Using ChatGPT 3.5 to Reformulate Word Problems for State Exam in Mathematics

Piret Luik[1][a], Carmen Keivabu[2] and Kerli Orav-Puurand[3][b]
[1]*Institute of Computer Science, University of Tartu, Narva mnt 18, Tartu, Estonia*
[2]*Väike-Maarja Upper-secondary School, Pikk 1, Väike-Maarja, Lääne-Viruma, Estonia*
[3]*Institute of Mathematics and Statistics, University of Tartu, Narva mnt 18, Tartu, Estonia*

Keywords: Mathematics, ChatGPT, Learning Outcomes, Evaluations.

Abstract: Although there are already studies on the use of large language models in education, the possibilities and potential of this are still not clear. Therefore, the purpose of this article is to find out whether ChatGPT is an effective tool for teachers in simplifying mathematical word problems, thereby improving students' learning outcomes. The study was conducted with 26 students who solved four state exam textual problems that were worded more clearly for the students, and then the students solved these reformulated problems a week later. Besides the tests a questionnaire of yes-no questions was used to obtain students ratings on the tasks. All the tasks were assessed according to the criteria of the assessment manuals of the state exam in mathematics The main results of the study showed that the students had overall better results with the task tests reformulated using ChatGPT compared to the state exam tasks. Based on the ratings, the students found some tasks to be clearer in the state exam tasks, while other tasks were more understandable to them in the versions reformulated by ChatGPT. In conclusion, ChatGPT has the potential to support mathematics teaching, but its effective use requires careful wording of tasks.

1 INTRODUCTION

Mathematics is a subject taught in general education schools from the first to the last grade and is necessary in everyday life as well as in other subjects. However, many students do not like mathematics, which is due to several factors including teaching methods, students' thinking ability, and understanding of the content of the subject (Gafoor & Kurukkan, 2015). A large number of students experience difficulties in solving mathematical word problems (also known as textual problems) due to the complex mathematical language (Sepeng & Madzorera, 2014). Similarly, the results of the OECD Program for International Student Assessment (PISA) (OECD, 2023) have shown that many students have difficulties solving mathematical word problems, especially understanding complex language and contexts. Word problems are defined as "verbal descriptions of problem situations wherein one or more questions are raised the answer to which can be obtained by the application of mathematical operations to numerical data available in the problem statement" (Verschaffel, et al., 2000). For example, a word problem is: Mary has 30 pencils, Lizzy has 17 fewer pencils. How many pencils do the girls have in total? In that case, student has to comprehend the problem, know what means less, total, write mathematical operation 30-17+30 and to solve this operation. Solving mathematical word problems is difficult because its main challenge is reading comprehension, i.e. understanding the problem and choosing the necessary solution strategies (Sepeng & Madzorera, 2014). Unlike everyday language, mathematical language is a language of symbols, concepts, definitions and theorems, and it causes difficulties for students (Haerani, et al., 2021; Ilany & Margolin, 2010). Students' struggles in solving mathematics word problems are also related to various background factors, including prior knowledge, motivation and social background (Langoban, 2020).

[a] https://orcid.org/0000-0002-9049-4192
[b] https://orcid.org/0009-0005-3829-5769

Luik, P., Keivabu, C. and Orav-Puurand, K.
Using ChatGPT 3.5 to Reformulate Word Problems for State Exam in Mathematics.
DOI: 10.5220/0013152200003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 1*, pages 180-190
ISBN: 978-989-758-746-7; ISSN: 2184-5026

Artificial Intelligence (AI) has spread to all areas of human activity, bringing significant changes and new scientific and ethical challenges. Education is not an exception in this development. ChatGPT (Chat Generative Pre-trained Transformer), a Large Language Model (LLM) disclosed by OpenAI, has attracted great interest in the possible applications of this technology in education (Karaca, 2024; Matzakos et al., 2023). The use of ChatGPT in mathematics learning offers significant benefits to both students and teachers (Manik, 2024). Generative Pre-trained Transformer (GPT) models, including GPT-3.5, are primarily designed for Natural Language Processing (NLP) tasks such as language generation, translation, and question answering. Although GPT models can perform simple arithmetic operations and recognize mathematical symbols and expressions, they are not specifically designed or optimized to solve complex mathematical problems (Wardat et al., 2023). Since existing language models, including ChatGPT, are not designed specifically for education, and even less so for mathematics, it is important to understand how language models can be effectively applied in education. It is especially important to find solutions to simplify mathematical word problems, which are often difficult for students.

2 BACKGROUND

2.1 Mathematical Language

Identifying the constituents of a text depends on an awareness of the role of form, words, and sentences in a text, especially knowledge of symbols and syntax. Syntax generally deals with the rules by which sentences and words are constructed (MacGregor & Price, 1999). The syntax of a mathematical language includes lists of symbols, rules for constructing language patterns, axioms, a deductive system, and theorems. Mathematical terms and symbols must be clearly defined. Each statement in a mathematical language is unambiguous - each mathematical pattern has a single definite structure determined by operational rules (Ilany & Margolin, 2010). In such a way, the language of mathematics is different from everyday or natural language. While in everyday language there is a word or pair of words for almost every object, in the language of mathematics the concepts are often more abstract and specific. Despite that, the results of studies indicate that good reading skills help students cope better with mathematical texts (Grimm, 2008; Ünal et al., 2023). Ilany and Margolin (2010) also argue that there is a bridge between mathematical language and natural or everyday language, and therefore knowledge gaps between mathematical language and natural language become particularly apparent when solving word problems.

Boulet (2007) emphasizes that the integrity of mathematical language is often related to three components: mathematical words (concepts), symbols and numbers. These three together define a mathematical language. Mathematical language relies on mathematical vocabulary, which includes technical terms (definable concepts), symbols, non-technical terms (non-definable concepts), and ambiguous words. These components of mathematical language create numerous challenges for students when solving problems related to tasks, especially word problems (Schleppegrell, 2007), because mathematical language needs to be learned and does not develop intuitively like natural language (Ilany & Margolin, 2010). When solving mathematical problems accompanied by text, the student is faced with two mixed languages: natural or everyday language and mathematical language. The main differences between natural language and mathematical language are primarily due to the fact that the structure of mathematical language is more precise and less flexible than that of natural language (Kane, 1970). In natural language, there are words that can be ambiguous, but in mathematical language, terms and symbols must be defined unambiguously so that every mathematical pattern has one deep structure that is determined by operational rules (Ilany & Margolin, 2010).

When solving textual tasks, reading and understanding the text content are very important and, in addition to the linguistic properties of the task text, reading competence also plays an important role in the solving process (Stephany, 2021). The linguistic structure of a text affects its comprehension, as the reader may be more familiar with certain language constructs, such as sentence structure and terminology, which better correspond to their expectations and experiences (Kintsch, 1994). In addition, Davis-Dorsey et al. (1991) have found that reframing word problems so that meaningful connections become clearer, such as comparisons in the text, improves their comprehension and solving process.

2.2 Language Models

Artificial intelligence is a branch of computer science that focuses on computer programs that can mimic human thinking abilities, such as teaching and

decision-making (Tashtoush, et al, 2024), relying on data, algorithms, and computer power to learn from experience, patterns, and make judgments (Opesemowo & Ndlovu, 2024). As part of the AI models, Large Language Models (LLM) have emerged in the field of computer-based language processing. These models are able to understand complex linguistic patterns and generate clear and context-appropriate responses, being effective tools for many tasks, including natural language processing, machine translation, and question answering (Hadi, et al., 2023).

The first language models were created in the 1950s and 1960s based on a set of rules and they used manually created grammar rules and properties to process the language, but they were limited in their capabilities and could not handle the complexity of natural language processing (Liddy, 2001). The development of language models has taken place in four stages: statistical models (Statistical Language Models, SLM), neural network models (Neural Language Models, NLM), contextualized word representations (Pre-trained Language Models, PLM) and large language models (Hadi, et al., 2023). The current LLMs include two particularly prominent language models: BERT (Bidirectional Encoder Representations from Transformers) by Google and GPT developed by OpenAI (Dao, 2023). OpenAI's ChatGPT (abbreviation of 'Chat Generative Pretrained Transformer'), based on GPT-3.5, is an excellent example of the capabilities of today's language models, providing answers and conversations resembling ordinary language (Hadi, et al., 2023). Following the success of GPT-3, OpenAI developed successor models such as InstructGPT, Codex, ChatGPT, and GPT-4 and the evolution continues (Kalyan, 2024).

A variety of LLMs have appeared: ChatGPT, Microsoft's Bing Chat (now Copilot), and Google's Bard (now Gemini). Chat GPT is popular for its user-friendliness and easy availability, offering a free GPT-3.5 version and a paid GPT-4 version, whereas Bing Chat with its limited access and browser compatibility may be less suitable for daily use and Google Bard is still in the early stages of development (Giannakopoulos et al., 2024). ChatGPT is capable of generating coherent, partially accurate, systematic and informative responses that integrate and preserve the topic and history of the conversation (Zhai, 2022) using AI and natural language processing (Opesemowo & Ndlovu, 2024) being capable of imitating human-like conversations (Lin, et al., 2023). ChatGPT can be asked for data, analysis and even opinions, and is able to continuously maintain a dialogue style that engages the user in a more natural way (Rahman & Watanobe, 2023).

2.2.1 Large Language Models in Education

The rapid development of AI including LLMs transforms education among many other sectors (Opesemowo & Ndlovu, 2024) and affects millions of students and teachers (Rudolph et al., 2023). Farrokhina et al. (2023) analysed different sources, using the SWOT analysis framework, to identify the strengths, weaknesses, opportunities and threats of ChatGPT in education. They declare that the strengths of Chat GPT include generating plausible answers, providing real-time personalized responses and self-improving capability, while the opportunities are increase of accessible information, decrease of teachers' workload, and facilitation of personalized and complex learning. However, on the negative side of the ChatGPT, they mention lack of deep understanding, risk of biases and high-order thinking skills and difficulties with evaluating the quality of the responses as weaknesses, whereas threats include lack of understanding of the context, undermining of academic integrity, plagiarism, and decline of higher-order cognitive skills (Farrokhnia et al., 2023). Using ChatGPT, teachers can create different forms of tests (Zhai, 2022), generate lesson plans, questions and answers for educational purposes (Yang, 2023). Kim and Adlof (2024) emphasize that teachers should use ChatGPT as a tool in creating a learning environment rather than as an end in itself.

The use of AI is becoming increasingly common in engineering education and is also important in mathematics education (Opesemowo & Ndlovu, 2024; Vintere et al., 2024). One promising development in this area is the use of technology such as the ChatGPT language model (Wardat et al., 2023). Although LLMs such as ChatGPT can perform mathematical calculations, they may not always be as accurate or efficient as dedicated mathematical software or hardware (Soygazi & Oguz, 2024; Wardat et al., 2023). From six presented word problems ChatGPT gave correct solutions to four, but an alternative NLP framework, LangChain, only to one (Soygazi & Oguz, 2024). However, ChatGPT can effectively explain mathematical theorems and concepts in an easy to understand language that is suitable for students, but sometimes the sentences are too long (Wardat et al., 2023). ChatGPT could also generate mathematical questions, which are highly relevant to the input context, but frequently repeats information from the context which leads to the generation of lengthy questions (Pham et al., 2024).

A LLM provides innovative methods for teaching and learning, as well as personalized learning experiences and intelligent instruction (Gan, et al., 2023; Kasneci, et al., 2023; Yang, 2023). However, in the study by Pham et al. (2024) it was revealed that if more complex questions need to be generated in math, ChatGPT often replicates the initial demonstration instead of increasing the difficulty of the question.

The first step of the mathematical word problems solving process is reading and comprehension (Çetġnkaya, et al., 2018; Sepeng & Madzorera, 2014). Readability is a measure of how easy a piece of text is to read and several readability formulas have been developed specifically to every language to determine readability level of texts (Çetġnkaya, et al., 2018). Most readability formulas use sentence length, word length, word familiarity, number of terms and/or word abstractness to calculate readability score of the text (Mikk, 2000). Comprehensibility is the extent to which the text as a whole is easy to understand and it is related to readability. Lower readability scores of texts in mathematical tasks indicated these texts might be less comprehensible for students (Çetġnkaya, et al., 2018).

Karaca (2004) used readability formulas in his study to calculate readability values of stories created with LLM. It was found that ChatGPT 3.5 and Gemini generated more appropriate stories for educational purposes, with the average readability of the stories increasing from easy to difficult based on the educational level. Using ChatGPT 3.5 the average number of words in text sentences for the upper-secondary level was 9.59, the average number of letters per word was 6.23 and the stories created were at a medium level of difficulty (Karaca, 2004).

Previous research has also explored the benefits, limitations and challenges of using LLMs in education. They emphasize the importance of well-defined strategies, critical thinking skills (Kasneci, et al., 2023; Opesemowo & Ndlovu, 2024), and ethical considerations (Baidoo-Anu & Ansah, 2023) including using AI to complete assignments and cheat on exams (Yang, 2023). Concerns about academic integrity and privacy violations are also addressed (Dao, 2023; Opesemowo & Ndlovu, 2024). In mathematics, however, ChatGPT can talk about the subject but lacks a deeper understanding of it (Wardat et al., 2023). Also, the results of the study by Rane (2023) show that although ChatGPT can provide useful solutions and explanations, there are challenges when it comes to creating accurate mathematical drawings and providing logical explanations.

3 GOAL AND RESEARCH QUESTIONS

As described above, several studies have shown the potential of using LLMs in education, including mathematics. As word problems in mathematics are difficult for learners, the use of ChatGPT can help simplify learning material for students by formulating word problems in a different way. Therefore, the aim of this study was to find out to what extent students' learning outcomes change when word problems from the state exam in mathematics are reformulated with hints using ChatGPT and how is the wording rated by students.

Three research questions were posed:

1. How does ChatGPT 3.5 reword mathematical word problems and what hints does it add?
2. What are the differences in results between the word problems used in the state exam and those reformulated with hints by ChatGPT?
3. How do students rate the wording of the word problems reformulated by ChatGPT compared to the word problems of the exam?

4 METHODOLOGY

4.1 Sample

The sample consisted of 11th grade (age from 17 to 18) students of two upper-secondary schools. The first school is located in an urban area and 15 students (7 boys and 8 girls) from this school participated. The second school is a rural school and 11 students (4 boys and 7 girls) participated in the study. A total of 26 students participated and all of them studied mathematics according to the same syllabus and had the same teacher, which ensured that they had a uniform background and comparable knowledge in the field under study. There were no students with special educational needs among the participants.

All parents and students were informed about the study. It was emphasized that the study is voluntary and the results would be presented only in a generalized form without linking them back to specific learners. It was also explained that the study is conducted in a mathematics class with the consent of the teacher and the students' participation will not affect the evaluation of their performance in mathematics.

4.2 Procedure

At first, four word problems from the tasks of the mathematics state exam were selected. The text content of the chosen problems included topics that the students had learned during the period of the study: speed/time/path length, system of equations, arithmetic sequence, and probability. The texts of the word tasks in these problems had to correspond to life situations and contained foreign words (for example, duathlon, tariff, etc.), mathematical terms (average, sector, etc.), question sentences, narrative sentences, and at least one drawing. These four word problems from the state exam formed Test A and, at first, the participating students solved Test A with the word problems from the state exam. The tests contained detailed information about both the study and the assignments to ensure the students had a clear understanding of their role.

The selected word problems were then reformulated with the ChatGPT 3.5 language model. The first prompt was as follows:

Word problems in mathematics are difficult for students to understand and cannot be solved. Difficulties are caused by long sentences, a lot of text, unfamiliar concepts. Please, taking this knowledge into account, reword the text of the math word problems so that the mathematical concepts in the text and also the more complex words are understandable to an upper-secondary level student and that the text is not ambiguous for the students.

As ChatGPT also added solutions to the problems, the next prompt was given:

Can you leave out the solution to the problem or replace the solution with hints?

These four reformulated word problems formed Test B, which was solved by students a week after solving Test A. Brown et al. (2008) claim that after 3 weeks students have forgotten nearly all of the test questions. In our case the wording was not the same and it was asked students have they solved the same test previously. All students answered ‘no’ meaning that they did not associated the test A and test B problems. In both instances they had 45 minutes to solve the tasks. Both tests were printed on A4 sheets and the students solved the problems on paper.

The tasks were assessed by the teacher, who is the second author of this paper according to the criteria of the assessment manuals of the state exam in mathematics. The first three problems had 10 points as the maximum score obtainable and up to 5 points could be awarded for successfully solving the last problem.

For answering the last research question, a questionnaire of yes-no questions was used. The questionnaire was created in the online Google Forms environment. After solving Test B, the questionnaire link was distributed to the students. The questionnaire was filled in by students using smart devices.

4.3 Data Analysis

The collected data were entered and organized in the MS Excel program. This involved entering the results of the students' paper solutions as well as coding the questionnaire responses. When the results of Test A and Test B and the questionnaire responses were entered, the names of the students were deleted from the data table.

The data was analysed using the program IBM SPSS Statistics 29.0.2.0. Descriptive statistics were used to perform a statistical analysis where the average results and standard deviations of the problem solutions were calculated for both the original and reformulated problems. As the variables were not normally distributed, a non-parametric Wilcoxon test was performed to assess whether there were statistically significant differences between the exam problems and those reformulated by ChatGPT. In addition, the chi-square test was used to analyse the differences in proportions.

5 RESULTS

5.1 Reformulated Word Problems

The process of reformulating the word problems and adding hints resulted, in most cases, in doubling the length of the original word problem (see Table 1). However, in the first two problems the sentences were shorter after reformulation than in the state exam.

In the first two word problems Chat GPT replaced the terms and explained the mathematical concepts. In the last two problems the problem itself was not reformulated, but hints and formulas were given.

Problem 1: The state exam problem required knowledge of the speed formula, understanding of the foreign word ‘duathlon’, knowledge of mathematical terms such as ‘smaller by 3’, ‘2 more’, ‘in half an hour’. In the reformulated problem the speed formula was given to students; the term ‘duathlon’ was replaced by a description that the boys participated in a competition where they had to run and ride a bike; the hint explained that ‘3 less’ requires subtraction and ‘2 more’ means addition; the phrase ‘30 minutes’ was used instead of ‘half an hour’.

Table 1: Characteristics describing the length of the word problems.

	Problem 1	Problem 2	Problem 3	Problem 4
Number of letters (SE)	430	395	562	515
Number of letters (CGPT)	905	717	1517	1148
Number of words (SE)	75	60	101	89
Number of words (CGPT)	175	108	274	190
Number of sentences (SE)	6	4	10	7
Number of sentences (CGPT)	15	10	20	14
Average number of letters per word (SE)	5.7	6.6	5.6	5.8
Average number of letters per word (CGPT)	5.2	6.6	5.5	6.0
Average word count per sentence (SE)	12.5	15.0	10.1	12.7
Average word count per sentence (CGPT)	11.6	10.8	13.7	13.6

SE – state exam
CGPT – ChatGPT

Problem 2: The state exam problem was about electricity consumption using specific concepts such as 'electricity package', 'day and night tariff' and 'kilowatt-hour', and requiring knowledge of mathematical terms such as 'smaller by 40' and 'twice as small'. To solve the problem, it was necessary to create and solve equations, calculate proportions and convert monetary units. In the reformulated problem the terms 'electricity package' and 'consumption' remained, but instead of the term 'day tariff' the phrase 'the price during the day' was used, and the night tariff was explained similarly. As a hint, it was said that it is necessary to prepare two equations and the variables in these equations were named. No information was given on the meaning of 'smaller by 40' or 'twice as small', nor about converting monetary units.

Problem 3: Solving the state exam problem required knowledge about scoring of the game using the terms 'successful move', 'failed move' and 'point', and knowledge of mathematical terms such as '5 more points', '3 points are deducted', '0.5 points more than last time'. It was necessary to understand and apply an arithmetic sequence and create and solve equations. In addition, the task required analytical skills to calculate the change in score. ChatGPT did not reword the problem, but added some hints: "if 3 points are deducted after the first unsuccessful move, 3.5 points after the second, etc", and advice on the steps of the solutions.

Problem 4: This state exam problem included a drawing and was about the calculation of probability using concepts such as 'sector' and 'probability'. To solve the task, it was necessary to apply mathematical skills, such as probability calculation, combinatorics and logical analysis. ChatGPT did not reword the problem, but provided a description of the drawing: "The first wheel of fortune has four sectors and the second wheel has three sectors", and the formula for calculating probability. However, the mathematical concept 'sector' was not explained. In addition to the description of drawing some hints were provided. For example, "By spinning two wheels at the same time, the player gets a score by adding up the points from both wheels."

5.2 Comparison of Test Scores

There was statistically significant difference only in the case of the last problem and in the total test score (see Table 2). In both cases students received significantly higher scores in Test B.

Table 2: Comparison of test scores.

	Test A M (SD)	Test B M (SD)	Z	p
Problem 1	3.6 (1.53)	3.6 (1.10)	0.140	0.913
Problem 2	0.2 (0.43)	0.5 (0.76)	- 1.490	0.127
Problem 3	3.7 (3,73)	4,7 (3.26)	-1.590	0.116
Problem 4	1.7 (1.46)	3.1 (2.08)	-3.051	0.002
Total score	9.2 (4.83)	11.8 (4.42)	-2.714	0.007

With all problems, there were students whose scores decreased in Test B and those whose scores increased in Test B compared with Test A, and some students who received equal scores in Tests A and B (see Table 3). Comparing the proportion of students who received higher score in Test A with those who received higher score in Test B, there was a statistically significant difference only in the case of the fourth word problem (chi-square=12.613, p<.001). In addition, 77% of students received a higher total score in Test B and there was a statistically significant difference compared to the proportion of students who received a higher score in Test A (chi-square=19.731, p<.001).

Table 3: Students' score change between Tests A and B.

	Higher score in Test A n (%)	Equal scores n (%)	Higher score in Test B n (%)
Problem 1	8 (31%)	13 (50%)	5 (19%)
Problem 2	4 (15%)	14 (54%)	8 (31%)
Problem 3	7 (27%)	7 (27%)	12 (46%)
Problem 4	2 (8%)	10 (38%)	14 (54%)
Total score	4 (15%)	2 (8%)	20 (77%)

5.3 Students' Ratings

The students were asked were the word problems more understandable in Test A or in Test B or some in Test A and some in Test B. In total, 16 (62%) of the participating students answered in the questionnaire that some word problems were more understandable in Test A and some in Test B. Students were asked about each problem: did it contain incomprehensible words, sentences or was the whole text incomprehensible, and was it easy to read (Figure 1)?

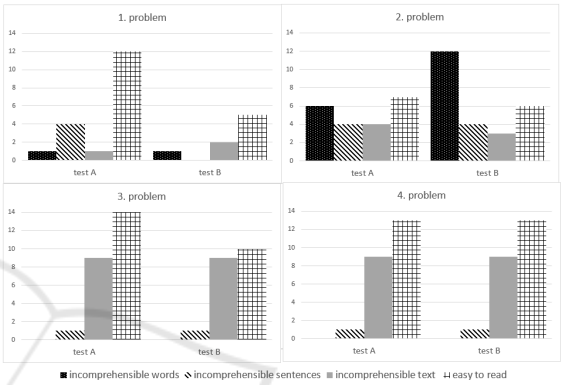

Figure 1: Students' ratings on the wording of the problems in Tests A and B (sample size 26).

The proportion of students who marked that Test A included incomprehensible sentences was statistically significant in comparison to Test B only in the case of the first word problem (chi-square=4.135, p=.042). At the same time, a statistically greater number of students rated that the first word problem in Test A was easier to read than in Test B (chi-square=4.237, p=.040). There were no other statistically significant differences between the ratings of Test A and Test B (in all cases p>.05).

As ChatGPT added hints, students were also asked if the provided hints were useful or confusing. With all the problems, there were students who marked that the hints were helpful and others who felt that the hints confused them (Table 4).

Table 4: Students' ratings on hints in Test B.

	Problem 1	Problem 2	Problem 3	Problem 4
Hints were helpful	7 (27%)	2 (8%)	5 (19%)	2 (8%)
Hints were confusing	8 (31%)	8 (31%)	5 (19%)	6 (23%)

6 DISCUSSION

The aim of this study was to find out to what extent the reformulation of the word problems from the state exam in mathematics with hints provided by ChatGPT changes students' learning outcomes and how students rate the rewording. Three research questions were posed.

As the first research question it was analysed how ChatGPT 3.5 rewords mathematical word problems and what hints does it add. It was interesting that, despite similar prompts, only the first two word problems were reworded. The text in the last two problems remained the same, and only hints were added. Similarly to the study by Pham et al. (2024) that found that ChatGPT often replicates the initial question instead of enhancing the question's difficulty, we can say the same about simplifying word problems. Nevertheless, the mathematical concept 'sector' in the last problem was explained by describing the drawing in the hint. Everyday concepts 'successful move' and 'failed move' in the third problem were not explained. Maybe it was assumed by the ChatGPT that these concepts are understandable for higher-secondary students, as concepts like 'duathlon' or 'daily rate' in the first two problems were replaced with the explanation. Using a different wording instead of concepts might help to solve the problems better, as it has been noted that understanding the text content plays an important role in the solving process (Stephany, 2021).

It was found that reformulated problems were longer. In addition, comparing the average number of letters per word showed that rewording by ChatGPT did not decrease the length of the words. However, in the case of the first two problems, the average number of words per sentence decreased, but the sentences were still longer than in the Karaca (2024) study. Therefore, similarly to the previous studies (Pham et al., 2024; Wardat et al., 2023), we conclude that the text in the reworded problems was too long and even sentences were shortened only in the first two word problems.

Answering the second research question about the differences in results between the word problems in the state exam and those reformulated with hints by ChatGPT, it was found that the total test score was statistically significantly higher in the test where the word problems were reformulated with hints by ChatGPT as opposed to the test with the original state exam word problems. Therefore, based on the comparison of the total scores of the test, it can be concluded that the reformulated word problems were easier for the students to solve. One possible explanation for the higher score of the reformulated tests in our study might be that the hints helped create meaningful connections in the text. For example, it was explained that '3 less' requires subtraction and '2 more' means addition. The solving process can be improved by reframing word problems so that clearer meaningful connections are created (Davis-Dorsey et al., 1991). Our result also supports the conclusion of Zhai (2022) that ChatGPT helps teachers create different forms of tests.

However, comparing the individual word problems, a statistically significant difference was found only in the case of the fourth word problem. It was an interesting result because the problem wording itself remained the same. More than half of the students improved their score in Test B. This problem was the only one with a drawing. As the content of the drawing was explained in one of the hints of the reworded problem, it might have helped with solving the problem. In the last problem the formula for calculating probability was provided in the hints but, similarly, the speed formula was given as one of the hints in the first problem and it did not help students there. As previous studies using ChatGPT in mathematical education have focused mainly on the possibility of using a LLM for solving mathematical problems (e.g. Rane, 2023; Soygazi & Oguz, 2024; Wardat et al., 2023), this kind of comparison is a novel understanding.

The results regarding the last research question about students' ratings on the wording of the word problems reformulated by ChatGPT compared to the word problems of the exam showed that, as expected, some word problems were more understandable in the original form and others after reformulation. In addition, the added hints can be helpful for some students while creating confusion in others. However, comparing the results of this and the previous research question, it can be concluded that students do not always perceive how comprehensible a word problem is. Even though significantly more students found some of the sentences in the state exam version of the first problem to be incomprehensible, there was no marked improvement in solving after the problem was reworded. On the other hand, rewording seemed to be helpful for solving the last word problem, but there were no statistically significant differences between the ratings of Test A and Test B.

7 CONCLUSIONS

Despite a growing trend of AI studies in education, the research on using LLMs in this field is still

limited. This study adds novel findings on how reformulation of mathematical word problems with hints using ChatGPT can change students' learning outcomes and how students rate the rewording. The main results of the study showed that, although there was no statistical difference in the results of the individual tasks, except for the fourth, there was a statistical difference in the aggregate results between the state exam tasks and the tasks reformulated by ChatGPT. The tasks reformulated using ChatGPT were overall more understandable and easier for students to solve than the exam tasks. The ratings revealed that more than half of the students who participated in the research found the wording of the state exam to be more comprehensible in some tasks and the reformulation by ChatGPT in others.

Based on the results of the study, the ChatGPT language model can be recommended to teachers for reformulating tasks or adding hints to tasks for students with learning disabilities. Students may also be advised to use a LLM to simplify more complex formulations.

The first limitation of the study is the small sample size, which prevents the results from being generalised. Also, only four textual tasks were used in the study and, despite the one-week time gap, the order in which the tasks were presented could also affect the results. Although there were some who received a worse result in all the tasks the second time, a different order of presenting the tasks could be used the next time. It was also not possible to use a control group in this study, which would have increased the reliability of the results. Therefore, an experiment with a control group could be conducted in future studies.

ACKNOWLEDGEMENTS

This work was supported by the Estonian Research Council grant "Developing human-centric digital solutions" (TEM-TA120).

REFERENCES

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *Journal of AI, 7*(1). http://dx.doi.org/10.2139/ssrn.4337484

Boulet, G. (2007). How Does Language Impact the Learning of Mathematics? Let Me Count the Ways. *Journal of Teaching and Learning 5*(1).

Brown, G., Irving, E., & Keegan, P. (2008). *An Introduction to Educational Assessment, Measurement and Evaluation*. Auckland, NZ: Pearson Education.

Çetġnkaya, G., Aydoğan Yenmez, A., Çelġk, T., & Özpinar, I. (2018). Readability of Texts in Secondary School Mathematics Course Books. *Asian Journal of Education and Training,* 4(4), 250-256. https://doi.org/10.20448/journal.522.2018.44.250.256

Dao, X. Q. (2023). Which large language model should you usein Vietnamese education: ChatGPT, Bing Chat, or Bard?. *SSRN Electronic Journal.* http://dx.doi.org/10.2139/ssrn.4527476

Davis-Dorsey, J., Ross, S., & Morrison, G. (1991). The Role of Rewording and Context Personalization in the Solving of Mathematical Word Problems. *Journal of Educational Psychology*, 83(1), 61-68. https://doi.org/10.1037/0022-0663.83.1.61

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Gafoor, A. K., & Kurukkan, A. (2015). Why High School Students Feel Mathematics Difficult? An Exploration of Affective Beliefs. Online Submission, Paper presented at the UGC Sponsored National Seminar on Pedagogy of Teacher Education, Trends and Challenges (Kozhikode, Kerala, India, Aug 18-19, 2015).

Gan, W., Qi, Z., Wu, J., & Lin, J.-W. (2023). Large Language Models. In *Education: Vision and Opportunities. IEEE International Conference on Big Data*.

Giannakopoulos, K., Kaklamanos, E. G., & Makrygiannakis, M. A. (2024). Evidence-based potential of generative artificial intelligence large language models in orthodontics: a comparative study of ChatGPT, Google Bard, and Microsoft Bing. *European Journal of Orthodontics, 46*. https://doi.org/10.1093/ejo/cjae017

Grimm K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology, 33*(3), 410-26. https://doi.org/10.1080/87565640801982486

Hadi, M. U., Al Tashi, Q., Qureshi, R., Shah, A., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Hassan, S. Z., Shoman, M., Wu, J., Mirjalili, S., & Shah, M. (2023). *Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects*. https://www.techrxiv.org/users/618307/articles/682263-large-language-models-a-comprehensive-survey-of-its-applications-challenges-limitations-and-future-prospects

Haerani, A., Novianingsih, K., & Turmudi, T. (2021). Analysis of Students' Errors in Solving Word Problems Viewed from Mathematical Resilience. *Journal of Physics: Conference Series, 1157*(4).

Ilany, B.-S., & Margolin, B. (2010). Language and Mathematics: Bridging between Natural Language and

Mathematical Language, *Creative Education, 1*(3), 138-148. https://doi.org/10.4236/ce.2010.13022.

Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6. https://doi.org/10.1016/j.nlp.2023.100048.

Kane, R. B. (1970). The readability of mathematics textbooks revisited. *The Mathematics Teacher, 63*(7), 579-581.

Karaca, M. F. (2024). Is Artificial Intelligence able to Produce Content Appropriate for Education Level? A Review on ChatGPT and Gemini. In *Proceedings of the Cognitive Models and Artificial Intelligence Conference (AICCONF '24)*. Association for Computing Machinery, New York, NY, USA, 208–213. https://doi.org/10.1145/3660853.3660915.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103,* https://doi.org/10.1016/j.lindif.2023.102274

Kim, M., & Adlof, L. (2024). Adapting to the Future: ChatGPT as a Means for Supporting Constructivist Learning Environments. *TechTrends 68*, 37–46.

Kintsch, W. (1994). Text Comprehension, Memory, and Learning. *American Psychologist, 49*(4). 294–303. https://doi.org/10.1037/0003-066X.49.4.294

Langoban, M. (2020). What Makes Mathematics Difficult as a Subject for most Students in Higher. *International Journal of English and Education, 9*(3), 214-220.

Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science*, 2nd Ed. NY. Marcel Decker, Inc.

Lin, C.-C., Huang, A. Y. Q., & Yang, S. J. H. (2023). A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022). *Sustainability*, *15*(5), 4012. https://doi.org/10.3390/su15054012

MacGregor, M., & Price, E. (1999). An exploration of aspects of language proficiency and algebra learning. *Journal for Research in Mathematics Education, 30*, 449-467. doi:10.2307/749709

Manik, E. (2024). The Role of the Teacher Taken by ChatGPT. *International Journal of Advanced Technology and Social Sciences, 2*(1), 1-10.

Matzakos, N., Doukakis, S., & Moundridou, M. (2023). Learning Mathematics with Large Language Models: A Comparative Study with Computer Algebra Systems and Other Tools. *International Journal of Emerging Technology in Learning, 18*(20), 51-71. https://www.learntechlib.org/p/223774/

Mikk, J. (2000). *Textbook: Research and Writing*. Peter Lang, Europäiser Verlag der Wissenschaften, Frankfurt am Main.

OECD (2023), *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en.

Opesemowo, O. A. G. & Ndlovu, M. (2024). Artificial intelligence in mathematics education: The good, the bad, and the ugly. *Journal of Pedagogical Research, 8(*3), 333-346. https://doi.org/10.33902/JPR.202426428

Pham, P. V. L., Duc, A. V., Hoang, N. M., Do, X. L., & Luu, A. T. (2024). ChatGPT as a math questioner? Evaluating ChatGPT on generating pre-university math questions. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing (SAC '24)* (pp. 65–73). Association for Computing Machinery. https://doi.org/10.1145/3605098.3636030

Rahman, M., & Watanobe, Y. (2023). ChatGPT for Education and Research: Opportunities, Threats, and Strategies. *Applied Sciences, 13*, 5783.

Rane, N. (2023). Enhancing mathematical capabilities through ChatGPT and similar generative artificial intelligence: roles and challenges in solving mathematical problems. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.4603237

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. Journal of Applied Learning & Teaching, 6(1), 364-389. https://doi.org/10.37074/jalt.2023.6.1.23

Schleppegrell, M. J. (2007). The Linguistic Challenges of Mathematics Teaching and Learning: A Research Review. *Reading and Writing Quarterly, 23*, 133-139.

Sepeng, P., & Madzorera, A. (2014). Sources of difficulty in comprehending and solving mathematical word problems. *International Journal of Educational Sciences, 6*(2), 217-225.

Soygazi, F & Oguz, D. (2024). An Analysis of Large Language Models and LangChain in Mathematics Education. In *Proceedings of the 2023 7th International Conference on Advances in Artificial Intelligence (ICAAI '23)*. Association for Computing Machinery, New York, NY, USA, 92–97. https://doi.org/10.1145/3633598.3633614

Stephany, S. (2021). The influence of reading comprehension on solving mathematical word problems: A situation model approach. In A. Fritz, E. Gürsoy & M. Herzog (Ed.), *Diversity Dimensions in Mathematics and Language Learning: Perspectives on Culture, Education and Multilingualism* (pp. 370-395). Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110661941-019

Tashtoush, M., Wardat, Y., AlAli, R., & Saleh, S. (2024). Artificial Intelligence in Education: Mathematics Teachers' Perspectives, Practices and Challenges. *Iraqi Journal for Computer Science and Mathematics*, 5(1), 60-77.

Ünal, Z., Greene, N., Lin, X., & Geary, D. (2023). What Is the Source of the Correlation Between Reading and Mathematics Achievement? Two Meta-analytic Studies. *Educational Psychology Review 35*(4).

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems*, Lisse, The Netherlands, Swets & Zeithinger.

Vintere, A., Safiulina, E., & Panova, O. (2024). AI-based mathematics learning platforms in undergraduate engineering studies: analyses of user experiences. In *Proceedings of 23th International Scientific Conference Engineering for Rural Development*.

Wardat, Y., Tashtoush, M., AlAli, R., & Jarrah, A. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education, 19*(7).

Yang, J. (2023). The impacts and responses of ChatGPT on education. In *Proceedings of the 7th International Conference on Education and Multimedia Technology (ICEMT '23)*. Association for Computing Machinery, New York, NY, USA, 69–73. https://doi.org/10.1145/3625704.3625732

Zhai, X. (2022). ChatGPT User Experience: Implications for Education. *SSRN Electronic Journal*. http://dx.doi.org/10.2139/ssrn.4312418.