frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 18 January 2024
DOI 10.3389/feduc.2023.1330486

Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving

Karen D. Wang[1]*, Eric Burkholder[2], Carl Wieman[1,3], Shima Salehi[1] and Nick Haber[1]

[1]Graduate School of Education, Stanford University, Stanford, CA, United States, [2]Department of Physics, Auburn University, Auburn, AL, United States, [3]Department of Physics, Stanford University, Stanford, CA, United States

OPEN ACCESS

EDITED BY
Jochen Kuhn,
Ludwig-Maximilians-Universität München,
Germany

REVIEWED BY
Peng He,
Michigan State University, United States
José Martín Molina-Espinosa,
Monterrey Institute of Technology and Higher Education (ITESM), Mexico

*CORRESPONDENCE
Karen D. Wang
✉ kdwang@stanford.edu

RECEIVED 02 November 2023
ACCEPTED 26 December 2023
PUBLISHED 18 January 2024

CITATION
Wang KD, Burkholder E, Wieman C, Salehi S and Haber N (2024) Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Front. Educ.* 8:1330486.
doi: 10.3389/feduc.2023.1330486

COPYRIGHT
© 2024 Wang, Burkholder, Wieman, Salehi and Haber. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The study explores the capabilities of OpenAI's ChatGPT in solving different types of physics problems. ChatGPT (with GPT-4) was queried to solve a total of 40 problems from a college-level engineering physics course. These problems ranged from well-specified problems, where all data required for solving the problem was provided, to under-specified, real-world problems where not all necessary data were given. Our findings show that ChatGPT could successfully solve 62.5% of the well-specified problems, but its accuracy drops to 8.3% for under-specified problems. Analysis of the model's incorrect solutions revealed three distinct failure modes: (1) failure to construct accurate models of the physical world, (2) failure to make reasonable assumptions about missing data, and (3) calculation errors. The study offers implications for how to leverage LLM-augmented instructional materials to enhance STEM education. The insights also contribute to the broader discourse on AI's strengths and limitations, serving both educators aiming to leverage the technology and researchers investigating human-AI collaboration frameworks for problem-solving and decision-making.

KEYWORDS

ChatGPT, GPT-4, generative AI models, problem-solving, authentic problems, STEM education, physics education

1 Introduction

The rapid advancement of Large Language Models (LLMs) has attracted substantial attention from both the general public and academia. LLMs, such as GPT-4 by OpenAI, can generate human-like textual responses to text-based queries in real-time. Since the public launch of ChatGPT in November 2022, there has been a growing body of research exploring its various capabilities, limitations, and implications across diverse disciplines and tasks. One such field is education, where LLMs have far-reaching implications for both instructional practices, or *how* we teach and assess; as well as for curriculum content, or *what* we teach and assess.

Broadly speaking, problem-solving refers to the process of finding the solution to a problem when the steps for solving are not known to the problem solver beforehand (Newell et al., 1972; Mayer, 1992). Extensive research has been conducted to study problem-solving in physics education (Ince, 2018). For example, using textbook-style physics problems, a seminal study by Chi et al. (1981) found that experts abstracted the physics principles underlying a problem as the basis for their problem-solving approach, while novices often based their approaches on the surface-level features. More recently, the physics education research community has begun to recognize the need for helping students acquire effective practices and strategies for handling real-world, authentic problems beyond traditional textbook exercises (Bao and Koenig, 2019; Burkholder et al., 2020). Such practices and strategies are key to prepare students for the complex challenges that they will encounter in their professional careers and daily lives.

In this study, we examined ChatGPT's capacity for solving problems from a college-level engineering physics course. ChatGPT by OpenAI is one of most accessible and publicly used LLM-based tools, and its most advanced underlying model to date is GPT-4. GPT-4 has outperformed previous models like GPT-3 in an array of standardized exams in disciplines such as law and medicine. Notably, it has achieved scores in the 66th to 84th percentile on the AP Physics 2 Exam (Achiam et al., 2023), which features problems that are mostly situated in abstract scenarios and provide all necessary data in the problem statement. However, the literature has so far offered limited insights into the capability of GPT-4 in solving problems that are in real-world contexts and/or do not provide all the data needed for reaching a solution. Consequently, the nuances of GPT-4's problem-solving capability, including the range of problems that it can effectively solve and the quality of the generated solutions, remain largely unknown.

Investigating GPT-4's problem-solving capability has multifaceted implications that extend from enhancing educational practices to fostering human-AI collaboration. First, a more nuanced understanding of how GPT-4 solves different types of problems can offer insights into how to design LLM-augmented instructional materials to support student problem-solving. In this study, we focused our attention on scientific problem-solving with the long-term goal of leveraging LLM-based tools to enhance Science, Technology, Engineering, and Mathematics (STEM) education. This focus stems from the recognition that, despite problem-solving being widely acknowledged as a fundamental learning goal in STEM education (NGSS, 2013), effective ways to teach problem-solving remain elusive and understudied. Second, as students start using LLM-based tools such as ChatGPT for their homework problems they need help with (Shoufan, 2023), they need to be educated about its affordances and limitations to make effective use of such tools for their own learning. Furthermore, beyond its value in educational settings, knowledge of GPT-4's problem-solving capability contribute to the broader discourse on human-AI collaboration. Understanding the areas where AI excels and where it currently falls short can inform the development of a human-AI collaborative problem-solving framework.

In this study, we pose the following research questions:

- How does ChatGPT's problem-solving capability vary across different types of physics problems?
- What are ChatGPT's common failure modes for different types of physics problems?
- To what extent can a standard prompt engineering technique improve ChatGPT's performance for different types of physics problems?

2 Background

Human problem-solving has been studied across diverse research traditions and domains, including cognitive psychology, information processing, and discipline-based education research (Newell et al., 1972; Chi et al., 1981; Reif and Heller, 1982; Bransford et al., 1986). As different types of problems call for distinct problem-solving strategies and bodies of knowledge, one's problem-solving capability may significantly vary across problem types. Similarly, to illuminate the problem-solving capability of AI models such as GPT-4, we must first explicate the characteristics of the problems given to the models, and study the performance of these models across different problem types.

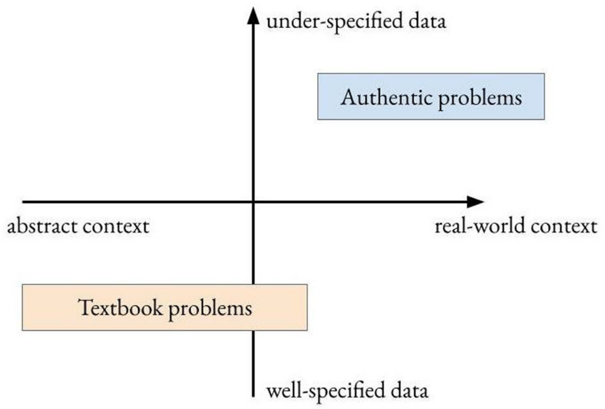

FIGURE 1
A two-dimensional plane to visualize the problem categorization framework. The x-axis represents the context dimension, ranging from abstract to real-world. The y-axis represents the data specificity dimension and ranges from well-specified to under-specified.

Our research group has done extensive work to characterize and assess authentic problem-solving expertise across science, engineering, and medicine domains (Salehi, 2018; Price et al., 2021, 2022). Drawing on these work, we now characterize problems in science and engineering domains along two dimensions: context and data specificity (Figure 1). The first dimension refers to the context where the problem is situated and spans from abstract to real-world. Abstract problems employ simplified, idealized scenarios that do not exist in the real world, such as frictionless planes and massless pulleys. On the other end of the spectrum are real-world problems that are based on scenarios that individuals may encounter in their daily lives or in professional settings. The second dimension is around the specificity of the data required to solve a problem. Well-specified problems provide all the data required for a solution, while under-specified problems lack some essential data, requiring the problem solver to determine what data is needed and how to obtain it for solving the problem. Textbook problems typically present well-specified data and may have either abstract or real-world context. These problems are designed to make it easier for learners to grasp and practice domain-specific concepts. In contrast, authentic problems bring with them the complexity and ambiguity that comes from real-world challenges and do not specify all the required data.

The above problem categorization framework is intended for analyzing problems that are knowledge-rich, or requiring the application of content knowledge from STEM disciplines. These problems differ from the classic knowledge-lean problems employed to study problem-solving in the information processing paradigm (Simon, 1973). The knowledge-lean problems, such as the Tower of Hanoi, are often termed as "well-defined" to indicate that they have clear initial and goal states and a set of clearly-defined operators for moving from the initial state to the goal state (Simon, 1978; Jonassen, 1997). It is important to differentiate "well-defined" and "ill-defined" from the "well-specified" and "under-specified" terminology we used in the problem categorization framework. The former terms capture the clarity of the initial and the goal states of a problem and the constraints on the possible operations to navigate from one to the other, while the latter terms are used for evaluating the quantity and clarity of data given in the problem statement.

Most existing research has focused on examining AI's performance in handling textbook-style problems that are well-specified and mostly abstract. For example, GPT-4 has performed well in standardized tests such as AP Biology, Chemistry, Environmental Science and Physics Exams (Achiam et al., 2023; Nori et al., 2023). The model also demonstrated proficiency surpassing average human performance in writing program functions that solely depend on existing public libraries (Bubeck et al., 2023). In contrast, there is a scarcity of research on how AI approaches authentic problems that are under-specified and situated in real-world contexts, even though such authentic problems are likely to constitute a significant share of the tasks that AI will encounter when deployed in the real world. Emerging research that ventures into the related domain has investigated AI's capacity for inductive reasoning, which involves identifying general principles from a small set of examples and applying these principles to novel situations (Gendron et al., 2023; Wang et al., 2023; Xu et al., 2023). Results of these investigations suggest significant room for improvement in AI's capability to make generalizations from specific instances.

While GPT-4's performance in solving textbook-style problems should not be extrapolated to its performance on authentic problems, a review of previous literature nonetheless provides insights into some of its common failure modes. One common flaw in GPT-4's performance is related to calculation errors. Previous studies have found that while the model can answer difficult high-school level math questions and discuss advanced mathematics concepts, it could also make basic errors in calculation (e.g., arithmetic mistakes) (Bubeck et al., 2023). Another limitation is the model's deficiency in critically evaluating its own solutions. This leads to failure in recognizing mistakes in its solution path (Bubeck et al., 2023; Zhang et al., 2023). A separate study employed GPT-3.5 and GPT-4 to answer open-domain questions, such as whether the New Orleans Outfall Canals are the same length as the Augusta Canal. The researchers summarized the models' failure modes in solving these problems into four categories: comprehension error, factualness error, specificity error, and inference error (Zheng et al., 2023). The study found that nearly half of the failures were due to factualness error, or the model lacking the necessary supporting facts to produce a correct answer, and another 25% of the failures were due to inference error, or the model failing to reason effectively.

In the context of physics education, a study reported that ChatGPT (based on the GPT-3 model) could narrowly pass a calculus-based college-level introductory physics course (Kortemeyer, 2023). One test used for evaluation was the Force Concept Inventory (FCI), which comprises well-specified multiple-choice questions. GPT-3 solved 60% (18 of 30) of the FCI items. Moreover, the researcher found that the model's performance variation was more influenced by the mathematics than the physics concepts involved. Similar to the above mentioned studies, this study found that ChatGPT had persistent problems with calculation, especially in manipulating and calculating formulas involving square roots.

In summary, the review of existing literature revealed a gap in our understanding of generative AI models' capability for solving real-world problems where data is often incomplete or ambiguous. The central aim of this paper is to explore how one AI model (GPT-4) perform across an array of real-world physics problems that vary in data specificity. Furthermore, we will investigate the model's common failure modes in solving these problems and evaluate whether a well-studied prompt engineering technique could improve AI's problem-solving performance.

3 Methods

3.1 Problems used in the study

A total of 40 homework problems from an engineering physics course taught by the second author were used in this study. The course is a calculus-based engineering physics 1 course taught at a public research-intensive university. The course is primarily taken by engineering, chemistry, and physics majors and covers an array of topics including static equilibrium (forces and torques), conservation of momentum and kinematics (linear and angular), conservation of energy, harmonic motion, mechanical waves, and fluid mechanics. This particular course was developed by the second author (Burkholder et al., 2022) and aims at developing students' problem-solving competencies. The course is designed on the theory of deliberate practice (Ericsson et al., 1993) and uses a template (Burkholder et al., 2020) to teach students real-world problem-solving skills.

The homework problems in this course are a mixture of textbook physics problems (for practice with basic calculations), problems that ask students to explain a physical phenomenon, and real-world problems that require students to make assumptions, seek out information, and make modeling decisions (Price et al., 2021). The real-world problems were designed to engage students in more deliberate reasoning with particular concepts rather than standardized procedures. They were also designed to have students practice turning real situations into manageable models, rather than providing simplified scenarios for them. We selected problems that were written by the second author specifically to engage students in real-world problem-solving, rather than the textbook or conceptual explanation problems. Based on our proposed problem categorization framework (Figure 1), we characterize these problems along the two key dimensions: context and data specificity. Regarding the context dimension, the problems are all situated in real-world contexts. For example, one problem involves calculating the total travel time for an elevator ascending to the top floor of the Salesforce Tower in San Francisco, and another involves selecting fishing lines that are strong enough to hang sculptures from the ceiling of an atrium of a new building.

Regarding the data specificity dimension, the problems used in this study span a spectrum from well-specified to under-specified. On one end of the spectrum are problems that provide all the data needed for solving, including values for key variables and parameters. On the other end are problems with under-specified or incomplete data, requiring the problem solver to determine what data is needed and how to collect the missing data. This variation in data specificity necessitates different levels of decision-making by the problem solver regarding data collection, which is a key practice for solving authentic problems as identified in our previous research (Salehi, 2018; Price et al., 2021). By incorporating this range of problem types, we are able to conduct a more comprehensive and nuanced evaluation of ChatGPT's problem-solving capability.

Table 1 presents two sample problems used in the study. Both problems are situated in real-world contexts. The first one is a well-specified problem where all data needed to solve the problem was provided in the problem statement. In contrast, the second one represents an under-specified problem where the problem statement does not provide any data, and necessitates the problem solver to collect all the required data through conducting an online query or making reasonable assumptions in order to solve the problem.

3.2 Experiments and analysis

We used ChatGPT with GPT-4 selected as the underlying model in the present study. The decision to use ChatGPT as opposed to running the experiments through OpenAI's API was grounded in the interest of face validity and ecological validity. Face validity refers to the appropriateness or relevance of a measurement method for its intended purpose (Nevo, 1985), while ecological validity refers to the extent to which the findings of a study can be generalized to the natural environments and real-world settings (Orne, 2017; Kihlstrom, 2021). Given students and instructors of STEM courses are more likely to use ChatGPT than to access the GPT-4 model directly through APIs, this methodological choice allows our study's findings to be more directly applicable to the common STEM educational settings where LLM-based tools are used.

Each problem statement of the 40 problems was pasted into the dialogue interface of ChatGPT, accompanied by the prompt of "solve the following physics problem." No additional guidelines or contextual knowledge was provided. If ChatGPT returned with queries or statements indicating that the problem could not be solved without additional information, a second prompt was put into the dialogue box (e.g., "please make reasonable assumptions about the missing information and solve the problem."). Once the model reached a final answer, its response was transferred to a centralized document for record and analysis. This approach was implemented to minimally influence ChatGPT's problem-solving approach and establish a baseline for its problem-solving capability.

In evaluating ChatGPT's performance on solving the physics problems in our data set, we adopted an approach similar to the one we use for grading students' worked-out solutions, where a single knowledgeable evaluator can effectively grade a student's solution. Our goal is to not only determine the accuracy of its final answers but also understand the steps it undertook to reach the answer. In cases where ChatGPT failed to reach the correct answer, we compared each step of its solution to the correct solution prepared by the course's lead instructor to determine where in the solution process it failed. The instructor's solutions loosely follow a template that we have devised to scaffold students' solving of authentic problems in physics and typically incorporate the following components:

- A clear representation of the problem highlighting its key features using a diagram or a set of bullet points
- Identification of the relevant physics concepts and formulas
- Noting all the information required for solving the problem, and for information not explicitly provided in the problem statement, noting how such information can be obtained through an internet search or a reasonable estimate based on prior knowledge
- Carrying out the necessary calculations to reach the correct answer.

It is important to note that we did not expect ChatGPT to follow the problem-solving template. Rather, the goal of comparing its solutions to the expert solutions is to determine the primary factor that led to the erroneous final answer. This analysis helps us ascertain whether the erroneous answer was due to a misrepresentation of the problem, a misapplication of physics principles, errors in the data used, or calculation errors. By pinpointing the chief reason behind the incorrect solutions, we aimed to gain a more nuanced understanding of the AI model's problem-solving potential and limitations. The results of this error analysis were recorded for each individual problem and collectively analyzed to identify patterns and recurrent themes in ChatGPT's problem-solving failures. To ensure both precision and thoroughness in the analysis of ChatGPT-generated solutions, the analysis was conducted by the first author, a researcher in STEM education, in close collaboration with the third author, who is an expert in physics and physics education.

Next, we examined whether simple prompt engineering could improve ChatGPT's problem-solving performance. In the context of AI research, prompt engineering refers to the process of designing, testing, and refining inputs given to AI models to enhance their performance (Liu et al., 2023). Prompting strategies such as zero-shot chain-of-thought, which involves literally telling the model to "think step-by-step," have demonstrated success in improving LLMs' performance in solving multi-step arithmetic word problems (Kojima et al., 2023). In the second phase of the study, we adopted a similar prompting strategy for the problems in our data set. Specifically, the prompt was updated to "solve the following physics problem step-by-step" just before presenting the problem statement to ChatGPT. The intention was to explore whether ChatGPT could decompose a problem into more manageable sub-problems and circumvent the errors it made during its initial problem-solving attempt. At the same time, it is important to note that the specific approach to decomposing

TABLE 1 Two sample problems used in the study.

The log cabin problem (well-specified)	The dresser tip-over problem (under-specified)
You are planning to build a log cabin and will need to pull the logs up a hill to the building site by means of a rope attached to a winch. In order to buy the rope, you need to know how strong the rope must be and decide to do a quick calculation for this. The heaviest of the logs weigh 500 lbs. You estimate the coefficient of friction between the log and the hill to be 0.8, and the hill you have to pull them up is at an angle of 30 degrees. How strong must the rope be?	IKEA has had some issues with children climbing their dressers by pulling on the drawers and getting hurt when the dressers tip over. Their solution has been to provide wall mounts that you can use to secure the top of the dresser to the wall. Figure out how strong this wall mount has to be to keep the dresser from tipping over. You should include an equation showing how your answer depends on the weight of the child and the size and weight of the dresser.

TABLE 2 ChatGPT's problem-solving performance grouped by the problems' data specificity.

Problem type	Number of problems	Number correctly solved	Accuracy rate
Well-specified	16	10	62.50%
Under-specified	24	2	8.33%

individual problems vary case-by-base, depending on the unique context and nature of each problem. Consequently, the efficacy of the prompt in improving the model's performance may not be consistent.

4 Results

In this section, we first present ChatGPT's problem-solving success rate without the use of prompt engineering. Next, we discuss three distinct failure modes based on a comparative analysis between ChatGPT's incorrect solutions and the instructor's correct solutions. Finally, we explore how prompt engineering impacted ChatGPT's problem-solving performance.

4.1 Problem-solving performance

Our analysis revealed a substantial difference in the ChatGPT's ability to solve well-specified vs. under-specified types of problems (Table 2). ChatGPT successfully solved 62.5% of the well-specified problems and only 8.3% of the under-specified problems. This discrepancy in accuracy rate was statistically significant (Fisher's exact test, $p < 0.001$).

At the same time, ChatGPT demonstrated a high level of proficiency in identifying the relevant physics concepts to apply based on the given problem statement. This capacity was evidenced by the model's consistent performance of outlining the relevant physics concepts at the beginning of the solutions it generated. ChatGPT's strength in this facet of problem-solving differs from typical human performance, as students often struggle to identify what concepts to apply as the starting point in solving unfamiliar problems. Additionally, students may struggle with complexities that arise in correctly identifying, applying and integrating domain knowledge learned at different times. In contrast, LLMs like GPT-4 are not constrained by such linear learning pathway, and their training data is likely to incorporate a more comprehensive range of domain-specific knowledge compared to what students learn in a typical college-level curriculum. This expansive knowledge base is one of AI's strengths in addressing real-world challenges.

4.2 Types of failure modes

A detailed table of all 40 problems used in the study, along with their underlying physics concepts and ChatGPT's performance, is available in the Supplementary material. In examining where all of ChatGPT's incorrect solutions diverged from the instructor's solutions, we identified three distinct types of failure modes. These failure modes can help us understand the underlying causes of the AI model's difficulties in problem-solving.

4.2.1 Failure to construct accurate models of the physical world

One of the failure modes is related to ChatGPT's failure to construct accurate models of the physical world based on the problem statement. Table 3 presents a problem from our dataset and ChatGPT's incorrect solution that falls into this type. The problem asks for an estimation of the force required from each of the 950 friction piles to prevent the Millennium Tower in San Francisco from sinking. This is an under-specified problem, as the problem statement does not specify the density of concrete. The highlighted texts in the problem statement are information that ChatGPT did not account for in its modeling of the problem.

A human problem solver could intuitively construct a model of the problem, that both the building and the concrete slab are on top of the 950 concrete piles, as depicted by the first model in Figure 2. This model helps the problem solver focus on the key objects and forces involved. The force required from each friction pile to support the building and concrete slab in addition to its own weight would therefore be: $F_{\text{friction}} = (W_{\text{building}} + W_{\text{slab}})/950 + W_{pile}$. In contrast, the second model in Figure 2 was created by the researcher based on ChatGPT's incorrect solution. ChatGPT failed to consider the weight of the concrete slab and piles in its calculations, indicating that it did not correctly conceptualize the spatial relationship between the building, the concrete slab, and the 950 piles. This led to an oversimplified and erroneous calculation of the force required from each pile.

Out of the 28 problems that ChatGPT did not solve, 14 can be attributed to this failure mode of it not being able to accurately model the physical world, especially regarding spatial relationships. Another example of this failure mode is the Dresser Tip-over problem that asks students to determine the force exerted by a wall mount to prevent a dresser from tipping over when children pull down on the drawers (see the problem in Table 1). Solving

TABLE 3 The Millennium Tower problem and ChatGPT's solution.

Prompt: Solve the following physics problem: The Millennium Tower in San Francisco was completed in 2009 and has received a lot of attention recently due to problems with the building settling and leaning. The building is 605 feet tall, has a base of about 20,000 square feet, and weighs about 7 tons per square foot of its foundation. The tower sits on a 10-ft thick concrete slab, which is in-turn supported by 950 friction piles, which are 14-inch square concrete pillars embedded in the bay sand. The piles are 80 feet long. Estimate the approximate force required from each friction pile to keep the building from sinking.
ChatGPT: To estimate the force each pile needs to exert, we need to determine the weight of the building and distribute that weight across the number of friction piles. The total weight of the building is given as 7 tons per square foot of its foundation. The foundation is ∼20,000 square feet. Therefore, the total weight is 7 tons/ft * 20,000 ft = 140,000 tons. Since 1 ton is equivalent to 2,000 pounds, the total weight of the building is 140,000 tons * 2,000 pounds/ton = 280,000,000 pounds. The force each pile needs to exert is 280,000,000 pounds/950 = 294,737 pounds (approximately), assuming an even distribution of the weight of the building.

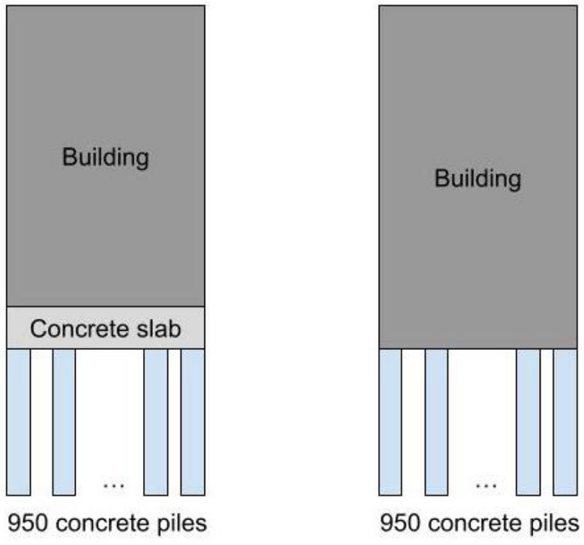

FIGURE 2
Two contrasting models for the Millennium Tower problem. The first one **(left)** is the accurate model and the second one **(right)** was created based on ChatGPT's incorrect solution.

the problem requires an accurate model representing the center of mass of the dresser, the pulling by a child, the location of the wall mount, and most importantly, the pivot point and the distances to the pivot point for each of the forces. One of its persistent challenges observed in our analysis is that ChatGPT cannot correctly identify the pivot point. In other words, it cannot correctly conceptualize how the dresser would fall if pulled by a child.

4.2.2 Failure to make reasonable assumptions about missing data

A second failure mode of ChatGPT in problem-solving was its limitations in making reasonable assumptions about missing data in under-specified problems. When presented with problems that lack complete data, ChatGPT usually attempted to fill in the missing data and generate a solution on its own. Only for two under-specified problems did it respond to the initial problem statement by stating that the problem doesn't provide all the necessary information to calculate the exact value for a solution. In such cases, a follow-up instruction was entered into ChatGPT's dialogue box, directing the model to make reasonable assumptions about the specific value it noted and solve the problem. Table 4 presents such an instance where ChatGPT explicitly listed the missing information.

ChatGPT was successful at filling in well-known factual information that was not provided in the problem statement, such as the speed of light in a vacuum, the speed of sound in water, the density of seawater, and even the height of the Salesforce Tower in San Francisco. However, its success rate dropped when the missing data involves more nuanced real-world understanding. In the Floating Duck problem presented above (Table 4), ChatGPT made an incorrect assumption about the density of a duck being 950 kg/m^3 even after receiving the prompt to make reasonable estimates. A more accurate density value can be inferred from the information that 20% of the duck's body is submerged in salt water, making its average density ∼260 kg/m^3.

Table 5 presents another problem that illustrates this failure mode. The problem asks a problem solver to estimate how long a driver should take to come to a stop while driving at 10 mph, so that a coffee cup that was accidentally left on the car's roof won't fly off. In analyzing ChatGPT's solution to this problem, we found that the model failed to provide a reasonable value for the friction coefficient between the coffee cup and the roof of the car (μ). While the typical ceramic-on-metal and metal-on-metal static coefficients of friction are between 0.4 and 0.6, ChatGPT assumed the friction coefficient to be 1, as implied by its equating acceleration to the gravitational constant g in the solution (highlighted in Table 5). This incorrect assumption led to an erroneous value of the horizontal acceleration, which in turn resulted in an underestimate of the time it takes to stop the car safely.

To its credit, ChatGPT did hedge on its solution by making the following statement at the end:

> ChatGPT: Please note that this calculation assumes that the friction between the cup and the roof is strong enough to keep the cup from sliding when the car is decelerating at a rate equal to g. The actual time could be much longer if the friction is weaker than assumed. This is a simplified model and doesn't account for all factors that might be present in a real-world situation.

This acknowledgment indicates that ChatGPT recognizes the potential error in it solution introduced by inaccurate assumptions. Nonetheless, we should be aware of its limitations, that it may

TABLE 4 An example of ChatGPT requesting more information for the Floating Duck problem.

Prompt: Solve the following physics problem: Off-shore oil-drilling can be controversial due to the possibility of oil spills. A sea duck can float on salt water with only about 20% of its body submerged below the surface. How thick would a layer of crude oil have to be for the duck to drown? For reference, the BP oil spill in 2010 reached a depth of 41.5 m.
ChatGPT: To solve this problem, we first need to understand the concept of buoyancy. Buoyancy is the force exerted on an object that is wholly or partly immersed in a fluid. In our case, the fluid is the combination of salt water and crude oil. The sea duck floats with only 20% of its body submerged in the water because the buoyancy force from the submerged part of the duck equals the weight of the duck. This means that the duck is in equilibrium. If a layer of oil were to cover the surface of the water, the duck would start to submerge into the oil, since oil is less dense than water. If the oil layer is thick enough, the duck could potentially sink below the surface and drown. To find the thickness of the oil layer that would cause the duck to drown, we need to find the point where the duck's buoyancy in the oil equals its weight. Here's where we run into a problem, though. To find this value, we need additional information that isn't provided in the problem: 1. The volume of the duck. This would give us the amount of oil (or water) displaced by the duck, which we could use to calculate buoyancy. 2. The density of the crude oil. Different types of crude oil can have different densities, and the density would affect the buoyancy. 3. The density of the duck. In order to figure out the weight of the duck, we would need to know both its volume and its density. This would allow us to figure out when the duck's weight equals the buoyancy force in the oil.
Prompt: Can you make some reasonable estimates and solve the problem?
ChatGPT: To make some reasonable estimates, let's say the following: 1. The volume of the duck: This can vary, but we can estimate an average duck to have a volume of 1 liter or 0.001 cubic meter. 2. The density of the crude oil: The density of crude oil varies based on its specific type and temperature, but a common value is around 850 kg/m^3. 3. The density of the duck: Most birds have a density slightly less than water due to their feathers and air-filled bones. Let's estimate it at about 950 kg/m^3, slightly less than the density of sea water which is about 1025 kg/m^3.

TABLE 5 The coffee cup on car problem and ChatGPT's solution.

Prompt: Solve the following physics problem: You are driving your electric car to work one morning and forget your coffee cup on the roof. As you leave your driveway, you speed up to 10 mph and then remember your coffee cup and decide you need to stop. If you stop too fast, the cup could go flying and hit a pedestrian. How long should you take to come to a stop to be sure the cup doesn't fly off the roof?
ChatGPT: To solve this problem, we need to consider the forces acting on the coffee cup, primarily gravity and friction. The friction force is given by the equation: $F_{\text{friction}} = \mu N$, where μ is the coefficient of friction and N is the normal force. However, we don't have a specific value for the coefficient of friction between the cup and the roof of the car. Let's assume that the coffee cup will stay on the roof if the car stops with an acceleration equal to or less than g (the acceleration due to gravity), which is ~9.81 m/s^2.

not make accurate assumptions about missing information from the outset. Such difficulty in making reasonable estimates for real-world quantities is also typical of novice students.

Another example that falls into this failure mode is ChatGPT's solution to a problem that involves estimating how the velocity of a merry-go-around would change once a person hops on after first pushing it around for 3 s. As the problem does not specify the initial speed of the merry-go-around, ChatGPT first estimated it to be around 14 mph through calculation. However, this estimate substantially exceeds the safe speed for merry-go-arounds, which are typically around 9 mph as stipulated by the US Consumer Product Commission. This example further illustrates that ChatGPT cannot always align and verify computational results with real-world standards and constraints. ChatGPT's solutions to eight problems had this failure mode.

4.2.3 Calculation error

While the first two failure modes are mostly associated with under-specified problems, the last failure mode, calculation error, has been observed in both well-specified and under-specified problems in the dataset. A total of six incorrect solutions, four for well-specified problems and two for under-specified problems, had this failure mode. The errors range from arithmetic to trigonometry, and underscore ChatGPT's well-known difficulties with mathematical computations.

TABLE 6 A problem for which ChatGPT reached the correct solution without using the data provided.

The horsepower problem
Engine power is sometimes expressed in terms of "horsepower." One horsepower was defined by James Watt, who observed that a horse could turn a mill wheel with a radius of 12 ft at a constant rate of 144 times per hour, exerting a nearly constant force of 800 N tangentially to the wheel. Derive the conversion for horsepower to kilowatts.

4.2.4 An idiosyncratic solution case

In addition to the three failure modes discussed above, we also identified an idiosyncratic case where ChatGPT reached the correct answer (therefore coded as correctly-solved) while completely disregarding the data provided in the problem. The problem is a well-specified one and involves the conversion of horsepower to kilowatts (Table 6).

Despite being provided with data in the problem statement, ChatGPT opted to utilize different data, that one horsepower is defined as the ability to lift 550 pounds one foot in one second, for its calculations. This behavior raises questions regarding how the underlying GPT-4 model potentially prioritizes its training data over new information in problem-solving.

4.3 Chain-of-thought prompt engineering

To what extent did prompt engineering enhance ChatGPT's problem-solving performance? In the second experiment, we

TABLE 7 ChatGPT's updated solution to the coffee cup on car problem after prompt engineering.

ChatGPT: We can analyze this scenario by considering the forces acting on the coffee cup. Since the car is accelerating in the opposite direction of its motion (decelerating), there will be a frictional force trying to keep the cup in place. Let's assume the coefficient of static friction between the cup and the car's roof is μ_s, and that this is the only force preventing the cup from sliding. 1. Find the maximum deceleration: The force of static friction is given by: $F_{\text{friction}} = \mu_s \cdot F_n$, where F_n is the normal force, which is equal to the weight of the coffee cup: $F_n = m{\cdot}g$. Here, m is the mass of the cup, and g is the acceleration due to gravity (9.8 m/s^2). The maximum static frictional force before the cup starts sliding is: $F_{\max} = \mu_s{\cdot}m{\cdot}g$. 2. Calculate the deceleration: This frictional force provides a maximum deceleration the car can undergo without the cup sliding. Using Newton's second law, $F = ma$: $\mu_s \cdot m \cdot g = m \cdot a$, Solving for a, we find: $a = \mu_s \cdot g$. 3. Find the time to stop: Now we have the maximum deceleration, we can calculate the time it will take to stop the car from 10 mph under this deceleration. Using the equations of motion: $v_f = v_i + at$, where $v_f = 0$ m/s (final velocity), $v_i = 10$ mph $= 4.47$ m/s (initial velocity), and $a = -\mu_s \cdot g$ (deceleration). Solving for t: $t = \frac{v_f - v_i}{a} = \frac{-4.47}{-\mu_s \cdot 9.8}$. We need the coefficient of static friction to calculate the exact value of t, which will tell us how long the driver should take to come to a stop to ensure the coffee cup doesn't fly off the roof. If we had that value, we could plug it into our equation to find the time. In real-world scenarios, the coefficient of static friction might vary widely depending on the materials in contact and the conditions, so without specific information, it's hard to give a precise time.
Prompt: Please make reasonable assumptions about the coefficient of friction and solve the problem.
ChatGPT: Certainly! To solve the problem, we'll make a reasonable estimate for the coefficient of static friction between the coffee cup and the car's roof. In general, the coefficient of static friction for smooth surfaces like a car roof and a typical coffee cup might range from 0.5 to 0.8. However, since a car roof is usually not designed to hold objects and the bottom of a coffee cup may be somewhat smooth, we'll use a more conservative estimate of $\mu_s = 0.6$. Now, we'll use our earlier formula for the time taken to stop: $t = \frac{-4.47}{-\mu_s \cdot 9.8}$ Substituting $\mu_s = 0.6$: $t = \frac{-4.47}{-0.6 \times 9.8} \approx 0.76$ s. So the driver should take ~0.76 s to come to a stop to ensure that the coffee cup doesn't fly off the roof. Of course, this is a simplified model, and in a real-world scenario, other factors such as wind resistance, the curvature of the roof, and the exact materials of the cup and roof would also influence the outcome.

applied the "solve the following physics problem step-by-step" prompt to all 40 problems in the dataset. Among the 12 problems that ChatGPT initially solved correctly, it generated consistent correct solutions for 11 of them under prompt engineering. However, ChatGPT made a calculation error involving trigonometry in one of the problems. Interestingly, in the idiosyncratic case where ChatGPT reached the correct answer without utilizing the given data in its initial solution, the step-by-step prompt helped it incorporate the data provided in the problem statement in its calculation for the correct solution.

Among the 28 problems that ChatGPT initially failed to solve, it was able to correctly solve three with the step-by-step prompting. Two of the three were related to the failure mode of ChatGPT not being able to construct accurate models about the real world. For the first one, ChatGPT initially did not subtract the weight of the water from a squid when it ejected water to create a form of jet-propulsion. For the other one, it initially treated a marble as a non-rotating block and did not account for the rotational kinetic energy as it rolled up a ramp. The last one of the three was associated with the initial failure mode of not being able to make reasonable assumptions about missing data, in this case the coefficient of friction in the Coffee Cup on Car problem. Table 7 presents ChatGPT's updated solution. The prompt of solving the problem step-by-step led to more precise and deliberate problem-solving as illustrated in this example. ChatGPT first broke down the solution process into discrete steps, then noted that without knowing the coefficient of static friction, a specific numerical answer could not be provided. After receiving a follow-up instruction through the dialogue input box to make reasonable assumptions, it chose a reasonable value of 0.6 for the friction coefficient and successfully solved the problem, unlike what it did in the absence of prompt engineering.

Overall, the results suggest that prompt engineering had a moderate effect on enhancing ChatGPT's problem-solving performance by constructing accurate models of the problem and making reasonable assumptions, though this effect is not statistically significant (Chi-squared (1) = 0.06, $p = 0.81$). It should also be noted that step-by-step prompts had no impact on reducing calculation errors.

5 Discussion

The present study found a marked difference in ChatGPT's problem-solving performance between well-specified and under-specified problems. The problems used in the study are all situated in real-world contexts and require the application of physics knowledge, yet differ in how much information is specified in the problem statement. ChatGPT performed better in well-specified problem, although it made occasional calculation errors. In contrast, it was far less accurate in solving under-specified problems. Two specific failure modes were observed: the first one being failure to construct accurate models of the physical world and reason about relationships between different variables in a model, and the second one being failure to make reasonable estimates or assumptions about the missing data. Prompt engineering produced a moderate improvement in ChatGPT's problem-solving performance. The prompt of solving a problem step-by-step proved moderately beneficial in guiding the AI model to be more deliberate and accurate in estimating missing data and constructing models of the problems, though it did not alleviate calculation errors.

5.1 Implications for education

The problem-solving process adopted by experts in science and engineering domains can be characterized as a series of interlinked decisions (Price et al., 2021). Utilizing this framework to analyze ChatGPT's performance on solving problems situated in real-world contexts, we note that ChatGPT (based on GPT-4 model) demonstrated proficiency in deciding on the relevant domain-specific concepts and formulas based on a problem statement. At the same time, it fell short in making several key decisions, including determining how to construct a suitable model of a problem, and deciding how to make reasonable assumptions or estimates about incomplete data.

These results have significant implications for STEM education, especially around how to leverage LLM-based tools like ChatGPT to help students develop expertise in problem-solving. First, the study identified facets of problem-solving where ChatGPT is indeed effective, namely identifying the relevant physics concepts needed for solving a problem based on the problem statement. This opens the possibility for ChatGPT to serve as a tutor for domain-specific problems and support students to pinpoint the essential knowledge underlying each problem and enhance their understanding of conceptual knowledge through problem-solving. This tutoring capability is particularly important as students struggle to decide on relevant physics concepts and formulas through analyzing the problem's statement, instead they rely on ineffective strategies such as searching for equations that contain the same variables to solve problems (Ogilvie, 2009; Burkholder et al., 2020). Given the capability of ChatGPT in deciding on relevant concepts, students can query ChatGPT with prompts such as "identify the relevant concepts associated with the following problem." However, one concern associated with this use case is that ChatGPT may generate articulate, plausible-sounding, yet incorrect solutions based on the identified concepts. This presents a risk of misleading students and inducing misconceptions. Therefore, it is crucial to educate students on the problem-solving capabilities of ChatGPT (e.g., identifying the relevant concepts) as well as its shortcomings (e.g., generating inaccurate solutions due to failure to construct accurate models, failure to make reasonable assumptions, or calculation errors, particularly in the case of under-specified problems).

Second, the findings point to what we should prioritize in STEM education in an era of increasingly powerful AI technologies. To prepare students for solving authentic problems in their professional and personal lives, STEM courses must place an emphasis on fostering effective decision-making practices. Specifically, students must have opportunities to practice making decisions related to construct appropriate models based on complex, real-world scenarios, as well as practice making decisions on what data is needed for solving a given problem, how to collect the data, and how to critically evaluate data quality. Mastery in these decisions will help students decompose complex, under-specified real-world challenges into a series of tractable, well-specified sub-problems for AI tools like ChatGPT to solve. The emphasis on developing problem-solving and decision-making expertise aligns well with the broader educational goal of preparing students to navigate a future of human-AI collaboration.

Lastly, our findings have immediate implications for how to design homework and exam problems that are resilient to automatic solving by tools like ChatGPT. The key strategy involves incorporating authentic problems into teaching and assessment materials. These problems are not solvable by ChatGPT alone, and necessitate students to make informed decisions on how to utilize ChatGPT as a tool. At the same time, students must remain actively involved in constructing accurate models in real-world contexts and handle under-specified information. The inclusion of such authentic problems allows for a more valid assessment of student competencies in STEM courses.

5.2 Implications for human-AI collaboration

This study also provides insights for the future of human-AI collaboration. While LLMs like GPT-4 can solve well-specified problems, albeit with occasional calculation errors, human intervention is needed to provide contextual understanding and nuanced judgement that AI currently lacks, particularly when navigating the complexities and ambiguities associated with authentic problems. This insight suggests a complementary relationship between human intelligence and artificial intelligence in addressing complex, authentic problems in the real world. Specifically, human experience and expertise can help construct accurate models of the physical world and make reasonable estimates or data collection plans for missing information. At the same time, AI's computational capability to instantly sift through vast knowledge bases and pinpoint the relevant domain knowledge constitutes an important asset to support human problem-solving.

6 Limitations

One potential limitation of this study is that we did not have two researchers independently evaluate all ChatGPT-generated solutions to assess inter-rater agreement. Instead, our analysis method mirrored the approach used in grading physics coursework, where a single knowledgeable evaluator compares students' solutions with expert-generated correct solutions. To mitigate the risk of bias, we also adopted a close, interdisciplinary collaboration in the analysis process.

Next, in evaluating ChatGPT's capacity for problem-solving, it is important to recognize the inherent limitations associated with the underlying algorithm's probabilistic nature. ChatGPT may generate different answers each time a problem is posed, and this variability presents a challenge in our analysis of its solutions. The different releases and incremental builds of the algorithm could further produce varied results. Therefore, the interpretation of our findings must consider the specific version of the algorithm utilized, which spans from May to August 2023. Additionally, the current study did not ask ChatGPT to generate solutions for identical problems and prompts multiple times. This absence of repetitive testing restricts our understanding of the tool's stability and reliability in providing consistent solutions. The probabilistic and evolving nature of LLMs underscore the need for continuous evaluation and validation of their problem-solving capabilities in future studies.

7 Conclusion

This study probed the capabilities and limitations of LLM-based technologies such as ChatGPT in solving authentic problems that are situated in real-world contexts and under-specified in terms of the requisite data. By focusing on the domain of physics, we were able to incorporate a diverse set of real-world scenarios into the problem set. The problem-solving practices and processes adopted to solve these physics problems are also applicable in the broader fields of science and engineering. Furthermore, the decision to include problems from well-specified to under-specified in terms of the amount of information provided in the problem statement led to a nuanced understanding of ChatGPT's capacity for solving different types of problems. The findings revealed that ChatGPT is adept at identifying relevant physics knowledge and applying it to solve well-specified problems. At the same time, its performance is less robust in modeling real-world complexities and making reasonable assumptions when data is missing in under-specified problems.

These findings lead to future studies to investigate how LLMs can be utilized in STEM education to support student learning, such as serving as personalized tutors to scaffold students in identifying the relevant knowledge for solving a problem. Additionally, the insights from this study shed light on what are the key competencies for students to develop to prepare for a future where AI can effectively address well-specified problems. These competencies include the ability to construct accurate and concise models of problems, make deliberate decisions regarding assumptions and estimates, and devise plans for data collection. Students' mastery of these competencies, in conjunction with the advancement of AI technologies, potentially pave the way for a future where human-AI collaboration can effectively address complex challenges in the real world.

Data availability statement

Full datasets from this study are not publicly available due to being used for course homework, and will be provided upon request by the corresponding author.

Author contributions

KW: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Writing—original draft, Writing—review & editing. EB: Conceptualization, Methodology, Writing—review & editing. CW: Conceptualization, Funding acquisition, Supervision, Writing—review & editing. SS: Conceptualization, Funding acquisition, Supervision, Writing—review & editing. NH: Conceptualization, Funding acquisition, Methodology, Supervision, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the Stanford Accelerator for Learning and the Stanford Institute for Human-Centered AI.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2023.1330486/full#supplementary-material

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). GPT-4 technical report. *arXiv preprint* arXiv:2303.08774.

Bao, L., and Koenig, K. (2019). Physics education research for 21st century learning. *Discip. Interdscip. Sci. Educ. Res.* 1, 1–12. doi: 10.1186/s43031-019-0007-8

Bransford, J., Sherwood, R., Vye, N., and Rieser, J. (1986). Teaching thinking and problem solving: research foundations. *Am. Psychol.* 41, 1078. doi: 10.1037/0003-066X.41.10.1078

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: early experiments with gpt-4. *arXiv* [Preprint]. doi: 10.48550/arXiv.2303.12712

Burkholder, E., Miles, J., Layden, T., Wang, K., Fritz, A., Wieman, C., et al. (2020). Template for teaching and assessment of problem solving in introductory physics. *Phys. Rev. Phys. Educ. Res.* 16, 010123. doi: 10.1103/PhysRevPhysEducRes.16.010123

Burkholder, E., Salehi, S., Sackeyfio, S., Mohamed-Hinds, N., and Wieman, C. (2022). Equitable approach to introductory calculus-based physics courses focused on problem-solving. *Phys. Rev. Phys. Educ. Res.* 18, 020124. doi: 10.1103/PhysRevPhysEducRes.18.020124

Chi, M. T., Glaser, R., and Rees, E. (1981). "Expertise in problem solving," in *Advances in the Psychology of Human Intelligence*, ed. R. J. Sternberg (Hillsdale, NJ: Erlbaum), 7–76.

Ericsson, K. A., Krampe, R. T., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychol. Rev.* 100, 363–406. doi: 10.1037/0033-295X.100.3.363

Gendron, G., Bao, Q., Witbrock, M., and Dobbie, G. (2023). Large Language Models are Not Abstract Reasoners. *arXiv preprint* arXiv:2305.19555.

Ince, E. (2018). An overview of problem solving studies in physics education. *J. Educ. Learn.* 7, 191–200. doi: 10.5539/jel.v7n4p191

Jonassen, D. H. (1997). Instructional design models for well-structured and iii-structured problem-solving learning outcomes. *Educ. Technol. Res. Dev.* 45, 65–94. doi: 10.1007/BF02299613

Kihlstrom, J. F. (2021). Ecological validity and "ecological validity." *Perspect. Psychol. Sci.* 16, 466–471. doi: 10.1177/1745691620966791

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. (2023). Large language models are zero-shot reasoners. *Adv. Neural Inform. Process. Syst.* 35, 2219922213.

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Phys. Rev. Phys. Educ. Res.* 19, 010132. doi: 10.1103/PhysRevPhysEducRes.19.010132

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 135. doi: 10.1145/3560815

Mayer, R. E. (1992). *Thinking, Problem Solving, Cognition*. New York, NY: WH Freeman/Times Books/Henry Holt & Co.

Nevo, B. (1985). Face validity revisited. *J. Educ. Meas.* 22, 287–293. doi: 10.1111/j.1745-3984.1985.tb01065.x

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*, Volume 104. Cliffs, NJ: Prentice-hall Englewood.

NGSS (2013). *Next Generation Science Standards: For States, by States*. Washington, DC: National Academies Press.

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. *arXiv preprint* arXiv:2303.13375.

Ogilvie, C. A. (2009). Changes in students' problem-solving strategies in a course that includes context-rich, multifaceted problems. *Phys. Rev. ST Phys. Educ. Res.* 5, 020102. doi: 10.1103/PhysRevSTPER.5.020102

Orne, M. T. (2017). "On the social psychology of the psychological experiment: with particular reference to demand characteristics and their implications," in *Sociological Methods*, ed. N. JK. Denzin (London: Routledge), 279–299. doi: 10.4324/9781315129945-26

Price, A., Salehi, S., Burkholder, E., Kim, C., Isava, V., Flynn, M., et al. (2022). An accurate and practical method for assessing science and engineering problem-solving expertise. *Int. J. Sci. Educ.* 44, 2061–2084. doi: 10.1080/09500693.2022.2111668

Price, A. M., Kim, C. J., Burkholder, E. W., Fritz, A. V., and Wieman, C. E. (2021). A detailed characterization of the expert problem-solving process in science and engineering: guidance for teaching and assessment. *CBE–Life Sci. Educ.* 20, ar43. doi: 10.1187/cbe.20-12-0276

Reif, F., and Heller, J. I. (1982). Knowledge structure and problem solving in physics. *Educ. Psychol.* 17, 102–127. doi: 10.1080/00461528209529248

Salehi, S. (2018). *Improving Problem-Solving through Reflection*. Stanford, CA: Stanford University.

Shoufan, A. (2023). Exploring students' perceptions of ChatGPT: thematic analysis and follow-up survey. *IEEE Access* 11, 38805–38818. doi: 10.1109/ACCESS.2023.3268224

Simon, H. A. (1973). The structure of ill structured problems. *Artif. Intell.* 4, 181–201. doi: 10.1016/0004-3702(73)90011-8

Simon, H. A. (1978). "Information-processing theory of human problem solving," in *Handbook of Learning and Cognitive Processes*, Volume 5, ed. W. K. Estes (Abingdon: Taylor & Francis), 271–295.

Wang, R., Zelikman, E., Poesia, G., Pu, Y., Haber, N., and Goodman, N. D. (2023). Hypothesis Search: Inductive Reasoning with Language Models. *arXiv preprint* arXiv:2309.05660.

Xu, Y., Li, W., Vaezipoor, P., Sanner, S., and Khalil, E. B. (2023). LLMS and the abstraction and reasoning corpus: successes, failures, and the importance of object-based representations. *arXiv* [Preprint]. doi: 10.48550/arXiv.2305.18354

Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. A. (2023). How Language Model Hallucinations can Snowball. *arXiv preprint* arXiv:2305.13534.

Zheng, S., Huang, J., and Chang, K. C.-C. (2023). Why does ChatGPT Fall Short in Providing Truthful Answers? *arXiv preprint* arXiv:2304.10513.