Computers and Education: Artificial Intelligence 6 (2024) 100194

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Application of ChatGPT for automated problem reframing across academic domains

Hafsteinn Einarsson *, Sigrún Helga Lund **, Anna Helga Jónsdóttir **

University of Iceland, Sæmundargötu 2, Reykjavík, 102, Iceland

ARTICLE INFO

Keywords:
Large language models
Artificial intelligence
Interdisciplinary education
Problem reframing
Personalized learning

ABSTRACT

This paper explores the potential of large language models, specifically ChatGPT, to reframe problems from probability theory and statistics, making them accessible to students across diverse academic fields including biology, economics, law, and engineering. The aim of this study is to enhance interdisciplinary learning by rendering complex concepts more accessible, relevant, and engaging. We conducted a pilot study using ChatGPT to adapt problems across 17 disciplines, evaluated through expert review. Our results demonstrate the significant potential of ChatGPT in reshaping problems for diverse settings, preserving theoretical meaning in 77.1% of cases, and requiring no or only minor revisions in 74% of cases. An evaluation performed by 23 domain experts revealed that in 73.6% of cases the reframed problem was considered to add educational value compared to a corresponding abstract problem and to represent a real-world scenario in 57.0% of cases. Furthermore, a survey involving 44 Computer Science students revealed a diverse range of preferences between original and reframed problems, underscoring the importance of considering student preferences and learning styles in the design of educational content. The study offers insights into the practicality and efficacy of employing large language models, like ChatGPT, to enhance interdisciplinary education and foster greater student engagement and understanding.

1. Introduction

Consider a scenario where complex concepts from one discipline are seamlessly adapted to the unique context of another, thereby facilitating understanding for students from diverse academic backgrounds. Advancements in large language models (LLMs), like ChatGPT, are making this prospective scenario increasingly feasible. This paper explores the potential of these models to reframe problems from probability theory and statistics in a way that resonates with students across various disciplines, including biology, economics, law, and engineering.

The exploration of problem reframing is motivated by the challenges that students face when navigating the complex interdisciplinary landscape of undergraduate science programs. These programs often require introductory courses in diverse STEM fields, making it a daunting experience for students. For a detailed discussion of the benefits of personalized learning, please refer to the literature review section.

Despite the benefits of tailored content, adapting problems to suit the diverse contexts of various disciplines within large, heterogeneous classrooms presents a significant challenge for instructors due to the substantial time and effort required. In response, our research explores the potential of current LLMs in automating the process of tailoring problems for different disciplines. By leveraging LLMs, we seek to reconcile the needs of interdisciplinary education with the constraints educators face, contributing to a more personalized, engaging, and effective learning environment for students across a wide range of academic fields. Indeed, studies in various educational settings have provided evidence that personalization can lead to improved learning outcomes and student satisfaction (Chen et al., 2005, Chen, 2011, Hsu et al., 2013, Kim et al., 2018, Jonsdottir et al., 2017, Tseng et al., 2008, Alhazmi et al., 2018, Sancenon et al., 2022).

This exploration is guided by several relevant research questions that explore the potential applications and limitations of ChatGPT in the context of interdisciplinary problem reframing:

* Principal corresponding author.
** Corresponding authors.
E-mail address: hafsteinne@hi.is (H. Einarsson).
URL: https://english.hi.is/staff/hafsteinne (H. Einarsson).

https://doi.org/10.1016/j.caeai.2023.100194
Received 8 May 2023; Received in revised form 23 November 2023; Accepted 9 December 2023
Available online 13 December 2023

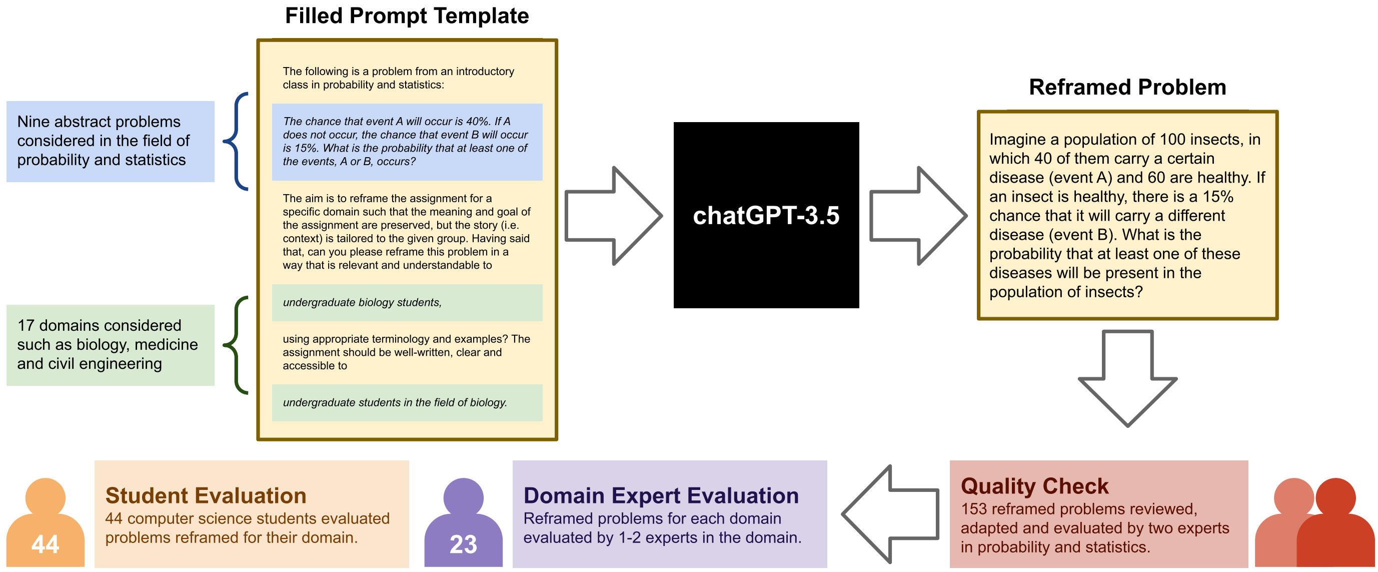

Fig. 1. An overview of the study design.

1. How effectively can ChatGPT reframe statistical problems for different academic disciplines while maintaining their educational value?
2. Are there any theoretical flaws or inaccuracies in the reframed problems generated by ChatGPT?

To address these questions, we present a pilot study that involves reframing problems for different settings using ChatGPT, covering nine problems in statistics and probability theory applied to 17 diverse domains of study. To perform the reframing, we present a simple prompt template that can benefit educators. Furthermore, we present the results of an evaluation performed by domain experts and results of a survey on students where their preference for reframed problems was evaluated. An overview of the study design can be seen in Fig. 1. We also discuss the potential benefits and challenges of using AI for this task. Our goal is to identify opportunities for future research and development in this area, contributing valuable insights into the practicality and efficacy of employing LLMs, like ChatGPT, to enhance interdisciplinary education and foster greater student engagement and understanding. To better contextualize our study, we will first delve into the existing literature on personalized learning, and the role of large language models in education. This literature review will provide a comprehensive understanding of the current state of these fields and highlight the gap our research aims to fill.

2. Literature review

The field of contemporary educational research is characterized by a multitude of themes, one of which is personalized learning. This literature review explores personalized learning, exploring the potential of tailoring educational content to potentially enhance student comprehension and engagement. It further explores the transformative role of large language models in education, and the emerging concept of automated problem reframing.

2.1. Personalized learning in education

Personalized learning, tailoring educational content to meet the unique needs and interests of each student, is identified as a key factor in improving student learning outcomes (Pane et al., 2015, 2017). This approach acknowledges that students have diverse learning styles, backgrounds, and interests, and that a one-size-fits-all approach to education may not be effective for all students (Tomlinson et al., 2003). A growing body of research supports the effectiveness of personalized learning regarding improved learning outcomes and student satisfaction (Chen et al., 2005, Chen, 2011, Hsu et al., 2013, Kim et al., 2018, Jonsdottir et al., 2017, Tseng et al., 2008, Alhazmi et al., 2018, Sancenon et al., 2022, Council et al., 2000, Hake, 1998, Trigwell et al., 1999).

In the context of the present study, Walkington (2013) found that students who received personalized instruction tailored to their out-of-school interests were faster and more accurate when solving mathematical problems than those who received traditional instruction. Similarly, Anand and Ross (1987) reported that personalizing problems in the context of friends, interests, and hobbies improved performance on standard and transfer problems. Cordova and Lepper (1996) showed that personalization significantly improved motivation, depth of engagement by students, amount learned in a fixed time period, and perceived competence and aspiration. However, it is also worth noting that simply presenting problems as stories can improve student performance (Koedinger & Nathan, 2004).

Despite the potential benefits, the large-scale implementation of personalized learning presents significant challenges, especially in the context of large, diverse classrooms. One of the key challenges is the need to adapt problems to cater to the diverse contexts of various disciplines, which can be time-consuming and labor-intensive for instructors. This challenge highlights the potential value of LLMs to automate the process of tailoring problems for different disciplines. Automated problem reframing holds the potential to benefit educators at all levels, as well as for content creators on learning platforms that offer problem sets to students, such as Webassign, Khan Academy, Coursera, and the tutor-web (Jonsdottir & Stefansson, 2014, Jonsdottir et al., 2021).

2.2. Large language models in education

Recent advancements in artificial intelligence, particularly in the realm of LLMs, offer promising implications for enhancing educational practices (Dwivedi et al., 2023, Gimpel et al., 2023). LLMs, like ChatGPT, have been trained on diverse datasets and can generate human-like text, making them useful for various educational tasks. This capability of the model has both positive and negative implications for education, and educators are even considering how to protect traditional approaches to teaching by creating LLM resistant exams (kaare Larsen, 2023). The integration of LLMs in educational settings remains a relatively nascent and under-explored domain. While some studies have highlighted the potential of LLMs for personalized learning such as

Table 1
Unmodified problem statements used in this work. For reframing these problem statements, the text was input verbatim as shown in the table, i.e. no specific equation formatting was used.

Problem title	Problem statement
Conditional probability	The chance that event A will occur is 40%. If A does not occur, the chance that an event B will occur is 15%. What is the probability that at least one of the events, A or B, occurs?
Normal distribution	Assume that X follows a normal distribution with a mean of 65 and a standard deviation of 8. Find k such that P(X < k) = 0.95.
Poisson distribution	Assume that X follows a Poisson distribution with lambda = 2. What is P(X > = 1)?
Binomial distribution	Assume that X follows a Binomial distribution with n = 8 and p = 0.3. What is P(X = 6)?
Central limit theorem	Assume you have 200 i.i.d. random variables, each with mean 1600 and standard deviation 200. What is the probability that their average will be greater than 1700?
Median	A variable was measured eleven times. The outcomes were: 5, 7, 24, 16, 3, 18, 9, 8, 18, 25, 5. What is the median of the measurements?
Comparing two means	In order to compare the means of two populations, 30 measurements are taken from each of the populations. The mean and standard deviation of the first sample are 23.5 and 8.5, while the mean and standard deviation of the second sample are 28.2 and 7.2. Is there a significant difference between the means of the two populations?
ANOVA	ANOVA is used to compare means in four populations. Five measurements are taken from each of the four populations. The total sums of squares is 1290, the treatment sums of squares is 190, and the error sums of squares is 1100. Is there a significant difference between the means of the four populations?
Regression	Data was collected on two continuous numerical variables, one independent and the other one dependent. The mean of the independent variable was 18.2 and its standard deviation was 3.1, whereas the mean of the dependent variable was 172.1 and its standard deviation was 21.4. The sample correlation coefficient between the two variables was 0.83. Using ordinary least squares, what are the estimated values of the intercept and the slope?

guided reading (Ochieng, 2023), automating the creation of educational content such as code explanations and programming assignments (MacNeil et al., 2022), and for teaching students to evaluate the output of LLMs (Mollick & Mollick, 2022), no studies to our knowledge have explored the opportunities of applying LLMs for reframing problems in an interdisciplinary manner in an educational setting.

2.3. Automated problem reframing

Although automated problem reframing is relatively under-explored, it is related to the concepts of style transfer and automated paraphrasing in text. Style transfer involves transforming a piece of writing into a specific style while preserving unrelated content Troiano et al. (2021), whereas paraphrasing involves expressing the same information using different words Zhou and Bhat (2021). Both techniques have shown promising results in various applications. However, it is their combination in models like ChatGPT, which employs generative pre-training and reinforcement learning with human feedback (Ouyang et al., 2022), that paves the way for more accurate and contextually relevant problem reframing.

3. Methods

In order to explore the capabilities and limitations of AI in reframing problems for different settings, we developed a prompting approach to guide the language model in generating revised versions of the problems. For the sake of clarity, we define a *prompt* as the input to a language model and a *prompt template* is input with specifically marked locations that need to be filled by a user. We will refer to the outputs of chatGPT and manual adaptions of them as *reframed problems*. Our approach involved a prompt template to help the language model understand the key elements of the original problem and the desired context or audience for the reframed version of the problem. In this section, we will describe our prompting approach in detail and provide examples of how it was applied to various problems.

3.1. Problem formulation

For the purpose of our study, we devised a collection of nine problems rooted in the principles of probability theory and statistics. These problems were intentionally crafted to encapsulate fundamental concepts typically encountered in an introductory course on these subjects. However, it's important to note that these problems were not derived from real-world scenarios, but rather were abstract in nature. This approach was adopted to keep the spotlight firmly on the task of problem reframing. For instance, one problem pertaining to binomial distributions required students to calculate the probability of the event that $X = 6$, given that X follows a binomial distribution with parameters $n = 8$ and $p = 0.3$. A comprehensive list of the original problem statements can be found in Table 1.

3.2. Prompting strategy

Considering that LLMs are trained to respond to textual inputs, the effectiveness of their output is often contingent on the specificity and relevance of the provided prompt. Consequently, the strategy for prompting is a crucial aspect of our research, as it steers the language model towards generating contextually appropriate adaptations of the problems. The objective of the prompts used in this study was to encapsulate the essential components of the original problem while delineating the intended context or target audience for the reframed version.

For the sake of explanation, we divide the prompt into three parts. The first part focuses on providing the language model with an understanding of the problem's subject matter. The second part states the original problem statement (with any relevant background information or definitions incorporated into the problem statement). The third part focuses on specifying the desired audience or context for the revised version of the problem. That includes providing information on the level of expertise or knowledge the audience has in the subject matter and what specific aspects of the problem are most relevant to them. The structure is as follows, with each part demarcated by two line breaks and placeholders encapsulated by angle brackets for problem statements, domain statements, and domains:

> *The following is a problem from an introductory class in probability and statistics:*
> ⟨**problem statement**⟩
> *The aim is to reframe the assignment for a specific domain such that the meaning and goal of the assignment are preserved, but the story (i.e. context) is tailored to the given group. Having said that, can you please reformat this problem in a way that illustrates its relevance and importance to* ⟨**domain statement**⟩*? The assignment should be well-written, clear and accessible to undergraduate students in the field of* ⟨**domain**⟩.

Table 2
Domain specific parts of the instructions for prompting the language model.

Domain	Domain statement
Architecture	architecture, using examples from the design, construction and history of buildings
Biology	undergraduate biology students, using appropriate terminology and examples
Business and marketing	business decision-making, using examples from the business world
Chemistry	chemistry, using examples from the chemical industry or laboratory research
Civil engineering	civil engineering, using examples from the construction and infrastructure industry
Computer science	computer science, using examples from the software development, artificial intelligence or computer systems
Economics	economics, using examples from financial and economic decision-making
Electrical engineering	electrical engineering, using examples from the electrical systems and devices industry
Food science	food science, using examples from the food industry, food production, or food preservation
History	history, using examples from historical events or historical figures
Industrial engineering	industrial engineering, using examples from the manufacturing and production industry
Law and pre-law	law, using examples from legal principles, case law or legal system
Mechanical engineering	mechanical engineering, using examples from the mechanics and thermodynamics industry
Medicine and pre-med	medicine, using examples from the medical field or patient care
Pharmacology	pharmacology, using examples from drug development and therapeutic treatments
Psychology	psychology, using examples from human behavior and mental processes
Social science	social science, using examples from the social and cultural context

The placeholders within the prompt called *domain statement* and *domain* refer to the information found in Table 2. We emphasize that it is crucial to provide specific and targeted prompts to guide the language model's output. Additionally, providing examples that illustrate the desired approach and give the language model a sense of the appropriate style and tone for the revised problem can further improve the output. Finding a good prompt can involve a lot of trial and error, and we encourage the reader to try out variations of the prompt template proposed in this work. It is worth noting that we did not use the language model as a conversational agent, meaning that we did not provide an answer to its response. The prompt was always used as the start of a new conversation. If the prompt is used in a conversational approach, i.e., by asking the model to reframe first problem 1, then problem 2 et cetera, then the stories will all be very similar. Starting a new conversation for each prompt led to greater diversity in the problem output, avoiding repetition of the same story with minor variations.

In Table 3, we present an illustrative example of the application of our prompting approach, along with a human-corrected adaptation that rectifies the inaccuracies in ChatGPT's output. It is noteworthy that the original problem's meaning is nearly preserved in ChatGPT's response. For instance, the response suggests that 40 insects are afflicted with disease A, implying a probability of one for the disease's presence in the population. This discrepancy is addressed in the human-corrected adaptation, underscoring the necessity for educators to discern these semantic deviations between the original and reframed problems. Such semantic inconsistencies are anticipated, given that there is no assurance of the correctness of the AI's response. Consequently, our study aims to evaluate the effectiveness of this method in reframing problems.

3.3. Evaluation of quality

To evaluate the quality of the revised problems generated by the language model, we used a human review process in which experts reviewed the output and rated it based on three criteria. Two experienced female professors in the department of mathematics undertook this review. The were aged between 35 and 44, specializing in probability and statistics, and each with a decade's worth of teaching experience in this domain. For each of the $17 \cdot 9 = 153$ reframed problems, the reviewers were asked to consider whether the meaning of the original problem was preserved, whether the revised problem made sense, and how much further modifications were required for the problem to be used in its intended setting. Table 4 details the scales used for these three criteria. Furthermore, the reviewers were given the instructions to not attempt to revise a problem if the theoretical meaning of the examples was not preserved. The reviewers were asked to use the scales to rate each revised problem, and the ratings were used to assess the overall performance of the language model in reframing problems for diverse contexts.

Crafting new problems can be both time-intensive and laborious. In recognition of this, the authors collaboratively devised this evaluation protocol to succinctly gauge the requisite effort needed to adapt the reformulated problems. The aim was to capture whether the meaning of the problem is preserved and how much manual work would be required to adapt the reframed problems as these factors would determine the usability of this approach. When a reframed problem was considered useless, it was not considered for downstream parts of this study and no further attempts were made to reframe the original problem with chatGPT-3.5. However, these problems were reframed with GPT-4 (see discussion).

Upon completing their individual evaluations, the reviewers consolidated their findings to generate a unified assessment for each problem. This process was underpinned by a consensus-driven dialog, ensuring that any disparities in their initial judgments were collaboratively reconciled to arrive at a definitive rating.

3.4. Domain expert evaluation

To evaluate the relevance and educational value of the reframed problems, we obtained expert feedback from 23 representatives covering all the 17 domains, including professors, an architect, and a third-year Ph.D. student in food science. Fig. 2 shows the distribution of the experts' age, gender, and teaching experience.

For the expert data collection, we designed 17 distinct surveys via Google Forms. Each expert only evaluated problems specific to their domain. They encountered both the original abstract problems and their tailored, reframed versions. For each of these pairs, the experts were prompted to express their level of agreement with two statements:

- The reframed problem represents a real-world scenario in your domain.
- The reframed problem adds educational value compared to the abstract problem.

The experts could respond on a Likert scale for each statement with the options: 1 - Strongly Disagree, 2 - Disagree, 3 - Neither Agree nor Disagree, 4 - Agree, and 5 - Strongly Agree. Answering each item was optional and missing data was left out from analyses. We had missing responses in 5 out of a total of 188 items. All domain experts were included in the analyses and missing responses were not treated in any specific manner.

Table 3
An example of applying our prompting strategy to a problem in probability theory framed for biology students with a human-written adaptation for applying the problem in the classroom.

Query	The following is a problem from an introductory class in probability and statistics: The chance that event A will occur is 40%. If A does not occur, the chance that event B will occur is 15%. What is the probability that at least one of the events, A or B, occurs? The aim is to reframe the assignment for a specific domain such that the meaning and goal of the assignment are preserved, but the story (i.e. context) is tailored to the given group. Having said that, can you please reframe this problem in a way that is relevant and understandable to undergraduate biology students, using appropriate terminology and examples? The assignment should be well-written, clear and accessible to undergraduate students in the field of biology.
Response	Imagine a population of 100 insects, in which 40 of them carry a certain disease (event A) and 60 are healthy. If an insect is healthy, there is a 15% chance that it will carry a different disease (event B). What is the probability that at least one of these diseases will be present in the population of insects?
Adaptation	Imagine a population of 100 insects, in which 40 of them carry a certain disease (event A) and the rest do not. If an insect does not carry the disease, there is a 15% chance that it will carry a different disease (event B). What is the probability that an insect chosen uniformly at random carries at least one of the diseases?

Table 4
Criteria used by a panel of experts to grade the reframed problems.

Criteria	Variable type	Scale
Is the theoretical meaning preserved?	Binary	Yes (The theoretical meaning of the original problem is maintained in the reframed problem); No (The theoretical meaning of the original problem is lost or significantly altered in the reframed problem)
How much are changes needed	Categorical	1 Useless (The reframed problem requires complete rewriting); 2 Major revision needed (The reframed problem requires significant changes to be usable; at least one minute of editing); 3 Minor revision needed (The reframed problem requires minor changes to be usable; less then one minute of editing); 4 Little to no revision needed (The reframed problem is nearly or completely ready for use as is)
Are the values appropriate?	Binary	Yes (The numerical values in the reframed problem are appropriate and make sense in the given context); No (The numerical values in the reframed problem are inappropriate or do not make sense in the given context)

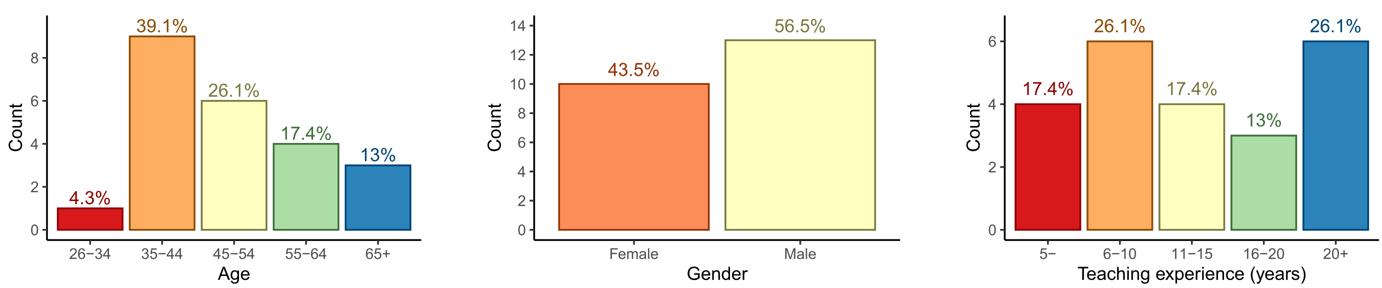

Fig. 2. Distribution of age, gender and teaching experience for the domain experts that reviewed the reframed problems.

3.5. Student evaluation

For a preliminary test of student engagement with the reframed problems, we sent a survey using Google Forms to 434 students in the department of computer science who had all completed at least one year of their three year program. 44 students responded to all questions in the survey (2 were excluded because they did not respond to all items) and 35 of those students had finished an introductory course in probability and statistics. All responses were used in the analysis and students were included in the analysis even though they did not respond to all items.

These students were presented with seven original abstract problems alongside their computer science-specific reframed versions. The ordering of the problems was not fixed, i.e., in 4 questions the abstract problem was presented first. For each pair, the problems were labeled as problem A and problem B. For each such paired set, students were prompted to indicate which version of the problem they preferred (A or B). More precisely, we used the following formulation: "As a student, which of the following problem formulations do you consider to be better, A or B?" At the survey's conclusion, participants were also given the opportunity to offer any additional insights or thoughts on the problem reframing process showcased within the survey.

3.6. Statistical methods

Association between categorical variables was estimated with Pearson's Chi-squared test with Monte Carlo simulated p-values, and Cramer's V was used to assess intra-reviewer agreement (Mangiafico, 2023). Association between a grade on Likert scale and categorical variables was estimated with ANOVA. All statistical analysis was performed in R version 4.2.2 (R Core Team, 2022), and figures were created with the ggplot2 package (Wickham, 2016).

4. Results

There was strong inter-rater agreement between the two reviewers, as indicated by the substantial concordance in Cramer's V values for the

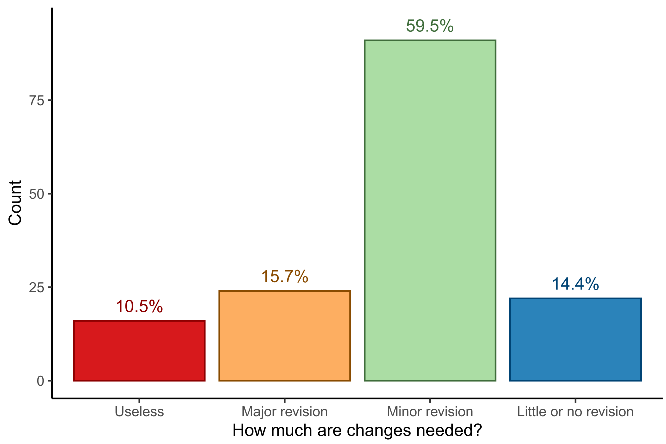

Fig. 3. Required changes for the generated problems.

preservation of theoretical meaning (0.80, 95% CI: 0.68-0.91), changes needed (0.58, 95% CI: 0.49-0.68), and appropriate values presented (0.76, 95% CI: 0.63-0.89). Consequently, the two reviewers convened to finalize an anonymous review for each problem, which the following sections are based on. The prompts, ChatGPT outputs, edited stories, and final reviewer evaluations are provided in the supplementary materials.

4.1. ChatGPT successfully reframes problems for diverse settings

ChatGPT generated valuable input for 137 of the 153 problems (89.5%). The vast majority (74%) of the generated problems required no or minor revisions, with only 16 problems (10.5%) deemed unsuitable by the reviewers (see Fig. 3).

The Social Sciences domain exhibited the highest proportion of reframed problems deemed ineffective (3 out of 9), although this ratio did not significantly differ from other disciplines ($p = 0.413$). Similarly, the type of problem did not significantly influence the proportion of ineffective reframed problems ($p = 0.746$). Fig. 4 presents the evaluation ratings for each problem statement, categorized by domain and problem type. It's worth noting that a problem was typically classified as ineffective by the review committee if the narrative context did not align well with the problem. Nevertheless, even these less successful reframing attempts by ChatGPT could serve as a source of inspiration for manual problem reframing, albeit at the cost of additional time required to revise the problem.

Below is a glimpse of the diversity present in the generated stories, which is evident from the examples of different topics covered. The stories touch upon a wide range of subjects, including the reign of King Henry VIII of England (history), fruit flies (biology), and the development of new software (computer science). The examples presented are unedited.

Conditional probability - History
Imagine that you are a historian studying the reign of King Henry VIII of England. During his reign, there were two potential outcomes for the fate of his political opponents: execution or banishment. Event A represents the chance that an opponent will be executed, which has been determined to be 40%. Event B represents the chance that an opponent will be banished, which has been determined to be 15% if event A does not occur. ...

Normal distribution - Civil engineering
Imagine you are a civil engineer working on a construction project. The project involves building a bridge that will span a large river. The height of the bridge is a critical factor, as it needs to be high enough to allow large ships and boats to pass underneath it. You have been tasked with determining the height of the bridge that will be required to ensure that at least 95% of the ships and boats that pass underneath it will be able to do so safely. ...

Poisson distribution - Social sciences
Imagine you are conducting a study on the number of crimes that occur in a certain neighborhood. You have collected data for the past year and have determined that the number of crimes follows a Poisson distribution with a rate of 2 crimes per day. You are now interested in determining the probability that there will be at least one crime in a given day. ...

Binomial distribution - Biology
Imagine that you are studying a population of fruit flies in a laboratory. You want to determine the probability of a certain genetic trait being present in a sample of 8 fruit flies. You know that the probability of this trait occurring in a single fruit fly is 0.3. You want to calculate the probability of exactly 6 out of the 8 fruit flies in your sample having the genetic trait. Using the binomial distribution, what is the probability of this outcome?

Central limit theorem - Computer science
Imagine you are working on a team responsible for developing a new software system. Your team has 200 developers working on the project, and each developer has been assigned a specific set of tasks to complete. The average time it takes for a developer to complete their tasks is 1600 hours, with a standard deviation of 200 hours. The project manager has set a deadline for the software to be completed, and the team must meet this deadline in order to launch the software on schedule. The question is, what is the probability that the average time it takes for all 200 developers to complete their tasks will be greater than 1700 hours? ...

Median - Economics
In economics, understanding the distribution of a variable is crucial for making informed financial and economic decisions. One measure of the center of a distribution is the median. Consider a company that has recorded the revenue from its last 11 fiscal quarters. The revenue figures were: $5 million, $7 million, $24 million, $16 million, $3 million, $18 million, $9 million, $8 million, $18 million, $25 million, $5 million. What is the median revenue for the company? ...

Comparing two means - Industrial engineering
In industrial engineering, it is important to understand how different production processes or manufacturing techniques affect the quality of the final product. In order to compare the effectiveness of two different production methods, 30 samples of the final product are taken from each method. The mean and standard deviation of the first sample are 23.5 and 8.5, while the mean and standard deviation of the second sample are 28.2 and 7.2. Is there a significant difference in the quality of the final product between the two production methods? ...

ANOVA - Medicine
In the medical field, it is important to understand if there are significant differences in the effectiveness of different treatments for a specific condition. For example, let's say we are studying the effectiveness of four different types of medication for treating high blood pressure. We randomly select five patients from each group and measure their blood pressure before and after taking the medication for a certain period of time. ...

Linear regression - Architecture
Imagine you are an architect designing a new building. One of the key factors you need to consider is the energy efficiency of the building. You decide to gather data on two important variables: the size of the building (in square feet) and the annual energy consumption (in kilowatt-hours). ...

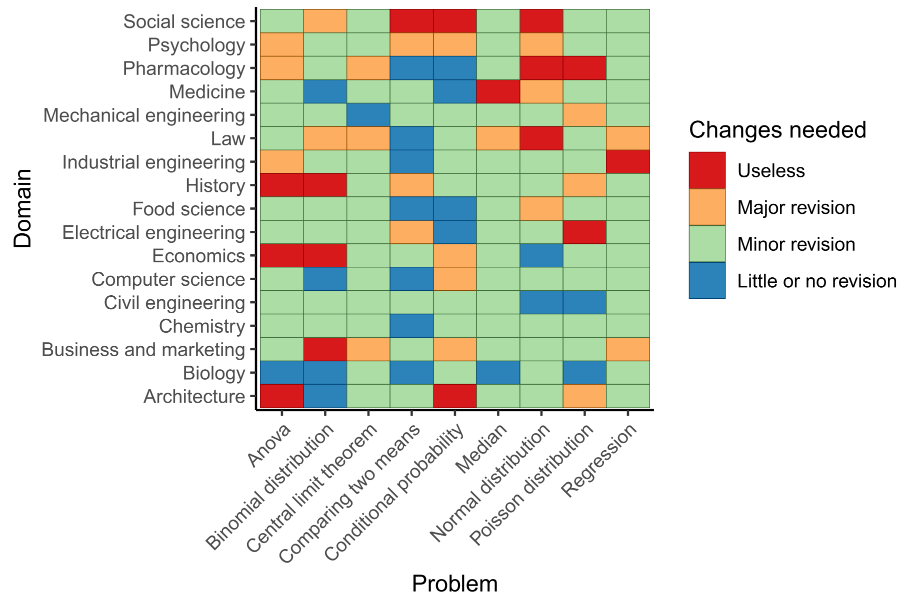

Fig. 4. The consensus on required modifications, as determined by the review panel, for the reframed problems, categorized by domain and problem type.

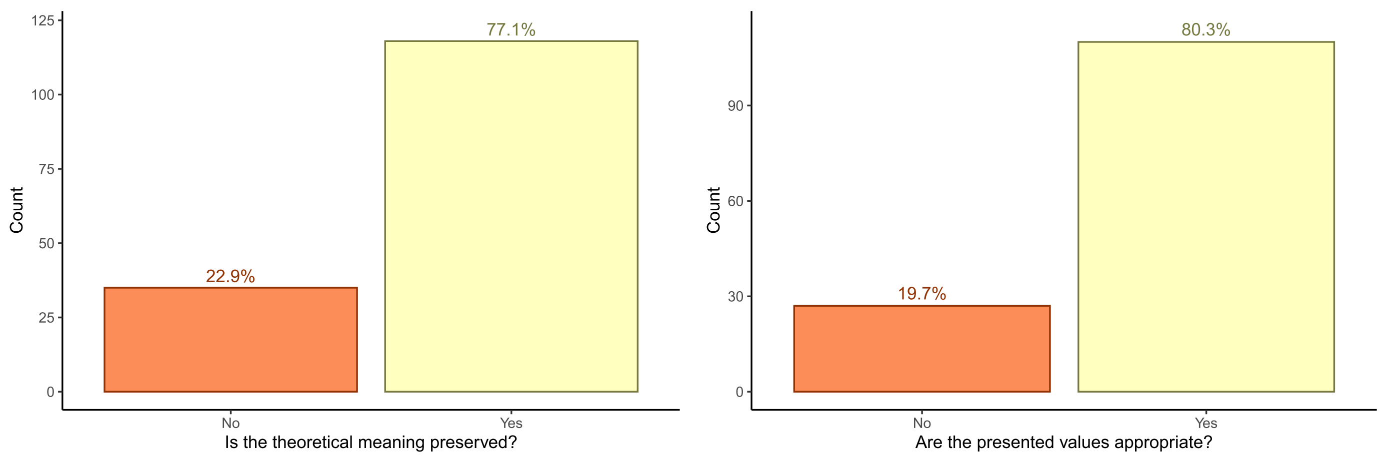

Fig. 5. Left: Preservation of theoretical meaning in the generated problems. Right:Frequency of problems generated with appropriate values.

4.2. Theoretical flaws in reframed problems are infrequent

The theoretical meaning was preserved in 118 out of 153 problems (77.1%), and 110 out of the 137 problems (80.3%) that were not classified as useless had appropriate numerical values (see Fig. 5). No significant statistical difference was found in whether the meaning was preserved nor if it had appropriate values by domain applied or problem type. For a complete overview by domain and problem type, see Fig. 6.

Within a specific problem, similar flaws were frequently observed in the generated problems. For instance, in the Conditional Probability problem set, the most common flaw was failing to differentiate between the probability of event A occurring, $P(A)$ and the conditional probability of event A occurring given that event B has occurred $P(A|B)$. An example of this can be seen below, where $P(B)$ is given instead of $P(B|\neg A)$ as in the original assignment:

... Imagine you are an investor considering two different investment options, A and B. Option A has a 40% chance of providing a positive return on investment, while option B has a 15% chance of providing a positive return. What is the probability that at least one of these options will provide a positive return on investment?. ...

The most frequent error encountered in the Normal distribution problems was that the probability requested in the generated problems was the complement of the probability in the original problem. Although this flaw might not pose a significant concern for educators as the generated problems can still be used, it is important to note that these problems differ from the original problem. An example of this can be seen below where the assignment asks for the value of k such that $P(X > k) = 0.95$ instead of $P(X < k) = 0.95$:

...Imagine that a chemist is measuring the purity of a chemical compound, and they want to know what proportion of the samples they collect will have a purity level that is higher than a certain threshold. The chemist has previously collected data on the purity levels of the compound and found that it follows a normal distribution with a mean of 65% and a standard deviation of 8%. The chemist wants to know what purity threshold, k, will ensure that 95% of the samples collected will have a purity level higher than that threshold. Using the information about the mean and standard deviation of the compound's purity levels, the chemist can use probability and statistics to find the value of k that meets this requirement. ...

The most common flaw identified in the inferential statistics problems was the imprecise usage of language when distinguishing between

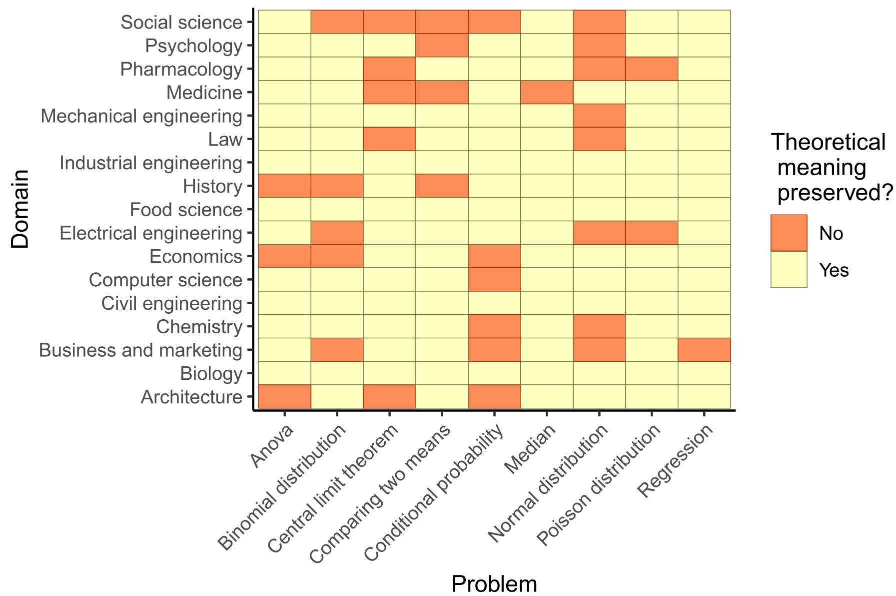

Fig. 6. Preservation of theoretical meaning by domain and problem type.

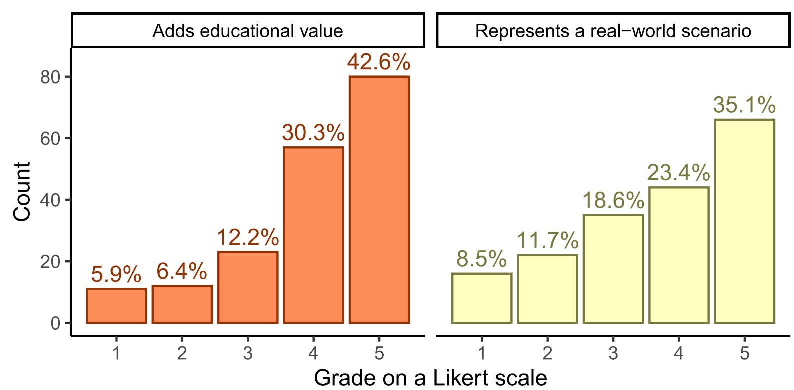

Fig. 7. Distribution of evaluations by domain experts on the reframed problems. A score of 1 on the Likert scale represents strong disagreement and a score of 5 represents strong agreement.

populations and samples. This issue is exemplified below, where the word group is used for both the sample and the population. Furthermore, the given problem also features inappropriate values where the mean height of soldiers is unusually low:

> *Imagine that you are a historian studying two different groups of people during different time periods in history. The first group is a group of 30 soldiers from the Revolutionary War, and the second group is a group of 30 soldiers from World War II. You want to compare the average height of the soldiers in each group to see if there is a significant difference between the two groups. The mean height for the Revolutionary War soldiers is found to be 23.5 inches, with a standard deviation of 8.5 inches. The mean height for the World War II soldiers is found to be 28.2 inches, with a standard deviation of 7.2 inches. Is there a significant difference in the average height of the soldiers between the two groups?* ...

Finally, we often observed attempts from ChatGPT to solve the problem, and we marked this as a minor revision required. Such solutions were often incorrect and to prevent this type of output, we recommend to ask the model not to provide a solution along with the reframed problem.

4.3. Evaluation by domain experts

Each expert reviewed abstract problems and corresponding reframed problems from their domain. They evaluated whether the reframed problems added educational value over the abstract versions and if they represented real-world scenarios (see Methods). Fig. 7 shows the overall distribution of responses. Experts considered the reframed problems to add educational value over abstract versions in 72.9% of cases (grades 4-5) and to represent real-world scenarios in 58.5% of cases.

Responses by age, gender and teaching experience are shown in Fig. 8. For the statement that reframed problems added educational value compared to the abstract problems, the average score varied by the level of teaching experience ($p < 0.001$). We noticed that experts with greater teaching experience had the highest proportion of participants agreeing strongly with the statement. We also analyzed responses by problem type with the results shown in Fig. 9. From this data, we conclude that the problems on comparing two means and on the normal distribution were most successful.

For the statement that the reframed problem represented a real-world scenario, the average score varied by the level of teaching experience ($p < 0.001$). Experts with more teaching experience most frequently strongly agreed with the statement as seen in Fig. 8. Fig. 9

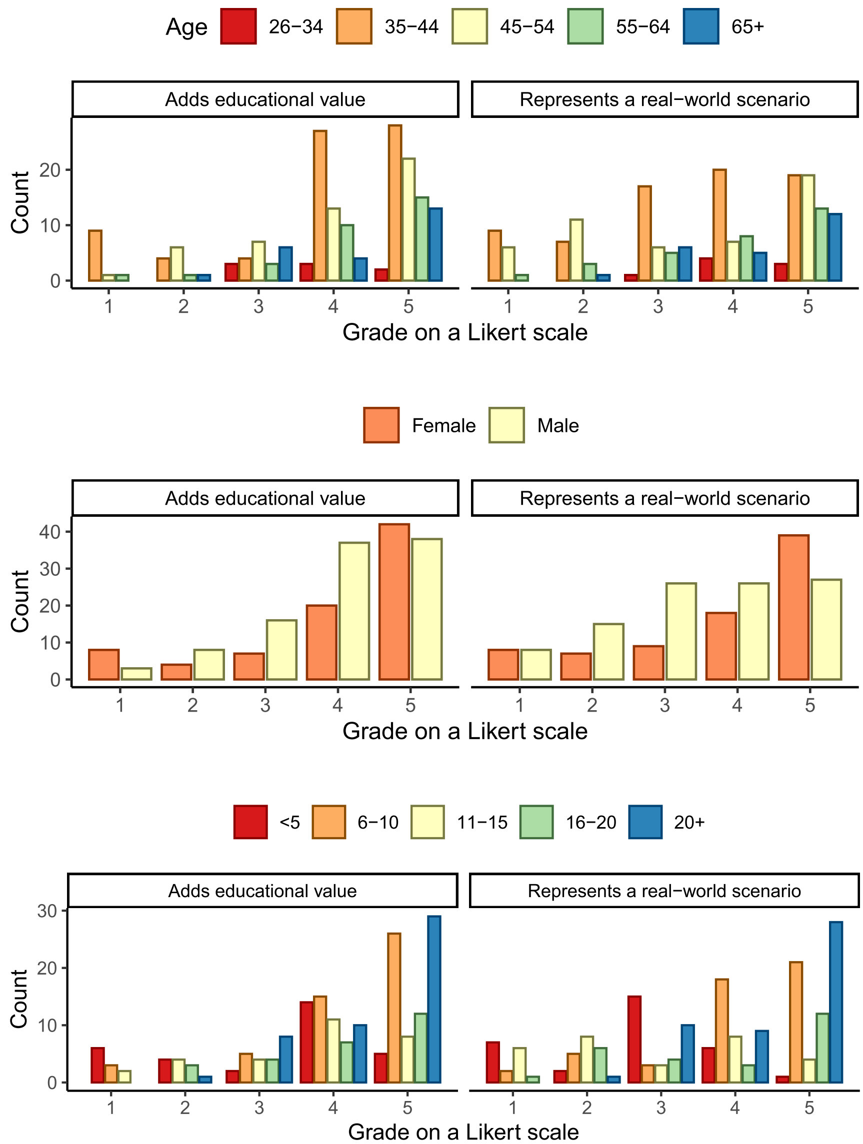

Fig. 8. Distribution of evaluations by domain experts by age, gender and teaching experience. A score of 1 on the Likert scale represents strong disagreement and a score of 5 represents strong agreement.

shows responses by problem type, indicating the reframed problems on comparing two means and normal distribution were most successful.

In conclusion, expert feedback suggests the reframed problems add educational value over abstract versions, and experts with more teaching experience held that opinion most strongly.

4.4. Assessing the student experience

To gauge the student experience, we conducted a survey involving 44 Computer Science (CS) students from the University of Iceland. The survey presented the students with seven tasks, each requiring them to express a preference between an original problem and its reframed counterpart tailored to the CS domain. The students were informed at the outset that the original problems had been reframed using ChatGPT and subsequently adapted and translated into Icelandic by a human.

The distribution of student preferences, as depicted in Fig. 10, does not reveal a clear inclination towards either original or reframed problems. The majority of students (13 out of 44) favored original problems exclusively, while the second largest group (8 out of 44) showed a preference for reframed problems only. When we analyzed the preferences according to problem type, we found no strong preference towards either original or reframed problems (see Fig. 11). The preference for reframed problems ranged from 27.3% (for conditional probability) to 59.1% (for binomial distribution).

In the final part of the survey, we invited students to share any additional thoughts they might have regarding the reframing of problems. Nine students took up this opportunity. The feedback we received was insightful and varied. Several students expressed concerns about the length of the reframed problems, indicating a preference for brevity and simplicity. This sentiment is reflected in the following translated responses:

- *I find that shorter questions are often better, but there could sometimes be a middle ground where more context is provided without it being an "essay".*
- *It's good to have "context" around concepts, but preferably if it's not too long, especially if it's something simple, then it might be better to have a simple explanation.*
- *There's no need to complicate matters with crazy descriptions in software development. You start to complicate the examples insanely much and they look incredibly hard, but when they are short they feel more simple.*

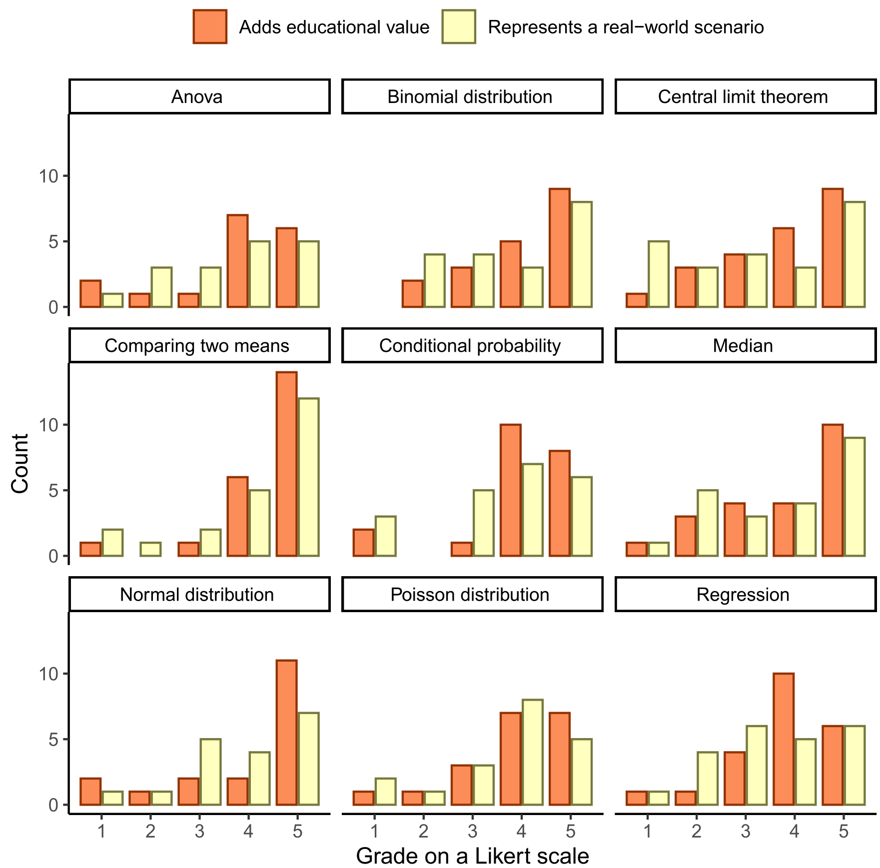

Fig. 9. Distribution of evaluations by domain experts on the reframed problems by problem type. A score of 1 on the Likert scale represents strong disagreement and a score of 5 represents strong agreement.

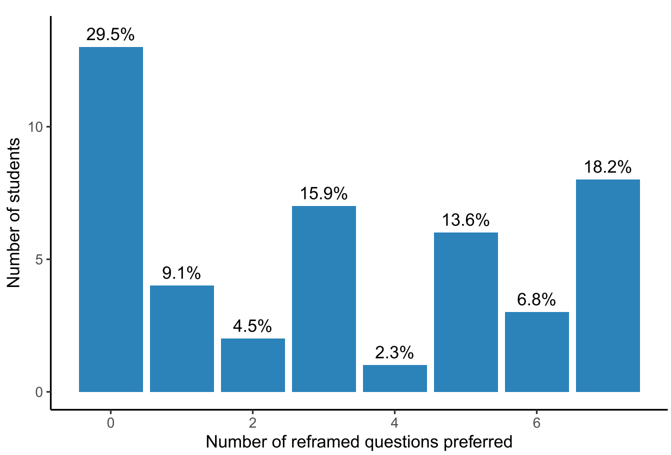

Fig. 10. Distribution of student preferences between reframed and original problem formulations.

- *We want to have as little "fluff" in the content as possible. Skip all the "assume" and "imagine" and just say directly "You are...". Then focus on what you want to teach, either to find patterns from information or to practice formulas.*

This preference for brevity is reflected in the preferences depicted in Fig. 11. The reframed problem on binomial distributions is only slightly longer than the original version. However, length alone does not fully account for the observed preferences. For instance, the reframed problem on Poisson distribution was four times longer than the original version, yet many students still preferred the reframed version. Other students expressed a more positive view of the reframing approach:

- *I feel it matters greatly to me as a student to be able to understand the purpose of what I am learning. Therefore, the examples where clear examples of the use of specific concepts or methods are used are much better for me than the examples that are just pure mathematics.*
- *Wonderful development in teaching methods is taking place here.*

One student expressed a preference for the original problem, citing their prior knowledge of the material as the primary reason. This student found comfort in the familiarity of the original problem, which aligned with their previous learning experiences.

On the other hand, two students expressed a different viewpoint. They found the original problems to be too "pure" in their mathematical notation, which they found off-putting. These students appreciated the reframed problems, as they provided a more relatable and accessible context. In other words, the reframed problems offered a more engaging and less intimidating approach to understanding complex concepts.

5. Discussion

Our exploration of the potential of LLMs, specifically ChatGPT, in reframing academic problems for diverse disciplines has yielded promising results. In this section, we discuss our findings in relation to our research questions, outline the implications of our study, and highlight its limitations and potential avenues for future research.

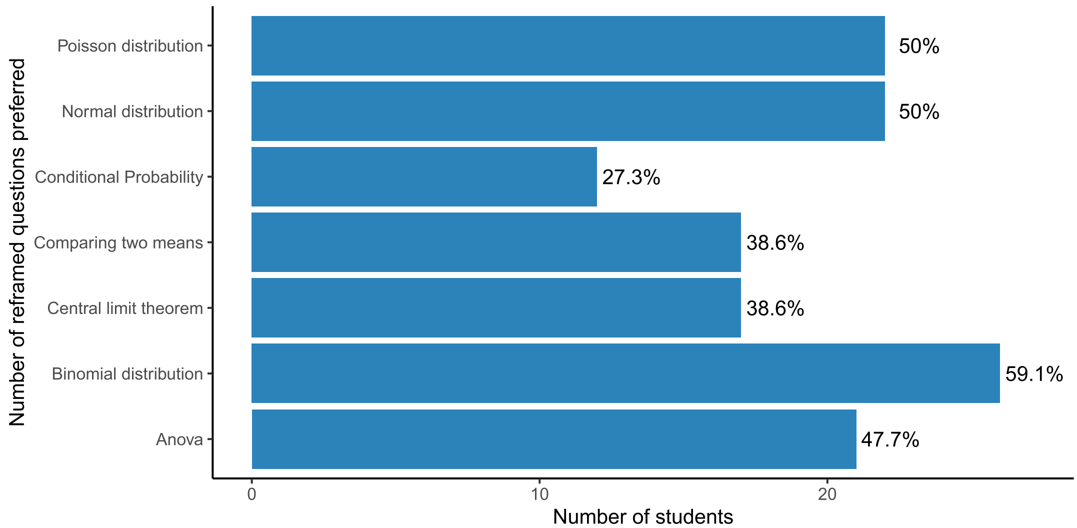

Fig. 11. Comparison of student preferences for reframed versus original problems, categorized by problem type.

5.1. Research question 1: effectiveness of ChatGPT in reframing problems

Our first research question sought to determine the extent to which ChatGPT can effectively reframe problems for different domains while preserving their educational value. Our findings indicate that ChatGPT is capable of reframing problems across a wide range of academic disciplines, with 77% of the reframed problems preserving their theoretical meaning and 74% requiring minimal revisions. This suggests that ChatGPT can serve as a valuable tool for educators seeking to create tailored educational content that resonates with students from diverse academic backgrounds.

However, it is important to note that while ChatGPT was effective in most cases, it still required expert review and revisions to ensure the accuracy and appropriateness of the reframed problems. This underscores the need for further refinement of AI models to reduce the need for human intervention. The advent of more advanced models, such as GPT-4, Anthropic's Claude, Google's Gemini and Pi, offers promising prospects for improving the performance of LLMs in this task.

Regarding the applicability of reframed problems to various domains, expert assessments disclosed that 57% of such problems authentically mirrored real-world scenarios within their respective areas. To enhance this percentage, one could provide the model with more explicit context during the reframing process, although this would necessitate increased effort on the part of educators. It is plausible that future, more sophisticated models will elevate this metric. Moreover, 73.6% of the time, experts believed that the reframed problems enhanced educational value in comparison to the original abstract problem. This endorsement was most pronounced among domain experts boasting over two decades of pedagogical experience, potentially signaling their recognition of the efficiency gains and opportunities AI-driven problem reframing offers.

In terms of the student experience, our survey involving 44 Computer Science students revealed a diverse range of preferences between original and reframed problems. This underscores the importance of considering student preferences and learning styles in the design of educational content. Some students expressed a preference for brevity and simplicity, while others appreciated the context and relatability provided by the reframed problems. This feedback provides valuable insights for educators and AI developers alike, highlighting the need for a balanced approach that combines clarity, brevity, and context in problem reframing.

While we garnered invaluable perspectives from domain specialists and students, we did not probe into the direct impact of reframed problems on actual learning outcomes, thus leaving a critical dimension of educational efficacy unaddressed in this pilot study. Previous research has shown that personalized assignments, particularly those contextualized with narratives, can bolster problem-solving capabilities Walkington (2013). Given the timeliness and overarching significance of this subject for the pedagogical community, we deem it essential to share our findings, fostering further exploration into how these foundational concepts might influence learning outcomes.

5.2. Research question 2: theoretical flaws in reframed problems

Our second research question asked whether there are any theoretical flaws or inaccuracies in the reframed problems generated by ChatGPT. Our findings reveal that theoretical flaws were somewhat infrequent (23%), with 96.1% of the reframed problems that were not classified as useless having appropriate numerical values. This reaffirms the potential of ChatGPT in providing accurate and contextually relevant problem formulations.

However, we did observe some recurring errors in the reframed problems, but there is no way to tell beforehand how the reframing process might fail, necessitating a careful review process. The need for expert review highlights the challenge of "distinguishing between real knowledge and convincingly written but unverified model output" that was discussed in the work of Kasneci et al. (2023). The need for a review process further aligns with observations that ChatGPT's mathematical abilities are not as stellar as those reported for other disciplines Frieder et al. (2023).

5.3. Implications and recommendations

The outcomes of our research have significant ramifications for the field of education, particularly in the context of interdisciplinary pedagogy. The demonstrated capability of ChatGPT to adapt academic problems across a wide range of disciplines heralds a new era of possibilities for educators. This offers an innovative method to deepen the comprehension of interdisciplinary learning, aligning with the broader educational objective of personalized learning, a concept that has been identified as a key factor in maximizing each student's learning potential (Zawacki-Richter et al., 2019, Roschelle et al., 2020, Zhang et al., 2019).

It is important to highlight that in our study, the primary interaction is between the educator and ChatGPT, with the only data required from students being their field of study. This underscores the potential of ChatGPT as a tool for educators, enabling them to create customized educational content without the need for extensive student data.

Furthermore, while all the work in this study could have been conducted without ChatGPT, it would have been a significantly more labor-intensive and subjective process. The use of ChatGPT streamlined the reframing of problems, making a task that would have been prohibitively time-consuming for most educators both feasible and efficient. This highlights the transformative potential of AI in education, particularly in tasks that require a high degree of customization and personalization.

However, the integration of ChatGPT into educational settings is not without its challenges. These include the generation of unverified or potentially biased output, necessitating meticulous review and moderation (Kasneci et al., 2023, Dwivedi et al., 2023). Moreover, educators bear the responsibility of validating the reframed problems. In preparing the material, educators should consider the four criteria proposed by Jacobs (1989): Are the problems valid within the disciplines, for the disciplines, beyond the disciplines, and do they contribute to broader outcomes? From our experience, ChatGPT generates reframed problems that are valid within the context of probability and statistics. However, assessing their relevance within a specific frame of reference can be challenging when they fall outside the educator's area of expertise. As we saw with domain expert evaluations, the stories were not always judged as representing a real-world scenario. Assessing the validity of the problems for the disciplines can also be challenging, as the reframed problems, although phrased in terms familiar to the student, may not necessarily contribute to the student's knowledge within that discipline. Evaluating the validity of the problems beyond the disciplines and their contribution to broader outcomes is challenging because it can require access to endpoints such as job performance that might be hard to quantify and compare. However, reframed problems hold the potential to help students forge connections between disciplinary issues and learn how to translate domain-specific problems into abstract problems, a skill that can transcend both disciplines and contribute to broader outcomes. Prior studies support this assertion, indicating that when reframed problems relate to out-of-school interests, students' ability to write symbolic equations from story scenarios significantly improved compared to students who received traditional instruction (Walkington, 2013).

To effectively harness the potential of ChatGPT, we propose several strategies for applying reframed problems in an educational context. These strategies aim to strike a balance between leveraging the capabilities of LLMs and ensuring the accuracy and appropriateness of the educational content.

5.3.1. Choice-based approach

Our survey revealed a wide range of student preferences, suggesting that one potential strategy could be to provide students with a choice between the original and reframed versions of problems. This approach would empower students to select the format that best resonates with their individual learning style, potentially fostering increased engagement and a deeper comprehension of the subject matter. This strategy not only acknowledges the diversity of student learning styles but also encourages students to take an active role in their learning journey.

Furthermore, this choice-based approach could serve as a valuable teaching tool, helping students to appreciate the complexity of longer problems and equipping them with the skills to effectively translate these problems into their abstract counterparts. However, it is important for educators to consider the potential benefits of not always offering a choice. This could challenge students to grapple with more complex problems, thereby enhancing their ability to translate intricate, real-world scenarios into abstract mathematical representations.

5.3.2. Narrative abstraction approach

Transforming abstract problems into relatable narratives can be a powerful pedagogical tool. However, it is essential to recognize that not all students might find this approach immediately beneficial, especially when they are at the beginning stages of their academic journey. Introducing stories or real-world scenarios can add an extra layer of complexity that might not always be conducive to learning for some individuals.

To tackle this, educators can use the following approach: before diving into the solutions, students could be trained to distill these narratives back into their abstract forms. This exercise can serve dual purposes. Firstly, it aids students in discerning the essential elements of a story and how they correlate with the abstract concepts they've encountered in class. Secondly, it provides students with a structured method to transition between real-world scenarios and mathematical representations, equipping them with a skill set that can be invaluable in numerous academic and real-world situations.

5.3.3. Supplementary material approach

Another approach could be to use the reframed problems as supplementary material to the original problems. This would provide students with an opportunity to see the practical applications of abstract concepts in various domains, thereby enhancing their understanding and appreciation of the subject matter. This approach could particularly benefit students who find traditional mathematical notation daunting, as the reframed problems provide a more accessible and engaging way to understand complex concepts.

5.3.4. Collaborative learning approach

Reframed problems could also be used in group discussions or collaborative learning activities. This method encourages students from various disciplines to collaborate in problem-solving, thereby fostering an environment of interdisciplinary learning. In this context, students not only learn from the problem at hand but also from each other's unique disciplinary perspectives, thereby deepening their understanding of the subject matter. This collaborative learning approach not only enhances students' problem-solving abilities but also cultivates essential collaboration skills, both of which are invaluable for their future professional endeavors.

5.3.5. Student-centered reframing approach

Finally, an innovative approach could involve introducing students to the prompting methodology used in this study. Students could be tasked with abstract problems and entrusted with the creation and refinement of a contextual narrative for each problem. They would then be expected to devise a solution that aligns with the narrative's style and submit both their reimagined problem and its corresponding solution. This pedagogical strategy not only empowers students to choose the context that best enhances their comprehension of the problem, but also fosters creativity and critical thinking skills. Furthermore, it provides an opportunity to educate students about the limitations of LLMs and their role in rectifying the model's output. While this approach may alleviate the burden of assignment creation for educators, it could introduce new complexities in the grading process. This strategy aligns with the efforts of other researchers who advocate for teaching students how to effectively interact with and utilize LLMs (Mollick & Mollick, 2022).

5.4. Limitations and future directions

5.4.1. Scope of the study

One of the primary limitations of our research pertains to the range of problem types and academic disciplines explored. The problems and domains we explored, while diverse, represent only a fraction of the vast expanse of interdisciplinary education. This limitation might have influenced our findings, as certain problem types or academic domains could present unique challenges or opportunities for problem reframing that were not captured in our study. Therefore, the results should be interpreted with caution, as they may not generalize to all problem types or academic domains.

Future research could expand the scope of the study by incorporating a broader array of problem types and disciplines. This would provide a more comprehensive understanding of ChatGPT's capabilities in reframing problems for interdisciplinary education. It could also lead to insights that could inform the development of improved prompting strategies or the fine-tuning of the model for better performance in certain domains.

5.4.2. Dependence on expert review

Another significant limitation is the dependence on expert review and revisions to ensure the accuracy and appropriateness of the reframed problems generated by ChatGPT. Despite the promising results, the need for expert intervention underscores the challenges of using AI in education and the importance of further refining AI models to reduce the need for human intervention.

More advanced models, such as GPT-4 (OpenAI, 2023), will likely perform better in this task and reduce the need for post-editing by the teacher. Preliminary tests with GPT-4 on the problems labeled as useless in our study have shown promising results, with all but one resulting in a useful reframing (see supplementary material).

5.4.3. Subjective assessment

Our study relied on subjective assessments by reviewers and students, which could be influenced by individual biases and expertise. This subjective evaluation could potentially limit the reliability and generalizability of our findings. Furthermore, the homogeneity of the student group could limit the generalizability of the student evaluation as students from other disciplines might have different preferences.

Future research could benefit from incorporating more objective evaluation metrics to assess the quality of reframed problems and a more diverse student population. For instance, future studies could measure student performance on the generated problems or use machine learning techniques to automatically evaluate the quality and relevance of reframed problems. In that regard, using GPT-4 for evaluation has been applied in several works already and shown to align well with human evaluations (Eloundou et al., 2023, Liu et al., 2023, Gao et al., 2023, Eldan & Li, 2023).

5.4.4. Language translation challenges

Another notable limitation of our study is the challenge of translating the reframed problems into languages other than English. While ChatGPT has demonstrated proficiency in several languages, its performance can vary significantly depending on the language in question. For instance, languages such as Icelandic, which have complex grammatical structures and a smaller corpus of training data, may pose considerable challenges for automated translation (Símonarson et al., 2021). This limitation is particularly relevant in the context of global education, where learning materials need to be accessible to students from diverse linguistic backgrounds.

The process of translating the reframed problems into other languages currently requires expert intervention, which can be time-consuming and resource-intensive. This underscores the need for further advancements in AI models to improve their multilingual capabilities and reduce the need for human intervention. Despite these challenges, the potential of LLMs in enhancing the accessibility and inclusivity of education across linguistic boundaries remains a promising avenue for future exploration.

5.4.5. Future directions

Our research opens up several avenues for future work and the question remains of how and if these domain-specific reformulated problems improve student learning outcomes. We are currently exploring if that is the case as it would provide empirical evidence on the effectiveness of tailored educational content in enhancing student understanding and engagement. Earlier research has illuminated the benefits of personalizing challenges through narrative contexts, showcasing enhanced problem-solving prowess among learners (Walkington, 2013). Yet, an unresolved question remains: Would students exhibit heightened engagement on an educational platform when presented with problems tailored to their specific domain as opposed to more abstract problems? It is conceivable that carefully reframed problems can increase the entertainment value of problems, and thereby improve engagement. The potential of LLMs in education is vast, and creativity in applying these models is the only limiting factor.

6. Conclusion

In this exploration, we delved into the capabilities of LLMs, specifically ChatGPT (version 3.5), to reframe academic problems across a spectrum of disciplines. Our findings illuminate the impressive potential of ChatGPT in this context, with 77% of reframed problems retaining their theoretical essence and 74% necessitating minimal revisions. This highlights the potential of ChatGPT as a potent tool for educators, enabling the creation of tailored educational content that can resonate with students from a wide array of academic backgrounds.

However, our study also underscores the necessity for expert review and refinement to ensure the accuracy and relevance of the reframed problems. Domain experts considered the problems to reflect a real-world scenario in 57% of the cases, and add educational value in 73.6% of cases. This underlines the challenges inherent in the application of AI in education and the importance of further honing AI models to minimize the need for human intervention.

Our research carries substantial implications, heralding new possibilities for interdisciplinary teaching and learning. By offering an innovative approach to deepen the understanding of interdisciplinary learning, our study aligns with the broader educational objective of personalized learning.

Moreover, our survey results reveal a pronounced preference among a substantial proportion of students for reframed problems, suggesting that this approach could potentially enhance student engagement and comprehension. As language models continue to evolve and improve, we anticipate that the quality and effectiveness of problem reframing will only increase, further enhancing its value in an educational context.

In conclusion, our study illuminates the promising potential of ChatGPT in reframing academic problems across diverse disciplines. While challenges remain, our findings offer valuable insights for educators and chart a course for future research in this area. Our work contributes to the burgeoning literature on the application of AI in education, proposing a novel approach to enrich interdisciplinary education. The encouraging results we have observed prompt us to contemplate the future trajectory of this research: How can we further leverage the power of AI to transform the way we learn and teach across disciplines? This question, we believe, will steer the next wave of research in this thrilling intersection of AI and education, unveiling new pathways for enhancing learning experiences and outcomes.

Ethics statement

This research involved human participants in two surveys. Prior to their participation, all participantswere informed that the data collected would be used for research purposes and would be handled anonymously. All procedures were performed in compliance with relevant laws and guidelines for research ethics approved by the University Council of the University of Iceland. The study was conducted with due consideration for the ethical treatment of all participants, ensuring that all data remained anonymous.

Open data

This study is committed to the principles of open data. Accordingly, the dataset supporting the conclusions of this article is available in the supplementary files accompanying the publication. This dataset has been anonymized to ensure the confidentiality and privacy of the study participants.

CRediT authorship contribution statement

Hafsteinn Einarsson: Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing. **Sigrún Helga Lund:** Conceptualization, Data curation, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Anna Helga Jónsdóttir:** Conceptualization, Data curation, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to revise the text for clarity and accessibility. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgements

The authors extend their heartfelt gratitude to the 44 students and the 23 domain experts who generously contributed their time and expertise to this study.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.caeai.2023.100194.

References

Alhazmi, S., Hamilton, M., & Thevathayan, C. (2018). Cs for all: Catering to diversity of master's students through assignment choices. In *Proceedings of the 49th ACM technical symposium on computer science education* (pp. 38–43). New York, NY, USA: Association for Computing Machinery.

Anand, P. G., & Ross, S. M. (1987). Using computer-assisted instruction to personalize arithmetic materials for elementary school children. *Journal of Educational Psychology*, *79*, 72.

Chen, C. M., Lee, H. M., & Chen, Y. H. (2005). Personalized e-learning system using item response theory. *Computers and Education*, *44*, 237–255.

Chen, L. H. (2011). Enhancement of student learning performance using personalized diagnosis and remedial learning system. *Computers and Education*, *56*, 289–299.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*, 715.

Council, N. R., et al. (2000). Expanded edition. National Academies Press. How people learn: Brain, mind, experience, and school.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., et al. (2023). "so what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, Article 102642.

Eldan, R., & Li, Y. (2023). Tinystories: How small can language models be and still speak coherent English? arXiv preprint, arXiv:2305.07759.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint, arXiv:2303.10130.

Frieder, S., Pinchetti, L., Griffiths, R. R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., & Berner, J. (2023). Mathematical capabilities of ChatGPT. arXiv preprint, arXiv:2301.13867.

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., & Wan, X. (2023). Human-like summarization evaluation with ChatGPT. arXiv preprint, arXiv:2304.02554.

Gimpel, H., Hall, K., Decker, S., Eymann, T., Lämmermann, L., Mädche, A., Röglinger, M., Ruiner, C., Schoch, M., Schoop, M., et al. (2023). Unlocking the power of generative AI models and systems such GPT-4-4 and ChatGPT for higher education.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*, 64–74.

Hsu, C. K., Hwang, G. J., & Chang, C. K. (2013). A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. *Computers and Education*, *63*, 327–336.

Jacobs, H. H. (1989). *Interdisciplinary curriculum: Design and implementation*. ERIC.

Jonsdottir, A. H., Bjornsdottir, A., & Stefansson, G. (2017). Difference in learning among students doing pen-and-paper homework compared to web-based homework in an introductory statistics course. *Journal of Statistics Education*, *25*, 12–20.

Jonsdottir, A. H., Lentin, J., Calian, V., Hafsteinsson, E. K., & Stefansson, G. (2021). Using real data for statistics education in an open-source learning environment. *Teaching Statistics*, *43*, 5–12.

Jonsdottir, A. H., & Stefansson, G. (2014). From evaluation to learning: Some aspects of designing a cyber-university. *Computers and Education*, *78*, 344–351.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, Article 102274.

Kim, D., Yoon, M., Jo, I. H., & Branch, R. M. (2018). Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea. *Computers and Education*, *127*, 233–251.

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*, 129–164.

kaare Larsen, S. (2023). Creating large language model resistant exams: Guidelines and strategies. arXiv:2304.12203.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). GPTEval: NLG evaluation using GPT-4 with better human alignment. arXiv preprint, arXiv:2303.16634.

MacNeil, S., Tran, A., Leinonen, J., Denny, P., Kim, J., Hellas, A., Bernstein, S., & Sarsa, S. (2022). Automatically generating cs learning materials with large language models. In *Proceedings of the 54th ACM technical symposium on computer science education V: Vol. 2* (p. 1176).

Mangiafico, S. (2023). Rcompanion: Functions to support extension education program evaluation. Retrieved from https://CRAN.R-project.org/package=rcompanion, r package version 2.4.21.

Mollick, E. R., & Mollick, L. (2022). New modes of learning enabled by AI chatbots: Three methods and assignments. Available at SSRN.

Ochieng, P. (2023). Are large language models fit for guided reading? arXiv preprint, arXiv:2305.10645.

OpenAI (2023). GPT-4 technical report. arXiv:2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint, arXiv:2203.02155.

Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2015). *Continued progress: Promising evidence on personalized learning*. Rand Corporation.

Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S., & Pane, J. D. (2017). Informing progress: Insights on personalized learning implementation and effects. research report RR-2042-BMGF, RAND Corporation.

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Roschelle, J., Lester, J., & Fusco, J. (2020). AI and the future of learning: Expert panel report. Digital Promise.

Sancenon, V., Wijaya, K., Wen, X. Y. S., Utama, D. A., Ashworth, M., Ng, K. H., Cheong, A., & Neo, Z. (2022). A new web-based personalized learning system improves student learning outcomes. *International Journal of Virtual and Personal Learning Environments*, *12*, 1–21.

Símonarson, H. B., Snæbjarnarson, V., Ragnarson, P. O., Jónsson, H., & Thorsteinsson, V. (2021). Miðeind's WMT 2021 submission. In *Proceedings of the sixth conference on machine translation, association for computational linguistics* (pp. 136–139). Online. Retrieved from https://aclanthology.org/2021.wmt-1.9.

Tomlinson, C. A., Brighton, C., Hertberg, H., Callahan, C. M., Moon, T. R., Brimijoin, K., Conover, L. A., & Reynolds, T. (2003). Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted*, *27*, 119–145.

Trigwell, K., Prosser, M., & Waterhouse, F. (1999). Relations between teachers' approaches to teaching and students' approaches to learning. *Higher Education*, 57–70.

Troiano, E., Velutharambath, A., & Klinger, R. (2021). From theories on styles to their transfer in text: Bridging the gap with a hierarchical survey. *Natural Language Engineering*, 1–60.

Tseng, J. C., Chu, H. C., Hwang, G. J., & Tsai, C. C. (2008). Development of an adaptive learning system with two sources of personalization information. *Computers and Education*, *51*, 776–786.

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*, 932.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. Retrieved from https://ggplot2.tidyverse.org.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, *16*, 1–27.
Zhang, K., Bonk, C. J., Reeves, T. C., & Reynolds, T. H. (2019). *MOOCs and open education in the Global South: Challenges, successes, and opportunities*. Routledge.
Zhou, J., & Bhat, S. (2021). Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 conference on empirical methods in natural language processing, association for computational linguistics* (pp. 5075–5086). Retrieved from https://aclanthology.org/2021.emnlp-main.