

Using ChatGPT to Innovate Mathematical Word Problems

^{1*}Jaurigue, Jerryco M., ¹Lacro, Romeo III P., ¹Laset, Leopoldo Jr. B., ¹Gaza, John Cedric C.

* jmjaurigue@up.edu.ph

University of the Philippines Rural High School
University of the Philippines Los Baños
Philippines

Abstract: *This study explores the capabilities of Generative Artificial Intelligence (GenAI) in innovating and rewriting mathematical word problems. Recognizing that rewriting word problems is a complex cognitive process that is often challenging and time-consuming for teachers, we leverage the growing capabilities and performance of GenAI. This study examines ChatGPT-4 as a support tool for rewriting Grade 7 mathematical word problems on linear equations in one variable, utilizing six innovation techniques: replacement, addition, modification, contextualization, inversion, and reformulation. We applied zero-shot, few-shot, and chain-of-thought prompts. Experts evaluated outputs for adherence to the target innovation technique (ADH), appropriateness for Grade 7 (APP), solvability (SOL), and competency alignment (COMP). We also computed Flesch Reading Ease and grade-level estimates. Results show strong performance for the replacement technique across prompt types, with consistent positive ratings and suitable readability. When applied using the addition technique, the outputs followed the intended form; however, experts often had reservations about appropriateness, solvability, and competency alignment because the added conditions increased cognitive load. Modification frequently rephrased rather than changed the structure. Inversion did not achieve true structure reversal, though items remained appropriate and solvable. Reformulation largely adhered but tended to yield variable-based tasks that were too abstract for Grade 7. Prompt effectiveness showed no single best strategy. Some tasks worked with zero-shot prompts, while others benefited from few-shot or chain-of-thought.*

1. Introduction

At the start of each academic year, educators look for fresh assessment ideas. In mathematics, teachers must decide whether to reuse old problems or replace them. Reuse risks item exposure, leaks, and reduced engagement. Yet some items remain effective and may only need a fresh look or small tweaks [1]. Composing new, complex problems is demanding because it is a core professional skill. A practical response is to derive new items from existing ones using structured innovation procedures [1]. In this study, we rewrite mathematical word problems using six classroom-tested innovation techniques [1], expanded in section 2.2.

Automating such rewrites helps teachers assemble question pools, create parallel forms for different sections, reduce opportunities for cheating, and tailor contexts to learner interests and domains such as health or agriculture. It also saves time and supports consistent alignment with curricular goals.

Research on automated question generation (AQG) for mathematical word problems has grown with advances in educational technology. Recent reviews summarize methods and trends in AQG [2]. Some studies generate word problems from equations using trained language models [3]. With accessible generative AI, non-programmers can now produce revisions through **prompt engineering**, where a large language model predicts text conditioned on a user's instructions [4]. In education, prompt engineering supports problem creation, assessment, feedback, knowledge representation, programming support, lesson planning, and chatbots [5]. Comparative work shows that optimal prompting depends on the dataset and task [6], with applications across K–12 STEM, including physics question answering with Mixture-of-Experts models [7], zero-shot, one-shot, and

chain-of-thought rewrites of math problems [8], expert-rated AI-supported problem solving using zero-shot, few-shot, chain-of-thought, and Ask-Me-Anything prompts [9], STEM Q&A generation evaluated for clarity and accuracy [10], and orchestration frameworks for solving equations, a known challenge for LLMs [11].

With all of these in mind, the present paper examines how ChatGPT can support the innovation of mathematical word problems, focusing on word-problems involving linear equations in one variable, and considers quality, grade-level appropriateness, and readability of the generated outputs.

2. Methodology

This study adopted an exploratory case study design to examine the potential of ChatGPT in innovating word problems involving linear equations in one variable. As an initial investigation into the integration of ChatGPT in mathematics instruction, the scope was deliberately limited to this specific type of word problem. The exploration focused on analyzing ChatGPT's outputs generated through three distinct prompt engineering strategies (one-shot, few-shot, and chain-of-thought prompting) in combination with six different word problem innovation techniques.

2.1. Problem Selection and Rewriting by the authors

The authors collaboratively selected four-word problems involving linear equations in one variable. The first problem was designated for rewriting using ChatGPT, while the second to fourth problems served as input examples for the few-shot and chain-of-thought prompting strategies. The sole criterion for selection was that each problem should be a typical word problem found in instructional materials, involve a linear equation in one variable, and be appropriate for Grade 7 learners. The selected problems can be seen below:

Problem 1 [12]. An artist invests in a tax-free bond paying 6%, and \$6000 more than three times as much in mutual funds paying 5%. Her total annual interest income from the investments is \$825. How much does she invest at each rate?

Problem 2 [13]. A father is three times as old as his son. In twelve years, he will be twice as old as his son. How old are the father and son now?

Problem 3 [12]. Two cars leave Iowa City, Iowa, at the same time and travel east on Interstate 80. One travels at a constant rate of 55 mph. The other travels at a constant rate of 63 mph. In how many hours will the distance between them be 24 mi?

Problem 4 [14]. Nick and Chloe have two grades of milk from their small dairy herd: one that is 24% butterfat and another that is 18% butterfat. How much of each should they use to end up with 42 litres of 20% butterfat?

The selected problems represent four common applications of linear equations: money and investment (Problem 1), age problems (Problem 2), distance-rate-time problems (Problem 3), and mixture problems (Problem 4). Guided by the innovation strategies proposed by [1], each of the second, third, and fourth problems was transformed into a set of "new" or modified problems. As a result, six new problems are generated, producing a total of eighteen. Each newly constructed

problem was then reviewed and validated, then discussed the feedback until they reached a consensus.

2.2. Prompt Engineering

After finalizing the set of original and modified problems, all prompts were written in advance before being submitted to ChatGPT. Each prompt was constructed following the structural components of effective prompt engineering as outlined by [15] and formatted according to the guidelines specific to each prompting technique as described by [4]. All prompts underwent multiple rounds of revision to ensure clarity, consistency, and alignment with the intended prompt engineering approach. Three main prompting techniques based on [15] are used:

1. Zero-shot approach: This is the simplest prompting method that needs no further examples and only needs a simple description for ChatGPT to work with.
2. One-shot / Few-shot approach: This method allows for giving one example (one-shot) or more examples (few-shot) for ChatGPT to understand the task that needs to be done.
3. Chain of thought: This method adds additional steps that serve as a guide for ChatGPT to understand how to get to the problem, which can also be integrated with the few-shot approach.

| Prompt |
|--|
| Rewrite this word problem by changing the given values: "An artist invests in a tax-free bond paying 6%, and \$6000 more than three times as much in mutual funds paying 5%. Her total annual interest income from the investments is \$825. How much does she invest at each rate?" |
| Output |
| An artist invests in a tax-free bond paying 7%, and \$8000 more than four times as much in mutual funds paying 4%. Her total annual interest income from the investments is \$1,200. How much does she invest at each rate? |

Figure 1. Prompt and output for a task that asks ChatGPT to innovate a problem using the replacement technique [1].

As shown in Figure 1, which is an example of the one-shot approach, the prompt used in this instance incorporated four key elements of prompt engineering: instruction, context, input, and output. The instruction directed ChatGPT to "rewrite this word problem," while the context specified that the revision should involve changes to the given values. The input was a standard algebraic word problem involving investments: "An artist invests in a tax-free bond paying 6%, and \$6000 more than three times as much in mutual funds paying 5%. Her total annual interest income from the investments is \$825. How much does she invest at each rate?"

Although the output was not explicitly defined in the prompt, ChatGPT generated a revised version that closely mirrored the original in terms of length, structure, and problem type. For other innovation types, the corresponding one-shot prompts followed the same general format but varied in their instructional component. For example:

- **Addition:** "Rewrite this word problem by adding a new given, constraint, or obstacle."
- **Modification:** "Rewrite this word problem by changing or modifying it while retaining the given values."
- **Contextualization:** "Rewrite this word problem to make it more relevant to Grade 7 Filipino students."
- **Inversion (Turning the Problem Around):** "Rewrite this word problem by reversing it, that is, use the original end goal as the given, and the given as the end goal."
- **Reformulation:** "Rewrite this word problem as a different type of mathematical problem."

Prompt

Consider these problems and their revisions:

Problem 1: Two cars leave Iowa City, Iowa, at the same time and travel east on Interstate 80. One travels at a constant rate of 55 mph. The other travels at a constant rate of 63 mph. In how many hours will the distance between them be 24 mi?

Revision 1: Two cars leave Iowa City, Iowa, at the same time and travel east on Interstate 80. One travels at a constant rate of 55 mph. The other travels at a constant rate of 63 mph. In how many hours will the distance between them be between 20 mi and 28 mi?

Problem 2: A father is three times as old as his son. In twelve years, he will be twice as old as his son. How old are the father and son now.

Revision 2: A father is three times as old as his son. In twelve years, he will be anywhere from 1.5 to 2 times as old as his son. How old are the father and son now?

Rewrite this word problem similar to the example above:

An artist invests in a tax-free bond paying 6%, and \$6000 more than three times as much in mutual funds paying 5%. Her total annual interest income from the investments is \$825. How much does she invest at each rate?

Indicate and explain the revisions made.

Output

An artist invests in a tax-free bond paying 6%, and \$6000 more than three times as much in mutual funds paying 5%. Her total annual interest income from the investments is between \$800 and \$850. How much does she invest at each rate?

Figure 2. Prompt and output for a task that asks ChatGPT to innovate a problem using the addition technique using few-shot learning with two examples.

Figure 2 presents the prompt used and the corresponding ChatGPT output for rewriting a given word problem using the addition innovation technique within a few-shot learning framework. The prompt's instruction is clearly stated as "rewrite this word problem similar to the example above," although it is not positioned at the beginning. Instead, the prompt opens with the context, established through two illustrative problem-revision pairs. The first one involves a distance-rate-time scenario, and the other an age-related problem. The revisions made for these two problems were modifying the numerical values and introducing ranges of more open-ended conditions. The input is still the word problem similar to Figure 1, and the output now reflects both the rephrased problem and the additional directive: "Indicate and explain the revisions made." The final instruction tells the model the expected task and also defines the structure of the desired output, that is, a revised version of the original problem and a written explanation detailing the specific modifications made.

Prompt

Consider these problems and their revisions:

Example 1

Two cars leave Iowa City, Iowa, at the same time and travel east on Interstate 80. One travels at a constant rate of 55 mph. The other travels at a constant rate of 63 mph. In how many hours will the distance between them be 24 mi?

Revision: Two cars leave Iowa City, Iowa, at the same time and travel east on Interstate 80. One travels at a constant rate of x mph. The other travels at a constant rate of y mph. Make a formula to find the time in hours it will take so that the distance between the two cars is z miles.

Here, the speeds were represented by variables x and y while the distance distance is represented by the variable z and the problem asked for the formula of time in terms of x , y , and z .

Example 2

Problem: A father is three times as old as his son. In twelve years, he will be twice as old as his son. How old are the father and son now?

Revision: A father is three times as old as his son. In twelve years, he will be twice as old as his son. Generalize the relationship between the father's age and the son's age given that the father is X times as old as his son now and will be Y times as old in Z years. How old could they be now based on these parameters?

Here, the problem asks to create a formula for the father's age and son's age given that the father's current age is X times his son's and will be Y times as old in Z years.

Example 3

Problem: Nick and Chloe have two grades of milk from their small dairy herd: one that is 24% butterfat and another that is 18% butterfat. How much of each should they use to end up with 42 litres of 20% butterfat?

Revision: Nick and Chloe need to mix two types of milk with different butterfat percentages to achieve a specific butterfat concentration in a final mixture. If they have one milk with a butterfat percentage of $A\%$ and another with $B\%$, and they want to produce C litres of milk with a butterfat percentage of $D\%$, how much of each type of milk should they use?

Here, the scenario is generalized to mixing two types of milk with unspecified butterfat percentages ($A\%$ and $B\%$) to achieve a specific butterfat concentration in a final mixture ($D\%$) for a specified volume (C liters), rather than using specific values for butterfat percentages and total volume.

Rewrite this word problem similar to the examples above:

An artist invests in a tax-free bond paying 6%, and \$6000 more than three times as much in mutual funds paying 5%. Her total annual interest income from the

investments is \$825. How much does she invest at each rate?

Indicate and explain the revisions made.

Output

An investor allocates funds between two investment types: one investment yields an interest rate of R_1 % and the other yields an interest rate of R_2 %. The amount invested in the second investment is F dollars more than three times the amount invested in the first investment. The total annual interest income from both investments is T dollars. Determine how much is invested at each interest rate.

Figure 3. Prompt and Output demonstrating Chain-of-Thought Prompting for problem reformulation innovation method.

One of the key distinctions between chain-of-thought prompting and few-shot prompting lies in the inclusion of explicit reasoning or explanation accompanying the task. In chain-of-thought prompting, the model is not only shown examples of task completions but is also guided to reflect or generalize based on the structure of those examples. In the case of the prompt shown in Figure 3, the context is constructed through three example problem-revision pairs, each followed by a brief explanation of the changes made. This layer of reasoning shows the chain-of-thought approach, as it guides the model to learn the underlying structure of the problem and apply it to a new case. Similar to Figure 2, the instruction in this prompt appears at the end, directing ChatGPT to “rewrite the word problem similar to the examples above” and to “indicate and explain the revisions made.” The input remains a word problem involving investment, while the output is expected to mirror the generalization pattern established by the examples.

All prompts were executed using the free version of ChatGPT-4.0, without logging into any personal accounts. This was done intentionally to minimize the influence of prior interactions and ensure that the outputs were generated based solely on the information and structure provided within each prompt.

2.3. Evaluation

Three mathematics teachers, all currently completing their doctoral degrees, were invited to evaluate the revised (or “new”) word problems. Two of the expert evaluators are PhD candidates specializing in Educational Measurement and Evaluation, while the third has already completed all academic requirements for a PhD in Mathematics Education. Given that there were seven prompts for each of the six innovation methods, each expert evaluated a total of 42 revised problems. For each item, they responded to the evaluation criteria using the options: Yes, Yes, but with reservations, or NO. The expert evaluators were asked to consider the following questions for each problem:

1. Does the revised word problem adhere to the specified innovation type or format?
2. Is the word problem appropriate for Grade 7 students?
3. Does the word problem have a clear and achievable solution?
4. Does the word problem align with the mathematics skills and knowledge expected at the Grade 7 level?

All generated word problems were analyzed for readability using the online tool available¹. Although other readability analyzers are available, this tool was selected because it has been used in prior studies examining outputs from GenAIs [15], [16]. While the tool provides several standardized readability scores, including the Gunning Fog Index, Flesch-Kincaid Grade Level, and Dale-Chall Score, only the Flesch Reading Ease (FRE) score was considered for this study.

Because different formulas can yield conflicting results, focusing on a single, widely recognized metric helped avoid multiplicity and inconsistent interpretations. The FRE score measures the overall ease of reading a text, with higher values indicating simpler and more accessible

¹ <https://datayze.com/readability-analyzer.php>

language. According to the website, a Flesch score between 60 and 70 is generally appropriate for middle school students, including those in Grade 7. For each revised problem, the corresponding FRE score was recorded along with the grade level most suitable based on that score.

3. Results

Table 1 compiles expert judgments and readability for all six innovation techniques across prompt types. For the *replacement* technique, experts marked ADH, APP, SOL, and COMP as Yes in all rows. Most FRE values matched Grade 7. One-shot and three-shot had lower FRE (69.3 and 60.4), which hints at harder vocabulary.

In addition, adherence was Yes across prompt types. Yet APP, SOL, and COMP were often Yes with reservation from 1-shot through CoT-3. These reservations can be attributed to the nature of the *addition* innovation technique itself, which involves introducing complexity such as variable conditions or value ranges to the original problem. According to Evaluator 2, while this approach promotes critical thinking and makes the problems more engaging, it may also challenge Grade 7 students, especially those who have not yet encountered linear inequalities. For instance, several of the revised problems required students to determine solutions across a range, which, while valid, might be unfamiliar to learners who are only beginning to master linear equations. Readability stayed near the Grade 7 band (mostly 72–74). An important observation relates to the zero-shot prompting output, which, although it received full “Y” marks in all evaluation categories, yielded a significantly lower FRE score (65.21), corresponding to an 8th or 9th-grade reading level. This may suggest that without the benefit of example patterns (as provided in few-shot and chain-of-thought prompts), the model may generate problems that are syntactically more complex or less aligned with the desired cognitive level.

For *modification*, adherence was frequently marked with reservation. It is essential in this technique that the revised versions retain the original given values but ask a different question or present a new problem structure. However, as noted by Evaluator 2: “*I don't think the problem was altered in Revisions 1, 2, 3, and 7. Revisions 4 and 6 mainly involved replacing values, rather than changing the structure of the problem. Also, Revision 4 presents a completely different context.*” This means that although some revisions preserved the original numerical values, they did not significantly modify the problem being asked. For example, in Revision 3 (which used few-shot prompting), the output from ChatGPT was: “*An artist invests in a tax-free bond paying 6%, and she invests \$6000 more than three times the amount she invested in the bond in mutual funds paying 5%. If her total annual interest income from both investments is \$825, how much did she invest in each?*” This is essentially a rewording of the original problem, not a meaningful modification in terms of question structure. Readability was moderately high for both zero-shot and few-shot prompts, which may impact accessibility for some learners. Interestingly, CoT prompting appeared to perform best in aligning with the goals of the modification innovation technique, suggesting that guided reasoning may help generate more effective revisions in the future.

For *contextualization*, APP, SOL, and COMP were positive across prompt types. ADH was often Yes with reservations, except for CoT-1 and CoT-2, which received full Yes. According to Evaluators 2 and 3, the main reservation was that some of the revised problems still used terms that might be unfamiliar to Grade 7 students, such as “tax-free bond” and “mutual funds”. They suggested using more familiar financial terms and converting dollar values to pesos to make the problems more relatable within the local context. One particularly interesting case is the problem generated by COT 1-shot, which received full “Y” ratings across all criteria but had the lowest FRE score (42.83), placing it at college-level readability. The problem was also the longest, yet evaluators found it

appropriately contextualized due to its use of familiar terms and a realistic scenario: "*Maria is a small business owner who invests in two different savings options. She puts her money in a local cooperative bank that offers an annual interest rate of 6%, and she invests ₱6000 more than three times as much in a community investment program offering 5% annual interest. Her total annual interest income from both investments is ₱825. How much does she invest in each savings option?*"

For *inversion* (turning the problem around), no prompt type achieved adherence. Even with examples, the model typically kept the original structure rather than exchanging the values given and what the question is asking for. Yet APP, SOL, and COMP were still Yes, which means the items were teachable and solvable even if the technique failed. Only CoT-2 had an APP with a reservation. FRE values were mainly at Grade 8–9, with zero-shot and CoT-1 closer to Grade 7.

For *reformulation*, adherence was strong (Yes in most rows). Zero-shot and CoT-2 were marked with reservations. APP, SOL, and COMP were often Yes with reservations because variable-only versions increased abstraction for Grade 7. According to Evaluator 2, the reformulated problems, while mathematically valid, were potentially too complex for typical Grade 7 learners due to the replacement of numerical values with variables. This shift from concrete to abstract representation may increase the cognitive demand of students who are still developing algebraic reasoning skills. For instance, solving for unknowns presented purely in variable form requires a level of abstraction that may not be developmentally appropriate for all learners at this stage. Zero-shot also differed in terms of form since it did not replace numbers with variables. All FRE values sat around Grade 8–9, reflecting longer sentences and heavier algebraic language.

4. Discussion

All the authors involved in rewriting selected mathematics problems agreed that the task is quite daunting. This initial exploration is relatively manageable compared to more complex problems that require innovation through the techniques outlined by [2]. Despite the challenge, we collectively found that generative AI tools like ChatGPT are surprisingly effective in supporting teachers with tasks such as rewriting word problems. These tools can serve a range of purposes, including providing students with varied exercises, creating individualized items, designing distinct test questions to minimize copying, and encouraging the development of problem-solving strategies by presenting problems in different forms, such as through modification, inversion, or reformulation.

The study shows that ChatGPT can revise mathematics word problems using different innovation techniques when given clear prompts, but the quality varies by technique and prompt type. The replacement technique produced the most consistent high-quality outputs, meeting all criteria with appropriate readability for Grade 7, which suggests that the task appears to be easy for ChatGPT. The addition technique also adhered to the intended format but often produced problems that were too complex for typical Grade 7 learners due to added conditions. Modification was less successful, as many outputs were simple rewordings or context changes rather than true structural changes. Contextualization outputs were generally appropriate but sometimes used unfamiliar terms, which expert evaluators suggested replacing with locally relevant ones. Inversion was the

Table 2. Expert evaluation and readability Scores for rewritten word problems using the six techniques

| Prompt Type | Replacement | | | | | | Addition | | | | | | Modification | | | | | |
|-------------|-------------|-----|-----|------|--------------|-----|----------|-----|-----|------|--------------|-----|--------------|-----|-----|------|--------------|-----|
| | ADH | APP | SOL | COMP | FRE | GL | ADH | APP | SOL | COMP | FRE | GL | ADH | APP | SOL | COMP | FRE | GL |
| Zero shot | Y | Y | Y | Y | 75.42 | 7 | Y | Y | Y | Y | 65.21 | 8/9 | YW | Y | Y | Y | 58.53 | 8/9 |
| 1 shot | Y | Y | Y | Y | 69.3 | 8/9 | Y | YW | YW | YW | 72.13 | 7 | YW | Y | Y | Y | 58.04 | 8/9 |
| 2 shot | Y | Y | Y | Y | 75.42 | 7 | Y | YW | YW | YW | 74.33 | 7 | YW | Y | Y | Y | 63.37 | 8/9 |
| 3 shot | Y | Y | Y | Y | 60.4 | 8/9 | Y | YW | YW | YW | 70.63 | 7 | Y | Y | Y | Y | 80.58 | 6 |
| COT 1 shot | Y | Y | Y | Y | 75.42 | 7 | Y | YW | YW | YW | 74.33 | 7 | YW | Y | Y | Y | 75.42 | 7 |
| COT 2 shot | Y | Y | Y | Y | 74.81 | 7 | Y | YW | YW | YW | 74.33 | 7 | YW | Y | Y | Y | 77.88 | 7 |
| COT 3 shot | Y | Y | Y | Y | 78.97 | 7 | Y | YW | YW | YW | 72.27 | 7 | YW | Y | Y | Y | 81.39 | 6 |

| Prompt Type | Contextualization | | | | | | Inversion | | | | | | Reformulation | | | | | |
|-------------|-------------------|-----|-----|------|--------------|---------|-----------|-----|-----|------|--------------|-----|---------------|-----|-----|------|-------------|-----|
| | ADH | APP | SOL | COMP | FRE | GL | ADH | APP | SOL | COMP | FRE | GL | ADH | APP | SOL | COMP | FRE | GL |
| Zero shot | YW | Y | Y | Y | 64.91 | 8/9 | N | Y | Y | Y | 71.8 | 7 | YW | YW | Y | YW | 65.1 | 8/9 |
| 1 shot | YW | Y | Y | Y | 57.71 | 8/9 | N | Y | Y | Y | 55.16 | 8/9 | Y | YW | Y | YW | 60.77 | 8/9 |
| 2 shot | YW | Y | Y | Y | 72.27 | 7 | N | Y | Y | Y | 62.13 | 8/9 | Y | YW | YW | YW | 52.58 | 8/9 |
| 3 shot | YW | Y | Y | Y | 57.52 | 8/9 | N | Y | Y | Y | 61.76 | 8/9 | Y | YW | YW | YW | 56.59 | 8/9 |
| COT 1 shot | Y | Y | Y | Y | 42.83 | College | N | Y | Y | Y | 71.28 | 7 | Y | YW | YW | YW | 54.83 | 8/9 |
| COT 2 shot | Y | Y | Y | Y | 63.23 | 8/9 | N | YW | Y | Y | 57.95 | 8/9 | YW | YW | YW | YW | 60.63 | 8/9 |
| COT 3 shot | YW | Y | Y | Y | 58.01 | 8/9 | N | Y | Y | Y | 68.09 | 8/9 | Y | YW | YW | YW | 52.93 | 8/0 |

ADH - Does the revised word problem adhere to the specified innovative type/format?; *APP* - Is the word problem appropriate for Grade 7 students?; *SOL* - Does the word problem have a clear and achievable solution?; *COMP* - Does the word problem align with the math skills and knowledge expected at the Grade 7 level?; *FRE* - Flesch Reading Ease; *GL* - Grade level based on *FRE*; *Y* - Yes, *YW* - Yes with reservations, *N* - No.

the most challenging technique: None of the outputs achieved proper structural reversal despite acceptable ratings for appropriateness and solvability. Reformulation showed strong adherence but often resulted in variable-based problems that were mathematically valid yet too abstract and linguistically demanding for Grade 7 learners.

Interestingly, we did not observe a consistent pattern in the effectiveness of prompt types across all innovation techniques. Some tasks worked well with simpler approaches such as zero-shot prompting, while others benefited more from few-shot or chain-of-thought (CoT) strategies. This does not mean that simpler prompts should always be preferred. For example, in reformulation, only the zero-shot prompt produced a question that did not require symbolic answers. Few-shot and CoT prompts tended to mimic the examples provided, producing variable-based problems. A similar trend appeared in contextualization, where more contextually appropriate results were achieved with more complex prompting strategies. These findings suggest that while prompt complexity does not always lead to better results, advanced prompting can guide ChatGPT to generate more nuanced and pedagogically sound problems.

Follow-up prompts can refine ChatGPT's outputs after initial problem generation. Although our study did not explore iterative prompting, such refinement can often resolve the challenges we observed. For example, experts highlighted complexity issues in addition, unfamiliar contexts in contextualization, and structural alignment in modification or inversion. A follow-up prompt such as "simplify language," "use local context," or "restructure the problem while preserving givens" could correct most of these issues.

5. Conclusion

This study investigated the potential of ChatGPT to support the innovation of mathematical word problems involving linear equations in one variable through various prompt engineering strategies. Results showed that ChatGPT can effectively generate and revise problems using a range of innovative techniques, with output quality largely influenced by the type of prompt used. Techniques such as replacement and contextualization performed well, particularly when guided by few-shot or chain-of-thought prompts, while more complex techniques like inversion and modification posed challenges in maintaining instructional intent and structural integrity. Although expert evaluations and readability scores confirmed the general suitability of several outputs for Grade 7 learners, some techniques raised concerns about age-appropriateness and cognitive demand. Despite these limitations, ChatGPT demonstrated strong potential as a teacher support tool for a more efficient instructional design.

Given the scope and limitations of this study, several future directions are proposed to deepen the understanding of generative AI's role in mathematics education and to expand its potential applications in classroom settings, such as: (1) exploring the use of GenAI in designing word problems for more complex topics such as systems of equations, quadratic functions, or geometry; (2) comparing outputs from different text-based LLMs (Gemini, Claude, DeepSeek, etc.); (3) pilot testing of the word problems to Grade 7 students, and; (4) extending the use of ChatGPT as a teacher support tool in instructional design.

Acknowledgements. The researchers extend their sincere gratitude to the three experts who generously dedicated their time and expertise in evaluating the outputs of this study. Appreciation is also given to the University of the Philippines Rural High School (UPRHS) for its continued support and encouragement throughout the research process.

References

- [1] C. P. Vistro-Yu, "Using innovation techniques to generate 'new' problems," in *Mathematical Problem Solving: Yearbook 2009, Association of Mathematics Educators*, Y. P. Lee, P. Y. Lee, and B. Sriraman, Eds., Singapore: World Scientific, 2009, pp. 185–207.
- [2] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari, "A Systematic Review of Automatic Question Generation for Educational Purposes," *Int. J. Artif. Intell. Educ.*, vol. 30, no. 1, pp. 121–204, Mar. 2020, doi: 10.1007/s40593-019-00186-y.
- [3] Z. Wang, A. S. Lan, and R. G. Baraniuk, "Math Word Problem Generation with Mathematical Consistency and Problem Context Constraints," Sept. 09, 2021, *arXiv*: arXiv:2109.04546. doi: 10.48550/arXiv.2109.04546.
- [4] L. Boonstra, "Prompt Engineering: Empowering Large Language Models for Real-World Applications," Google Cloud, v4, Jan. 2024. [Online]. Available: <https://cloud.google.com/blog/products/ai-machine-learning/prompt-engineering-empowering-large-language-models>
- [5] E. Chen, D. Wang, L. Xu, C. Cao, X. Fang, and J. Lin, "A Systematic Review on Prompt Engineering in Large Language Models for K-12 STEM Education," Oct. 14, 2024, *arXiv*: arXiv:2410.11123. doi: 10.48550/arXiv.2410.11123.
- [6] M. Son, Y.-J. Won, and S. Lee, "Optimizing Large Language Models: A Deep Dive into Effective Prompt Engineering Techniques," *Appl. Sci.*, vol. 15, no. 3, Art. no. 3, Jan. 2025, doi: 10.3390/app15031430.
- [7] K. Addala, K. D. P. Baghel, C. Kirtani, A. Anand, and R. R. Shah, "Steps are all you need: Rethinking STEM Education with Prompt Engineering," Dec. 23, 2024, *arXiv*: arXiv:2412.05023. doi: 10.48550/arXiv.2412.05023.
- [8] K. Norberg *et al.*, "Rewriting Math Word Problems with Large Language Models," 2023.
- [9] S. Schorcht, N. Buchholtz, and L. Baumanns, "Prompt the problem – investigating the mathematics educational quality of AI-supported problem solving by comparing prompt techniques," *Front. Educ.*, vol. 9, May 2024, doi: 10.3389/educ.2024.1386075.
- [10] K. W. Chan *et al.*, "Automatic item generation in various STEM subjects using large language model prompting," *Comput. Educ. Artif. Intell.*, vol. 8, p. 100344, June 2025, doi: 10.1016/j.caeai.2024.100344.
- [11] S. C. E. Fung, M. F. Wong, and C. W. Tan, "Chain-of-Thoughts Prompting with Language Models for Accurate Math Problem-Solving," in *2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*, Oct. 2023, pp. 1–5. doi: 10.1109/URTC60662.2023.10534945.
- [12] M. L. Lial, J. Hornsby, and T. McGinnis, *Beginning Algebra*, 11th ed. Boston: Pearson Education, Inc., 2012.
- [13] J. A. Marasigan, *Experiencing Mathematics (Math XP) Series I: Introductory Algebra*. Philippines: Vibal Publishing House, Inc., 2006.
- [14] T. Berg, "Mixture and Solution Word Problems," in *Intermediate Algebra*, BC Campus Open Education, 2020.
- [15] W. Zhao *et al.*, "Assessing the adherence of large language models to clinical practice guidelines in Chinese medicine: a content analysis," *Front. Pharmacol.*, vol. 16, p. 1649041, July 2025, doi: 10.3389/fphar.2025.1649041.
- [16] R. Olszewski, K. Watros, M. Mańczak, J. Owoc, K. Jeziorski, and J. Brzeziński, "Assessing the response quality and readability of chatbots in cardiovascular health, oncology, and psoriasis: A comparative study," *Int. J. Med. Inf.*, vol. 190, p. 105562, Oct. 2024, doi: 10.1016/j.ijmedinf.2024.105562.