematic difference in the distribution of outliers across the two datasets, we did not perform comparisons on time to completion between data from the ChatGPT-4 rewrites and the original problems.

Errors Per Problem

For *Analyzing Models of 2-Step Equations—Integers,* there were no significant differences in error rates based on rewrite condition, $t(9827) = 0.75$, $p = 0.46$. However, there was an interaction between rewrite condition and reader status such that the effect of receiving the LLM rewrites was greater for emerging readers than for non-emerging readers, $t(9876) = 2.06$, $p = 0.04$. Although the effects differed based on reading ability, there were no significant simple effects. Emerging readers did not show significant differences in error rates based on rewrite condition, $z = 1.59$, $p = 0.11$, nor did non-emerging readers, $z = 1.39$, $p = 0.16$.

For *Analyzing Models of 2-Step Equations—Rationals* there was a marginally significant effect of rewrite condition such that students made 0.06 fewer errors (95%: CI:[-0.12, 0.00]) per problem on average when completing ChatGPT-4 rewrites as compared to the human rewrites, $t(12280) = -1.90$, $p = 0.06$. However, there was no significant interaction based on reading ability, $t(12280) = -0.49$, $p = 0.63$.

Comparisons to the original problems showed that error rates for ChatGPT-4 rewritten problems were reduced. For *Analyzing Models of 2-Step Equations—Integers,* students made 0.31 fewer errors (95% CI:[-0.40, -0.22]) per problem, $t(8486) = -160.49$, $p < 0.001$, and for *Analyzing Models of 2-Step Equations—Rationals*, they made 0.48 fewer errors (95% CI:[-0.58, -0.39]) per problem,

$t(8204) = -10.19$, $p < 0.001$. There were no additional significant interactions with reading ability for either workspace, all p-values were greater than 0.48.

Number of Problems Completed

There were no significant differences in the number of problems students completed prior to achieving mastery nor in their interactions with reading ability for either workspace at the 0.05 level when comparing ChatGPT-4 rewrites to human rewrites. However, students did complete 0.55 fewer problems (95% CI:[-0.29, -0.81]) per workspace when they received the ChatGPT-4 rewrites as compared to the original problems, $t(8486) = -4.09$, $p < 0.001$. There were no additional interactions with reading ability, $t(8486) = 0.36$, $p = 0.72$.

Simpler Workspaces (Experiment A) Summary

Student performance across rewrite conditions was broadly similar. However, there was one notable marginally significant difference which might suggest that the human rewritten problems had an overall better effect on student outcomes than the ChatGPT-4 rewritten problems. Emerging readers were marginally more likely to be promoted (i.e., not master) the *Analyzing Models of 2-Step Equations—Integers* workspace when they received the ChatGPT-4 revised problems than when they received the human revised problems. However, this same effect was not true for the *Analyzing Models of 2-Step Equations—Rationals* workspace where comparisons of error rates showed a marginal effect that favored the ChatGPT-4 rewrites. Given these two opposing marginal effects and the null results for the remaining measures, we conclude that the effect of the ChatGPT-4 rewrites on student outcomes was similar to the human rewrites. Indeed, as with the human rewrites, direct comparison between performance on the ChatGPT-4 rewritten and the original problems showed that the revised problems resulted in higher rates of mastery, that students required fewer problems to reach mastery, and that students made fewer errors per problem when they received the ChatGPT-4 revised problems.

Complex Workspaces (Experiment B)

In Experiment B, the GPT-4 revisions were compared to the original set of problems. The workspaces in Experiment B were also believed to be more complex. We find support for that in comparing median times to complete *Simpler* versus *Complex* workspaces. The *Simpler* workspaces took a median time of 16.78 min whereas the *Complex* workspaces took a median time of 65.57 min. This additional 54.31 min (95% CI:[53.55, 55.06]) required on average to complete the *Complex* workspaces as compared to the *Simpler* workspace reflects the additional steps and complexity in the workspaces reported in Experiment B, $t(39741) = -140.90$, $p < 0.001$. Table 2 provides descriptive statistics.

Promotion Rate

For the *Modeling the Constant of Proportionality* workspace, there were no significant effects based on whether the problems were revised by GPT-4, all p-values were greater than 0.22.

For *Modeling Linear Relationships Using Multiple Representations*, the odds of promotion were 0.93 times (95% CI:[0.87, 0.99]) lower when students received the GPT-4 rewritten problems as compared to the original problems, $z = -2.41$, $p = 0.02$. There was no significant effect of the interaction between rewrite condition and reading ability, $z = 1.43$, $p = 0.15$.

Time to Completion

For *Modeling the Constant of Proportionality* students spent 1.59 fewer minutes (95% CI:[-2.79, -0.39]) completing the workspace when the they received the GPT-4 rewritten problems compared to the original problems, $t(514) = -2.59$, $p = 0.01$. There was no significant effect of the interaction between rewrite condition and reading ability, $t(26702) = -1.54$, $p = 0.12$.

For *Modeling Linear Relationships Using Multiple Representations*, there were no significant differences in the time it took to complete the workspace based on rewrite condition, all p-values were greater than 0.71.

Errors Per Problem

For *Modeling the Constant of Proportionality*, students who received the GPT-4 rewrites made 0.12 more errors (95% CI:[0.05, 0.19]) per problem on average than students who received the original problems, $t(40750) = 3.40$, $p < 0.001$. The interaction between reading ability and rewrite condition was not significant, $t(40750) = 1.00$, $p = 0.32$.

For *Modeling Linear Relationships Using Multiple Representations*, students who received the GPT-4 rewrites made 0.16 more errors (95% CI:[0.10, 0.21]) per problem on average than students who received the original problems, $t(384) = 5.36$, $p < 0.001$. The interaction between reading ability and rewrite condition was not significant, $t(16490) = 0.88$, $p = 0.38$.

Number of Problems Completed

For both *Modeling the Constant of Proportionality* and *Modeling Linear Relationships Using Multiple Representations* workspaces there were no significant effects related to the rewrite condition, all p-values were greater than 0.12.

Complex Workspaces (Experiment B) Summary

Results for the *Complex* workspaces were mixed. For both workspaces, students made more errors on problems that were revised according to the GPT-4 workflow. However, this increase in errors was not coupled with lower rates of mastery.

Indeed for *Modeling Linear Relationships Using Multiple Representations* promotion rates were lower among students who received the GPT-4 rewritten problems, meaning that these students mastered the content at a greater rate. Ultimately, mastery of the content is the primary goal, and we take this effect as promising despite the increased rate of error on the path to mastery. Further, students who mastered the *Modeling the Constant of Proportionality* workspace were able to do so more quickly when they received the GPT-4 rewritten problems. Notably, this was not accompanied by a reduction in the number of problems they needed to complete and may therefore reflect that students solved the individual problems more quickly.

Overall, when comparing both to the original problems, the GPT-4 revisions to the *Complex* workspaces had less effect on student outcomes than the ChatGPT-4 revisions to *Simpler* workspaces. Even when significant, effect sizes in Experiment B were small in comparison to the effects from Experiment A and results were not consistent across workspaces. We therefore interpret the findings as evidence that interventions to improve readability must be tested on student outcomes to determine to what extent those improvements affected learning.

General Discussion

We set out to test the ability of GPT-4 to improve the readability of math word problems and to subsequently improve student math outcomes, especially among emerging readers who may struggle to learn math as a result of their difficulty with reading word problems. GPT-4 successfully improved the readability of the problems, but the effect this had on student outcomes was varied. In two workspaces, rewrites performed by ChatGPT-4 had similar effects to rewrites performed by human authors and resulted in significant improvement in student outcomes over the original content. In a second set of workspaces, there were some signs that the GPT-4 rewritten content improved student mastery rates but these problems also resulted in higher error rates compared to the original problems. The discrepancy in results suggests that improving the readability of word problems improves outcomes for some types of problems more than for others. Further, it highlights the need for additional work in this area before LLMs can be used to improve the readability of text more broadly.

Effects of Readability on Math Outcomes

We expected revisions to the content to differentially affect emerging versus non-emerging readers as in Almoubayyed et al. (2023a). This expectation was not supported. Where effects were observed, interactions were rare and simple effects did not suggest noteworthy differences between groups. The null result here suggests that all students regardless of their reading skill are affected similarly by changes to the readability of math word problems.

Student outcomes in the workspaces labeled *Simpler* were sensitive to improvements in the readability of math word problems. Across the board, these revisions led to improved outcomes over the original content and similar

outcomes to the human revised content. Effects for the *Complex* workspaces were both more muted and mixed in the direction of the effect. The key difference between the *Simpler* and more *Complex* workspaces was the extent to which all of the steps relied on a mixture of reading and mathematical reasoning. In the *Simpler* workspaces, all steps within the problem required readers to attend closely to the story in the text in order to understand the formation of the equation. Thus, each response required careful reading. In the *Complex* workspaces, students needed to comprehend the text to extract appropriate variables but then they had to perform additional steps which did not involve the text (e.g., plotting points on a graph). Outcomes for these more strictly mathematical steps may be less affected by changes to the readability of the word problem.

One puzzling finding was the increased rates of error for both of the more complex workspaces. One possibility is that the GPT-4 rewrites, despite improving readability scores, were less clear. However, in one of these workspaces, students mastered the content at a greater rate, so it seems unlikely that the revisions had a negative effect on student learning more broadly. It is not immediately clear why this discrepancy in results would arise. The increased error rates did not negatively affect mastery rates nor increase the time students spent in the workspace which suggests that the increased error was small enough so as to not indicate differences in learning. Nevertheless, systematic differences based on which type of problem students received were present. It is still possible that increased readability in the problems resulted in more careless errors as might occur if students' ease-of-process heuristics led them to be overconfident in their responses (e.g., Son & Metcalfe, 2000). In future work, we plan a closer analysis of which skills showed the greatest discrepancy in errors and how those may relate to changes in the LLM revisions. This type of close analysis will be important both to inform how LLMs can be used for this work and to understand the relationship between text readability and performance on math word problems more broadly.

Limitations

A few limitations associated with implementing interventions in active learning environments should be noted. First, *Simpler* and *Complex* workspaces were revised using different prompts. The differential effects could be related to the differences in prompting rather than differences in how changing the readability of the problems affects outcomes. While this is a distinct possibility, we think this is unlikely to be driving the effect for two reasons. First, readability metrics suggest that the readability of the GPT-4 revised *Complex* workspaces were similarly improved as the ChatGPT-4 revised *Simpler* workspaces. Second, the problems revised by GPT-4 were subject to the same quality assurance process as those revised by ChatGPT-4. Human raters read the problems and rejected ones they believed had introduced errors or reduced readability of the problem. The rejected problems were revised again until a satisfactory version was produced. Although human reviewers rejected a greater percentage of the problems

revised by GPT-4, all problems that were included in the experiment ultimately passed this inspection. Thus, it seems likely that results are related more to how the workspaces are affected by improvements in readability rather than differences in the quality of questions the two different prompts produced.

Future Directions

The relationship between reading ability and solving math word problems is well established (Almoubayyed et al., 2023c; Daroczy et al., 2015; Greisen et al., 2021; Helwig et al., 1999; Koedinger & Nathan, 2004). Indeed, in this study emerging readers had lower performance across all metrics. It is less clear how improvements to the readability of math word problems affect this relationship. Broad changes to readability improved student outcomes for some math word problems but not others. Although aspects of readability like sentence length, word frequency, and lexical diversity have been largely predictive of reading comprehension, the comprehensibility of math word problems may be more dependent on other factors like iconicity (i.e., displaying events in temporal order). LLMs offer an opportunity to iterate improvements, targeting more precise changes in word problems and deploying those changes more broadly to directly test how changes to the readability relate to math learning and particularly to discover if a specific readability metric is useful for capturing the readability of math word problems. Such work may help add precision to understanding how specific findings related to the readability of a text are affecting student performance.

Conclusion

GPT-4 was able to produce revisions to math word problems which were comparable to human revisions of the same problems. This was true both in assessments of their readability and in their effect on student performance in MATHia. However, the extensive prompting required in this first attempt did not provide a scalable solution to revising additional workspaces. By implementing API calls to GPT-4 within Python code we were able to create a workflow that substantially reduced the human labor needed to make large-scale changes in educational materials. The revisions produced by this prompt show similar improvements to the readability of the problems and all problems were approved by human reviewers before deployment. Nevertheless, effects on student performance for this second set of revisions were more muted with only the revisions for one of the two workspaces improving student rates of content mastery. Further, despite the increased rate of mastery, the LLM revisions also led to higher error rates.

The results here highlight the importance of exercising caution when making generalizations about the impact of improved readability on student learning outcomes. Supporting emerging readers may require more than making problems more readable in general. Nevertheless, the results here also illustrate the potential for using LLMs both to improve educational materials and to understand what types of

improvements are most efficacious. By including LLMs in the workflow, offloading the most labor-intensive aspects of the work (writing and coding), we can generate materials more efficiently and deploy experiments more rapidly than using human labor alone. While we report the results from 4 workspaces here, we have already deployed GPT-4 to revise an additional 14 workspaces (encompassing over 1600 problems). We are optimistic that the rich data set we will generate across multiple workspaces will reveal more about how improvements in readability affect learning gains for specific skills.

Appendix A

Complex Workspaces Prompt

First Pass

You are a math problem revision bot. Your job is to revise math problems and make them easier for 6th graders to read.\n.

The text you will be given is not the final version the student will read. It contains placeholder text which begins with $. The placeholder text will later be substituted for actual text. You should leave the placeholder intact. Remember the placeholder text represents words. Be careful not to duplicate text that will be inserted by the placeholder. In particular, placeholders that contain the word "phrase" often represent the unit, so you should not insert a unit following these placeholders.\n.

It is critical that values like [_value_slope] and [/_value_slope] ${ORT} are retained in the output.\n.

In your revision, prioritize using simple but precise vocabulary (e.g., 'nests' is better than 'bird homes' and 'was' is better than 'had been'). Pronouns, especially "it" can be vague. Restate what "it" is instead of using the pronoun.\n.

Sometimes the original problem may contain abstractions like 'Area K'. Make them concrete. For instance, 'Area K' might be a 'field' and 'Area L' a 'backyard.'\n.

Do not rewrite parts of the text that are already simple. Leave simple words that are specific like "buy" and "spend" as they are. Your revisions should never increase the complexity of the text.\n.

You have a bad bias towards reorganizing sentences by adding clauses to the beginning of the sentence. Avoid adding clauses at the beginning of sentences (e.g., 'During …,', 'In …,'). Use shorter sentences instead.\n.

Stay in active voice, always. Define active voice before you start revising.\n.

You should change the text as much as necessary to improve readability, but keep in mind that the scale must stay the same. Size is important because nesting sometimes occurs in problems where there are, for example, a number of schools in a district or a number of teams in a league. Changing league to team will throw off the numbers. Instead change hard words like 'leagues' to words with equivalent sizes like 'clubs.'\n.

You can remove text (including placeholders) that are not relevant to the problem.\n.

For reference, the values for some placeholders have been provided. These are just to help you predict appropriate words. Do not substitute the placeholders in the text with their values.\n.

Before you begin, write a few sentences about what makes a text easy for 6th graders to read. Then write a few more sentences about how you might improve the readability of the text provided. Focus on the reading difficulty. Don't comment on placeholder text. Assume placeholder text will not be seen by the students. You should list a few vocabulary words from the text that are ok to keep as is and then list a few vocabulary words which are difficult. Remember, placeholder text should not be listed among the vocabulary to change.

Subsequent Passes (when necessary)

Your revision did not improve the readability enough. Please try again. Here are some tips to keep in mind: \n.

1. The text should have low lexical diversity. Use the same words rather than diversifying the vocabulary.\n
2. Keep sentences short. You have a tendency to use a lot of clauses in your sentences. Break long sentences into shorter ones.\n
3. Choose the most likely words. This might mean changing the meaning of the text. It's ok to completely change the topic to make the text more clear.\n
4. Be precise. Use words like 'bought' over 'got.' You like to use the word "use" a lot. You're over using it. Be more precise. For example, don't use "use" in place of "spend."\n
5. Make sure the topic is one for which 6th graders have sufficient background knowledge. If not, change the topic.\n
6. Balance providing context with removing information (words and phrases) that is not critical to completing the math problem. All placeholder text should be considered critical.
7. Maintain the same structure you did previously labeling this problem < Revised Problem 1 > .

Appendix B

Tables for all mixed effects models run as part of this study are provided below. Contrast coding was used in all models with tables specifying how to read the direction of the effect. In all cases, the direction of the effects reflects receiving the LLM rewrites or status as an emerging reader (Tables 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25).

Table 4 Human versus LLM rewrites comparison for Analyzing Models of Two-Step Linear Equations Integers

	Odds of being Promoted			
Predictors	*Odds Ratios*	*CI*	*z*	*p*
(Intercept)	0.10	0.09 – 0.11	-31.46	**< 0.001**
LLM Rewrites	1.18	0.99 – 1.39	1.90	0.057
Emerging Readers	2.99	2.60 – 3.43	15.49	**< 0.001**
LLM Rewrites × Emerging Readers	1.21	0.93 – 1.58	1.39	0.164
Random Effects				
σ^2	3.29			
$\tau_{00\ \text{School Id}}$	0.77			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	0.01			
$\rho_{01\ \text{School Id}}$	-0.57			
ICC	0.19			
$N_{\ \text{School Id}}$	476			
Observations	11362			
Marginal R^2 / Conditional R^2	0.054 / 0.234			

Table 5 Human versus LLM rewrites comparison for Analyzing Models of Two-Step Linear Equations Rationals

	Odds of being Promoted			
Predictors	*Odds Ratios*	*CI*	*z*	*p*
(Intercept)	0.05	0.04 – 0.06	-36.10	**< 0.001**
LLM Rewrites	1.05	0.84 – 1.32	0.46	0.645
Emerging Readers	3.68	3.12 – 4.34	15.46	**< 0.001**
LLM Rewrites × Emerging Readers	0.93	0.67 – 1.28	-0.46	0.646
Random Effects				
σ^2	3.29			
$\tau_{00\ \text{School Id}}$	0.71			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	0.14			
$\rho_{01\ \text{School Id}}$	0.05			
ICC	0.19			
$N_{\ \text{School Id}}$	375			
Observations	13307			
Marginal R^2 / Conditional R^2	0.073 / 0.245			

Table 6 Original versus LLM rewrites comparison for Analyzing Models of Two-Step Linear Equations Integers

Predictors	Odds of being Promoted			
	Odds Ratios	*CI*	*z*	*p*
(Intercept)	0.15	0.14 – 0.16	-55.92	**< 0.001**
LLM Rewrites	0.82	0.72 – 0.94	-2.88	**0.004**
Emerging Readers	3.63	3.18 – 4.15	18.96	**< 0.001**
LLM Rewrites × Emerging Readers	0.98	0.75 – 1.27	-0.18	0.859
Observations	9525			
R^2 Tjur	0.043			

Table 7 Original versus LLM rewrites comparison for Analyzing Models of Two-Step Linear Equations Rationals

Predictors	Odds of being Promoted			
	Odds Ratios	*CI*	*z*	*p*
(Intercept)	0.08	0.08 – 0.09	-49.87	**< 0.001**
LLM Rewrites	0.64	0.52 – 0.77	-4.57	**< 0.001**
Emerging Readers	4.50	3.71 – 5.48	15.18	**< 0.001**
LLM Rewrites × Emerging Readers	0.78	0.53 – 1.15	-1.26	0.209
Observations	8734			
R^2 Tjur	0.038			

Table 8 Human versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Integers

Predictors	Time to Completion in Minutes			
	Estimates	*CI*	*t*	*p*
(Intercept)	26.78	25.80 – 27.76	53.41	**< 0.001**
LLM Rewrites	0.32	-0.65 – 1.28	0.64	0.519
Emerging Readers	6.71	5.51 – 7.91	10.95	**< 0.001**
LLM Rewrites × Emerging Readers	0.72	-1.20 – 2.64	0.73	0.464
Random Effects				
σ^2	416.02			
$\tau_{00\ \text{School Id}}$	45.63			
$\tau_{11\ \text{School Id. Emerging Readers}}$	16.69			
$\rho_{01\ \text{School Id}}$	0.74			
ICC	0.08			
$N_{\ \text{School Id}}$	464			
Observations	10262			
Marginal R^2 / Conditional R^2	0.017 / 0.101			

Table 9 Human versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Rationals

	Time to Completion in Minutes			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	20.35	19.59 – 21.10	53.09	**< 0.001**
LLM Rewrites	-0.01	-0.68 – 0.66	-0.02	0.983
Emerging Readers	4.45	3.52 – 5.38	9.34	**< 0.001**
LLM Rewrites × Emerging Readers	-1.08	-2.42 – 0.27	-1.57	0.116
Random Effects				
σ^2	255.65			
$\tau_{00\ \text{School Id}}$	25.50			
$\tau_{11\ \text{School Id. Emerging Readers}}$	14.39			
$\rho_{01\ \text{School Id}}$	0.58			
ICC	0.08			
$N_{\ \text{School Id}}$	368			
Observations	12605			
Marginal R^2 / Conditional R^2	0.013 / 0.095			

Table 10 Human versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Integers

	Error Rate			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	3.48	3.40 – 3.55	87.64	**< 0.001**
LLM Rewrites	0.03	-0.05 – 0.11	0.75	0.455
Emerging Readers	0.86	0.76 – 0.96	17.23	**< 0.001**
LLM Rewrites × Emerging Readers	0.16	0.01 – 0.31	2.06	**0.039**
Random Effects				
σ^2	2.61			
$\tau_{00\ \text{School Id}}$	0.30			
$\tau_{11\ \text{School Id. Emerging Readers}}$	0.12			
$\rho_{01\ \text{School Id}}$	0.15			
ICC	0.11			
$N_{\ \text{School Id}}$	464			
Observations	10262			
Marginal R^2 / Conditional R^2	0.042 / 0.145			

Table 11 Human versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Rationals

	Error Rate			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	2.76	2.69 – 2.83	74.62	**< 0.001**
LLM Rewrites	-0.06	-0.12 – 0.00	-1.90	0.058
Emerging Readers	0.61	0.53 – 0.68	15.20	**< 0.001**
LLM Rewrites × Emerging Readers	-0.03	-0.15 – 0.09	-0.49	0.626
Random Effects				
σ^2	1.99			
$\tau_{00\ \text{School Id}}$	0.27			
$\tau_{11\ \text{School Id. Emerging Readers}}$	0.08			
$\rho_{01\ \text{School Id}}$	0.41			
ICC	0.11			
$N_{\ \text{School Id}}$	368			
Observations	12604			
Marginal R^2 / Conditional R^2	0.029 / 0.139			

Table 12 Originals versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Integers

	Error Rate			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	3.68	3.63 – 3.72	160.49	**< 0.001**
LLM Rewrites	-0.31	-0.40 – -0.22	-6.85	**< 0.001**
Emerging Readers	1.09	1.00 – 1.18	23.85	**< 0.001**
LLM Rewrites × Emerging Readers	0.07	-0.11 – 0.24	0.71	0.478
Observations	8490			
R^2 / R^2 adjusted	0.074 / 0.073			

Table 13 Originals versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Rationals

	Error Rate			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	2.96	2.91 – 3.00	124.61	**< 0.001**
LLM Rewrites	-0.48	-0.58 – -0.39	-10.18	**< 0.001**
Emerging Readers	0.61	0.51 – 0.70	12.76	**< 0.001**
LLM Rewrites × Emerging Readers	0.06	-0.13 – 0.25	0.64	0.525
Observations	8208			
R^2 / R^2 adjusted	0.047 / 0.046			

Table 14 Human versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Integers

	Number of Problems Completed			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	11.15	10.91 – 11.38	93.92	**< 0.001**
LLM Rewrites	-0.01	-0.24 – 0.22	-0.11	0.913
Emerging Readers	1.99	1.69 – 2.30	12.74	**< 0.001**
LLM Rewrites × Emerging Readers	-0.11	-0.56 – 0.34	-0.47	0.642
Random Effects				
σ^2	22.54			
$\tau_{00\ \text{School Id}}$	2.58			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	0.08			
$\tau_{11\ \text{School Id. Emerging Readers}}$	1.52			
ρ_{01}	-0.31			
	0.48			
ICC	0.10			
$N_{\text{School Id}}$	464			
Observations	10262			
Marginal R^2 / Conditional R^2	0.027 / 0.123			

Table 15 Human versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Rationals

	Number of Problems Completed			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	9.71	9.49 – 9.93	86.91	**< 0.001**
LLM Rewrites	-0.05	-0.23 – 0.13	-0.59	0.557
Emerging Readers	1.64	1.39 – 1.88	12.96	**< 0.001**
LLM Rewrites × Emerging Readers	-0.11	-0.46 – 0.25	-0.58	0.562
Random Effects				
σ^2	18.05			
$\tau_{00\ \text{School Id}}$	2.41			
$\tau_{11\ \text{School Id. Emerging Readers}}$	0.99			
$\rho_{01\ \text{School Id}}$	0.48			
ICC	0.11			
$N_{\text{School Id}}$	368			
Observations	12604			
Marginal R^2 / Conditional R^2	0.023 / 0.132			

Table 16 Original versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Integers

	Number of Problems Completed			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	11.45	11.32 – 11.59	170.69	**< 0.001**
LLM Rewrites	-0.55	-0.81 – -0.29	-4.09	**< 0.001**
Emerging Readers	2.35	2.08 – 2.61	17.49	**< 0.001**
LLM Rewrites × Emerging Readers	0.10	-0.43 – 0.62	0.36	0.719
Observations	8490			
R^2 / R^2 adjusted	0.039 / 0.039			

Table 17 Original versus LLM Rewrites for Analyzing Models of Two-Step Linear Equations Rationals

	Number of Problems Completed			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	10.17	10.03 – 10.31	141.69	**< 0.001**
LLM Rewrites	-1.19	-1.47 – -0.91	-8.29	**< 0.001**
Emerging Readers	1.92	1.64 – 2.21	13.40	**< 0.001**
LLM Rewrites × Emerging Readers	-0.47	-1.03 – 0.09	-1.64	0.101
Observations	8208			
R^2 / R^2 adjusted	0.035 / 0.035			

Table 18 Original versus LLM Rewrites for Modeling the Constant of Proportionality

	Odds of being Promoted			
Predictors	*Odds Ratios*	*CI*	*t*	*p*
(Intercept)	0.02	0.02 – 0.02	-52.76	**< 0.001**
LLM Rewrites	1.14	0.93 – 1.39	1.24	0.216
Emerging Readers	3.52	3.08 – 4.03	18.25	**< 0.001**
LLM Rewrites × Emerging Readers	0.98	0.75 – 1.28	-0.14	0.891
Random Effects				
σ^2	3.29			
$\tau_{00\ \text{School Id}}$	1.19			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	0.10			
$\rho_{01\ \text{School Id}}$	0.15			
ICC	0.27			
$N_{\ \text{School Id}}$	782			
Observations	42471			
Marginal R^2 / Conditional R^2	0.063 / 0.316			

Table 19 Original versus LLM Rewrites for Modeling Linear Relationships Using Multiple Representations

	Odds of being Promoted			
Predictors	*Odds Ratios*	*CI*	*t*	*p*
(Intercept)	0.32	0.29 – 0.34	-27.95	**< 0.001**
LLM Rewrites	0.93	0.87 – 0.99	-2.41	**0.016**
Emerging Readers	2.47	2.32 – 2.63	27.94	**< 0.001**
LLM Rewrites × Emerging Readers	1.09	0.97 – 1.24	1.43	0.153
Random Effects				
σ^2	3.29			
$\tau_{00\ \text{School Id}}$	0.90			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	0.01			
$\rho_{01\ \text{School Id}}$	-0.06			
ICC	0.22			
$N_{\ \text{School Id}}$	956			
Observations	33565			
Marginal R^2 / Conditional R^2	0.035 / 0.243			

Table 20 Original versus LLM Rewrites for Modeling the Constant of Proportionality

	Time to Completion in Minutes			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	84.50	82.41 – 86.60	79.05	**< 0.001**
LLM Rewrites	-1.59	-2.79 – -0.39	-2.59	**0.010**
Emerging Readers	20.87	19.64 – 22.11	33.15	**< 0.001**
LLM Rewrites × Emerging Readers	-1.87	-4.26 – 0.52	-1.54	0.124
Random Effects				
σ^2	2772.33			
$\tau_{00\ \text{School Id}}$	639.51			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	3.07			
$\rho_{01\ \text{School Id}}$	-0.99			
ICC	0.19			
$N_{\ \text{School Id}}$	780			
Observations	41477			
Marginal R^2 / Conditional R^2	0.023 / 0.206			

Table 21 Original versus LLM Rewrites for Modeling Linear Relationships Using Multiple Representations

	Time to Completion in Minutes			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	83.11	81.22 – 85.00	86.22	**< 0.001**
LLM Rewrites	-0.22	-1.75 – 1.32	-0.28	0.782
Emerging Readers	11.79	10.20 – 13.38	14.57	**< 0.001**
LLM Rewrites × Emerging Readers	0.57	-2.50 – 3.63	0.36	0.717
Random Effects				
σ^2	2547.99			
$\tau_{00\ school\_id}$	484.89			
$\tau_{11\ school\_id.ProblemCondition1}$	2.34			
$\rho_{01\ school\_id}$	0.82			
ICC	0.16			
$N_{\ school\_id}$	907			
Observations	26493			
Marginal R^2 / Conditional R^2	0.007 / 0.166			

Table 22 Original versus LLM Rewrites for Modeling the Constant of Proportionality

	Error Rate			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	6.49	6.36 – 6.63	92.69	**< 0.001**
LLM Rewrites	0.12	0.05 – 0.19	3.40	**0.001**
Emerging Readers	1.54	1.44 – 1.64	29.41	**< 0.001**
LLM Rewrites × Emerging Readers	0.07	-0.07 – 0.21	1.00	0.317
Random Effects				
σ^2	9.12			
$\tau_{00\ School\ Id}$	2.80			
$\tau_{11\ School\ Id.\ Emerging\ Readers}$	0.54			
$\rho_{01\ School\ Id}$	0.71			
ICC	0.21			
$N_{\ School\ Id}$	780			
Observations	41474			
Marginal R^2 / Conditional R^2	0.036 / 0.243			

Table 23 Original versus LLM Rewrites for Modeling Linear Relationships Using Multiple Representations

	Error Rate			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	4.44	4.36 – 4.51	115.33	**< 0.001**
LLM Rewrites	0.16	0.10 – 0.21	5.36	**< 0.001**
Emerging Readers	0.51	0.44 – 0.59	12.85	**< 0.001**
LLM Rewrites × Emerging Readers	0.05	-0.06 – 0.16	0.88	0.380
Random Effects				
σ^2	3.34			
$\tau_{00\ \text{School Id}}$	0.77			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	0.01			
$\tau_{11\ \text{School Id. Emerging Readers}}$	0.23			
ρ_{01}	0.32			
	0.52			
ICC	0.17			
$N_{\text{School Id}}$	907			
Observations	26492			
Marginal R^2 / Conditional R^2	0.012 / 0.182			

Table 24 Original versus LLM Rewrites for Modeling the Constant of Proportionality

	Number of Problems Completed			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	10.63	10.51 – 10.75	175.54	**< 0.001**
LLM Rewrites	0.03	-0.05 – 0.12	0.74	0.457
Emerging Readers	1.25	1.16 – 1.33	28.30	**< 0.001**
LLM Rewrites × Emerging Readers	-0.13	-0.30 – 0.03	-1.56	0.120
Random Effects				
σ^2	13.60			
$\tau_{00\ \text{School Id}}$	1.82			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	0.01			
$\rho_{01\ \text{School Id}}$	-0.86			
ICC	0.12			
$N_{\text{School Id}}$	780			
Observations	41474			
Marginal R^2 / Conditional R^2	0.018 / 0.134			

Table 25 Original versus LLM Rewrites for Modeling Linear Relationships Using Multiple Representations

	Number of Problems Completed			
Predictors	*Estimates*	*CI*	*t*	*p*
(Intercept)	13.81	13.65 – 13.98	162.21	**< 0.001**
LLM Rewrites	-0.09	-0.26 – 0.07	-1.09	0.274
Emerging Readers	0.90	0.72 – 1.08	9.81	**< 0.001**
LLM Rewrites × Emerging Readers	0.14	-0.18 – 0.45	0.84	0.401
Random Effects				
σ^2	26.79			
$\tau_{00\ \text{School Id}}$	3.16			
$\tau_{11\ \text{School Id. LLM Rewrites}}$	0.19			
$\tau_{11\ \text{School Id. Emerging Readers}}$	0.42			
ρ_{01}	0.22			
	0.07			
ICC	0.11			
$N_{\text{School Id}}$	907			
Observations	26492			
Marginal R^2 / Conditional R^2	0.005 / 0.113			

Acknowledgements We would like to thank Dr. Scott Crossley, Yahan Yan, and Joon Suh Choi for providing the version of the SentenceBERT model that was used in this work and the code for using it. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R324A210289 to Center for Applied Special Technology (CAST). The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Author Contributions All authors contributed to the study conception and design. Material preparation and analysis were performed by Kole A. Norberg, Kyle Weldon, and Husni Almoubayyed. All authors served in review of the materials. Time/cost analysis was performed by Logan De Ley. Data collection was executed by April Murphy. Conceptualization was motivated by Steve Ritter. The first draft of the manuscript was written by Kole A. Norberg and all authors contributed to subsequent versions of the manuscript. All authors read and approved the final manuscript.

Funding The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through grant R324A210289 to Center for Applied Special Technology (CAST). The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Data Availability Data from schools which permit sharing de-identified data is available upon request from the first author.

Declarations

Competing Interests The authors have no competing interests to report for this work.

References

Ali, R., Tang, O. Y., Connolly, I. D., Zadnik Sullivan, P. L., Shin, J. H., Fridley, J. S., & Telfeian, A. E. (2023). Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery, 93*(6), 1353–1365. https://doi.org/10.1227/neu.0000000000002632

Almoubayyed, H., Bastoni, R., Berman, S. R., Galasso, S., Jensen, M., Lester, L., ... & Ritter, S. (2023a). Rewriting Math Word Problems to Improve Learning Outcomes for Emerging Readers: A Randomized Field Trial in Carnegie Learning's MATHia. In *International Conference on Artificial Intelligence in Education* (pp. 200–205). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36336-8_30

Almoubayyed, H., Fancsali, S. E., Ritter, S. (2023b) Generalizing predictive models of reading ability in adaptive mathematics software, in: Proceedings of the 16th International Conference on Educational Data Mining, EDM2023.

Almoubayyed, H., Fancsali, S. E., Ritter, S. (2023c). Instruction-embedded assessment for reading ability in adaptive mathematics software. In Proceedings of the 13th International Conference on Learning Analytics and Knowledge, LAK '23, Association for Computing Machinery, New York, NY, USA.

Anthropic (2023). Model Card and Evaluations for Claude Model: Technical Report.

Arbel, Y. A., & Becher, S. I. (2023). How smart are smart readers? LLMs and the future of the no-reading problem. In The Cambridge handbook on emerging issues at the intersection of commercial law and technology (Elvy & Kim, Eds., forthcoming 2024). https://doi.org/10.2139/ssrn.4491043

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2016). lme4: Mixed-effects modeling with R; 2010. (8 April 2015) http://lme4.r-forge.r-project.org/book/.

Bestgen, Y., & Vonk, W. (2000). Temporal adverbials as segmentation markers in discourse comprehension. *Journal of Memory and Language, 42*(1), 74–87. https://doi.org/10.1006/jmla.1999.2670

Betts, E. (1946). *Foundations of reading instruction*. American Book Company.

Butler, J. J., Harrington, M. C., Tong, Y., Rosenbaum, A. J., Samsonov, A. P., Walls, R. J., & Kennedy, J. G. (2024). From Jargon to Clarity: Improving the Readability of Foot and Ankle Radiology Reports with an Artificial Intelligence Large Language Model. *Foot and Ankle Surgery*. https://doi.org/10.1016/j.fas.2024.01.008

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology, 103*(2), 429. https://doi.org/10.1037/a0022824

Chall, J. S., & Dale, E. (1995). *Readability revisited, the new Dale-Chall readability formula*. Brookline Books.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Choi, J. S., & Crossley, S. A. (2022, July). Advances in Readability Research: A New Readability Web App for English. In *2022 International Conference on Advanced Learning Technologies (ICALT)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICALT55010.2022.00007

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278. https://doi.org/10.1007/BF01099821

Crossley, S. A., Skalicky, S., & Dascalu, M. (2019). Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading, 42*(3–4), 541–561. https://doi.org/10.1111/1467-9817.12283

Crossley, S., Choi, J. S., Scherber, Y., & Lucka, M. (2023). Using Large Language Models to Develop Readability Formulas for Educational Settings. In International Conference on Artificial Intelligence in Education (pp. 422–427). Cham: Springer Nature Switzerland.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes, 54*(5–6), 340–359. https://doi.org/10.1080/0163853X.2017.1296264

Daroczy, G., Wolska, M., Meurers, W. D., & Nuerk, H. C. (2015). Word problems: A review of linguistic and numerical factors contributing to their difficulty. *Frontiers in Psychology, 6*, 348. https://doi.org/10.3389/fpsyg.2015.00348

Duffy, T. M. (1985). Readability formulas: What's the use?. In *Designing usable texts* (pp. 113–143). Academic Press. https://doi.org/10.1016/B978-0-12-223260-2.50011-6

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., Schatschneider, C., & Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic,

algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology, 98*(1), 29. https://doi.org/10.1037/0022-0663.98.1.29

Fuchs, L. S., Gilbert, J. K., Fuchs, D., Seethaler, P. M., & Martin, B. N. (2018). Text comprehension and oral language as predictors of word-problem solving: Insights into word-problem solving as a form of text comprehension. *Scientific Studies of Reading, 22*(2), 152–166. https://doi.org/10.1080/10888438.2017.1398259

Gomez-Rodriguez, C., & Williams, P. (2023). A confederacy of models: A comprehensive evaluation of LLMs on creative writing. ArXiv, abs/2310.08433.

Greisen, M., Georges, C., Hornung, C., Sonnleitner, P., & Schiltz, C. (2021). Learning mathematics with shackles: How lower reading comprehension in the language of mathematics instruction accounts for lower mathematics achievement in speakers of different home languages. *Acta Psychologica, 221*, 103456. https://doi.org/10.1016/j.actpsy.2021.103456

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research, 93*(2), 113–125. https://doi.org/10.1080/00220679909597635

Hoeks, J. C., Stowe, L. A., & Doedens, G. (2004). Seeing words in context: The interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19*(1), 59–73. https://doi.org/10.1016/j.cogbrainres.2003.10.022

Huang, C.-Y., Wei, J., & Huang, T.-H. K. (2024, May 11). Generating educational materials with different levels of readability using LLMs. In *In2Writing 2024*, Honolulu, HI.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329. https://doi.org/10.1037/0033-295X.87.4.329

Keene, E. O., & Zimmermann, S. (1997). *Mosaic of thought: Teaching comprehension in a reader's workshop*. Heinemann.

Kincaid, J. P., Fishburne Jr, 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. Technical Report, Naval Technical Training Command, Millington, TN, Research Branch, 1975.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *The Journal of the Learning Sciences, 13*(2), 129–164. https://doi.org/10.1207/s15327809jls1302_1

Lenth, R. (2022). emmeans: Estimated marginal means, aka least-squares means. R package version 1.7.2.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56

Metcalfe, J. (2011). Desirable difficulties and studying in the region of proximal learning. Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork, (pp. 259–276).

Miller, D. (2002). *Reading with meaning teaching comprehension in the primary grades*. Stenhouse Publishers.

Millis, K. K., & Just, M. A. (1994). The influence of connectives on sentence comprehension. *Journal of Memory and Language, 33*(1), 128–147. https://doi.org/10.1006/jmla.1994.1007

Morris, D., Trathen, W., Gill, T., Perney, J., Schlagal, R., Ward, D., & Frye, E. M. (2019). Reading Instructional Level from a Print-Processing Perspective. *Reading & Writing Quarterly, 35*(6), 556–571. https://doi.org/10.1080/10573569.2019.1598311

Mounla, G., Bahous, R., & Nabhani, M. (2011). The Reading Matrix© 2011. *Reading, 11*(3), 279–291.

Mugaanyi, J., Cai, L., Cheng, S., Lu, C., & Huang, J. (2024). Evaluation of large language model performance and reliability for citations and references in scholarly writing: Cross-disciplinary study. *Journal of Medical Internet Research, 26*, e52935.

National Center for Education Statistics. (2022). National Achievement-Level Results. https://www.nationsreportcard.gov/reading/nation/achievement/?grade=8

Norberg, K. A. (2022). *Avoiding miscomprehension: A metacognitive perspective for how readers identify and overcome comprehension failure*, Doctoral dissertation, University of Pittsburgh.

Norberg, K. A., Almoubayyed, H. et al. (2023, July 7). Rewriting Math Word Problems with Large Language Models. In: AIED2023 Empowering Education with LLMs workshop, Tokyo, Japan https://ai4ed.cc/workshops/aied2023

OpenAI (2023), GPT-4 Technical Report.

O'reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43*(2), 121–152. https://doi.org/10.1080/01638530709336895

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*(3), 228–242. https://doi.org/10.1016/j.learninstruc.2008.04.003

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Ritter, S., Murphy, A., Fancsali, S. E., Fitkariwala, V., Patel, N., & Lomas, J. D. (2020). UpGrade: An open source tool to support A/B testing in educational software. In *Proceedings of the First Workshop on Educational A/B Testing at Scale (at Learning@ Scale 2020)*.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive Tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review, 14*, 249–255. https://doi.org/10.3758/BF03194060

Saravia, E. (2022). Prompt Engineering Guide. https://github.com/dair-ai/Prompt-Engineering-Guide

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 204. https://doi.org/10.1037/0278-7393.26.1.204

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837.

Zainurrahman, Z., Yusuf, F. N., & Sukyadi, D. (2024). Text readability: Its impact on reading comprehension and reading time. *Journal of Education and Learning (EduLearn), 18*(4), 1422–1432.

Zheng, G., Fancsali, S. E., Ritter, S., & Berman, S. (2019). Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *Journal of Learning Analytics, 6*(2), 153–174. https://doi.org/10.18608/jla.2019.62.11

Zwaan, R. A. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1196. https://doi.org/10.1037/0278-7393.22.5.1196

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.