Education and Science

Vol 51 (2026) No 225 227-264

Original Article

Examining chatbot-generated responses on heat and temperature: misconceptions, consistency, and conceptual change

Zübeyde Demet Kırbulut Güneş [1], Bilal Güneş [2]

Abstract

This study examined whether responses generated by chatbots (ChatGPT-3.5, ChatGPT-4, and Bard) about heat and temperature match misconceptions identified in the literature and how these responses compare to those of learners. The study also addressed the effect of Conceptual Change Texts (CCTs) on chatbot-generated responses about heat and temperature, focusing on their relevance to prompt engineering. Heat and Temperature Four-tier Misconception Test (HTMCT) and CCTs were utilized from a previous study that investigated the effectiveness of CCTs in remedying misconceptions about heat and temperature held by pre-service physics teachers. The HTMCT, consisting of 20 items, was designed to diagnose misconceptions about heat and temperature held by pre-service physics teachers as identified in the literature, with each misconception being assessed using multiple items. In this study, the HTMCT was used to diagnose the chatbots' responses of the heat and temperature concepts before and after the implementation of CCTs. In addition, in-depth interviews with the chatbots were conducted to elaborate on their responses. Pre-service physics teachers in the prior study exhibited misconceptions about heat and temperature, which were effectively remediated by CCTs, leading to significant overall improvements. Similarly, this study found that chatbot-generated responses, except those from Bard, were prone to misconceptions. ChatGPT-4 consistently generated responses that aligned with the scientific paradigm, unlike the other two chatbots. However, pre- and post-test data revealed that ChatGPT-4-generated responses were prone to a misconception, specifically that equal amounts of heat supplied to different substances will result in the same final temperature, and these responses consistently reflected this misconception. Both ChatGPT-3.5 and Bard showed improved performance between the pre- and post-test data, despite providing inconsistent responses. While chatbots could generate responses that accurately expressed concept definitions, they struggled with drawing conclusions based on multiple scientific

Keywords

Science education
Artificial intelligence chatbots
Prompt engineering
Heat and temperature
Misconception
Conceptual change approach
Four-tier test

Article Info

Received: 09.03.2024
Accepted: 02.06.2025
Published Online: 01.31.2026

DOI: 10.15390/ES.2026.2515

[1] Gazi University, Gazi Faculty of Education, Department of Mathematics and Science Education, Ankara, Türkiye, zdgunes@gazi.edu.tr

[2] Gazi University, Gazi Faculty of Education, Department of Mathematics and Science Education, Ankara, bgunes@gazi.edu.tr

concepts, applying concepts to real-world scenarios, and engaging in complex reasoning. In this study, while the algorithms underlying the chatbots remain undisclosed, the post-test responses for all chatbots showed a notable decrease in incorrect responses and improved alignment with scientific knowledge, suggesting a positive influence of CCTs, akin to findings from the prior study.

Introduction

The COVID-19 pandemic has led to a widespread shift toward the use of online technologies for teaching and learning (Sartika et al., 2021; Whalley et al., 2021). Artificial intelligence (AI) has been making its way into our daily lives, including through computers and mobile devices (Ramos et al., 2008). AI is a general term that encompasses several technologies and methods, such as machine learning, deep learning, and natural language processing (Baker & Smith, 2023). Large language models (LLMs) based on generative artificial intelligence, such as ChatGPT and Google's Bard, have gained widespread attention in education (Alasadi & Baiz, 2023; Labadze et al., 2023; O'Dea, & O'Dea, 2023). These generative chatbots perform human-like cognitive tasks by analyzing very large amounts of data (Zawacki-Richter et al., 2019).

In science education, generative chatbots are being explored for their potential to support teaching and learning. Previous studies have investigated their use in various contexts, including assessing chatbot responses to exam questions (Clark, 2023; Fergus et al., 2023), using chatbot-generated responses as prompts for student critique (Exintaris et al., 2023), integrating chatbots in chemistry labs (Humphry & Fuller, 2023), and analyzing chatbot reasoning (Talanquer, 2023). Despite this growing interest, an underexplored area is the role of chatbots in addressing one of the most persistent challenges in science education: misconceptions about fundamental scientific concepts such as heat and temperature, which are essential for understanding thermodynamic principles and their applications across disciplines (Tiberghien, 1994). Building on this foundation, this study investigated chatbot responses (ChatGPT-3.5, ChatGPT-4, and Bard) related to the concepts of heat and temperature, examining their alignment with misconceptions identified in the literature and comparing these responses to those of learners. This is of particular significance for the field of science education, as it provides insight into how instructors can utilize chatbots to design formative assessment activities that elicit responses, challenge assumptions, and enhance learner reasoning. Students can discern accurate information and correct errors by subjecting chatbot responses to misconception tests to critical evaluation, thereby fostering a deeper and more robust understanding of scientific concepts. This manner of instruction facilitates the remediation of misconceptions and enhances the overall learning experience by encouraging critical thinking and active engagement with the subject matter (Exintaris et al., 2023; Rudolph et al., 2023). In this study, we adapted Çelik's (2022) methodology by employing a misconception test, originally constructed based on the literature on misconceptions about heat and temperature, to assess chatbot-generated responses and by utilizing Conceptual Change Texts (CCTs) to address these misconceptions. This approach enabled a comparative analysis to determine whether chatbot responses about heat and temperature aligned with the misconceptions identified in the literature and among pre-service physics teachers and to assess the effectiveness of CCTs in addressing these misconceptions across two different contexts: pre-service physics teachers and chatbot-generated responses.

Prompt engineering ensures that chatbots generate accurate responses by refining the way questions or inputs are structured. While some chatbots, like Bard, can update their responses based on input data, the key goal of prompt engineering remains consistent: to enhance the precision and relevance of chatbot-generated information (Brown et al., 2020). Similarly, CCTs are widely used in science education to remedy misconceptions and present accurate scientific concepts (Duit & Treagust, 2003). Both prompt engineering and CCTs share the objective of presenting clear, scientifically sound information (Duit & Treagust, 2003; Liu et al., 2023). This study investigated how CCTs could refine chatbot responses about heat and temperature, with the aim of improving the quality of educational content provided by AI-driven tools.

Background

This section provides an overview of the key concepts underlying this study. In particular, it focuses on (1) misconceptions about heat and temperature, exploring their prevalence and importance in science education; (2) prompt engineering, highlighting its role in optimizing chatbot responses for educational purposes; and (3) conceptual change texts (CCTs), highlighting their usefulness in addressing learners' misconceptions. Complementing the background discussion, the research aim section outlines the aims of the study and provides the basis for addressing the guiding questions of the study.

Misconceptions About Heat and Temperature

Perceptions are initially formed by observing one's environment and nature. For instance, without any other input, someone who observes the process from sunrise to sunset may conclude that the Earth remains motionless while the sun rises in the east and falls in the west. Indeed, in the past, scientists and philosophers have drawn such conclusions based on their observations of nature. During the 17th and 18th centuries, scientists concluded that all combustible substances contained an invisible matter called phlogiston. According to this theory, phlogiston was thought to escape from a burning substance and dissolve into the air. However, over time, scientific advancements have refined our understanding of the motion of the Earth and other planets in the solar system and led to a more accurate comprehension of combustion events, one that does not reference the outdated concept of phlogiston.

Misconceptions, like those about the Earth's motion or phlogiston, can create cognitive conflict, which affects how new information is processed. Learners often need to recognize and resolve these conflicts for effective learning (Posner et al., 1982). Some observations can encourage misconceptions in certain subjects, one of the best examples being the concept of heat. The misconception is that heat is matter and a quantity that a substance can possess, which is similar to the ideas of ancient philosophers and scientists. In this example, the belief is that heat is a tangible substance, akin to air or fluid, that can be added to or removed from objects. Studies have shown that such misconceptions about heat, temperature, force, and motion are persistent and can lead to systematic errors in understanding (Driver et al., 1994). Many learners also have difficulty distinguishing between "heat" and "temperature" and often use them as synonyms (Aydoğan et al., 2003; Çelik, 2022; Güneş, 2021).

Prompt Engineering

Prompt engineering is a fundamental technique of natural language processing (NLP), crucial for guiding large language models (LLMs) such as ChatGPT-3.5, ChatGPT-4, and Bard, to generate desired outputs. This technique is essential for optimizing the performance and utility of these models in various applications, including chatbots and automated content generation. By refining input prompts, researchers and practitioners can elicit more accurate, relevant, and contextually appropriate responses from LLMs. Carefully designed prompts leverage the inherent capabilities of LLMs more effectively, enabling them to perform complex tasks with minimal additional training. In the realm of maximizing the potential of large language models, prompt engineering is indispensable. Practitioners can substantially enhance the quality and relevance of model outputs by emphasizing clarity, context, structure, and iterative refinement. This technique also improves LLM performance and broadens their applicability across diverse fields, from automated content generation to educational tools (Brown et al., 2020).

Conceptual Change Text (CCT)

CCTs are one of the most effective methods for remediating misconceptions (Chambers & Andre, 1997; Çelik, 2022; Hynd & Alvermann, 1986; Wang & Andre, 1991). CCTs are based on conceptual change models. Researchers have proposed various conceptual change models (Chi, 1992; diSessa, 1988; diSessa, 2008; Mortimer, 1995; Posner et al., 1982; Ueno, 1993; Vosniadou, 1992). According to Posner et al. (1982) model, which is one of the most widely used conceptual change models in education, for conceptual change to occur, a learner must be dissatisfied with their current conception, then the learner must recognize a new concept as intelligible, plausible, and fruitful. CCTs introduce possible misconceptions in a given context and explain flaws to prompt dissatisfaction. They present scientific concepts clearly, making them plausible, and offer various contexts to enhance understanding and applicability (Chambers & Andre, 1997; Güneş, 2021). CCTs represent a promising intervention method to remedy misconceptions about heat and temperature (Aydoğan et al., 2003; Çelik, 2022). A comparison of prompt engineering and the use of CCTs in science education reveals a common underlying approach: methodically designing the input, whether prompts for language models or instructional texts for students, to enhance the quality and effectiveness of the output. Both techniques emphasize the importance of clarity, context, and iterative refinement to achieve their respective goals. In essence, prompt engineering improves chatbot performance, while CCTs enhance learning by addressing and remedying misconceptions, thereby improving conceptual understanding in educational settings (Duit & Treagust, 2003; Liu et al., 2023).

Research Aim

The ability to distinguish between true and false and between scientific and non-scientific information on the internet is becoming increasingly important with the widespread use of AI, making Information and Communication Technology (ICT) skills essential. Many countries, with the goal of all internet users having these skills, have systematically included ICT skills in their curricula. Presentations of the concepts of heat and temperature, which are directly related but not the same, on the internet are based on scientific and non-scientific information sources, such as in textbooks (Leite, 1999) and visual and written media (Allchin, 2023). For these reasons, we carried out a study on heat and temperature to check if the information in AI-based applications reinforced misconceptions about heat and temperature. A clearer comprehension of AI responses concerning cognitive features can aid in successfully integrating of emerging technologies in educational settings. We aimed to investigate whether the information provided to users by AI-based applications aligns with the misconceptions identified in the literature, how these responses compare to those of learners, and the effect of CCTs on chatbot-generated responses about heat and temperature. In this study, "learners" specifically refers to pre-service physics teachers, as examined in Çelik's (2022) study, where their conceptual understanding of heat and temperature provided a valuable benchmark for identifying and addressing misconceptions. The research questions that guided this study are as follows:

- Do chatbot-generated responses about heat and temperature correspond to misconceptions identified in the literature?
- Do chatbot-generated responses match pre-service physics teachers' misconceptions about heat and temperature?
- How do conceptual change texts affect chatbot-generated responses about heat and temperature concepts?
- Is a chatbot consistent in its responses to the concept test?

Method

The ChatGPT-3.5, ChatGPT-4, and Google Bard 1.0.0 chatbots were used in the study. The Heat and Temperature Four-tier Misconception Test (HTMCT) (Güneş, 2020) was used to diagnose the chatbots' responses regarding heat and temperature concepts (See Appendix for the HTMCT). First, we opened a new chat page in all chatbots to administer the HTMCT as a pre-test on 3 November 2023. Following this, we provided CCTs to remedy the chatbots' responses in case the chatbot-generated responses were in line with the misconceptions identified in the literature. The chatbots were tasked with offering suggestions for enhancing their responses, revealing varying degrees of receptiveness to learning. While ChatGPT-3.5 and ChatGPT-4 acknowledged the inability to integrate information from the prompt to update LLM, they still identified avenues for refining their responses. In contrast, Bard exhibited a high level of openness to learning, expressing readiness to adapt and improve based on the provision of accurate and high-quality information with the prompt. Therefore, because of their structured design, the CCTs were used as a tool to guide the chatbot's responses about the concepts of heat and temperature, not as a training model. In this study, the CCT, which consists of five distinct steps, was presented as a single, unified textual prompt rather than as separate stages, as is standard practice in such applications. By leveraging this unique approach, the CCT is believed to enhance the clarity and accuracy of chatbot responses. Additionally, responses indicated that remedying inaccuracies might require several days or even a week, thus justifying the one-week interval between the pre- and post-test assessments. One week later, on 9 November 2023, the HTMCT was administered as a post-test. In-depth interviews were then conducted to address the chatbot-generated responses that were consistent with the misconceptions. All tasks were completed on the same chat page. The pre- and post-test data were coded independently by the authors of this article, and five discrepancies between coders out of 120 coded data (20-item HTMCT for all three chatbots and for both pre- and post-test data) were observed and resolved through discussion; finally, the coding agreed upon by both authors was implemented.

Heat and Temperature Four-tier Misconception Test (HTMCT)

The HTMCT was developed as part of a master's thesis supervised by the second author of this study (Güneş, 2020). The HTMCT measures each misconception using more than one item. There are 20 items in the test. In the first tier of the test, one option is the correct response, the other options are misconceptions and distractors. The HTMCT was designed to diagnose seven common misconceptions (Çelik, 2022; Güneş, 2021) held by pre-service physics teachers, as identified in the literature (see Table 1).

Table 1. Heat and Temperature Misconceptions

Misconception Number	Misconceptions
MC[a]1	The skin is a good thermometer.
MC2	A temperature of 40°C is twice that of 20°C.
MC3	When equal amounts of heat are supplied to different substances, they will reach the same final temperature.
MC4	Heat and temperature are the same.
MC5	When two liquids are mixed, the temperature of the resulting mixture is equal to the sum of the temperatures of the individual liquids.
MC6	The temperature of any substance that receives heat increases. / The temperature of any substance that gives off heat decreases.
MC7	Woolen clothes give heat to you.

[a]MC stands for misconception

In the first tier of the test, multiple-choice questions were asked about the misconceptions shown in Table 1. In the second tier, the level of confidence in the responses given in the first tier was questioned. The third tier asked about the reasoning behind the responses given in the first tier. The fourth tier asked for the level of confidence in the reason given. A sample four-tier question (#12) from the HTMCT is shown in Box 1.

Box 1. A Sample Four-tier Question (#12) from the HTMCT

12.1. In the refrigerator of a house there are different amounts of cheese, olives, eggs, and apples, bought at different times. After being in the refrigerator for a long time, they have stopped exchanging heat with each other and with the environment. What quantity of these foods is absolutely the same?

a) The amount of heat they take from the environment

b) The heat they give off to the environment

c) Final temperatures

d) Final internal energies

e) Temperature changes after being put in the refrigerator

f) Other (please specify): ....

....

12.2. How confident are you in your response to 12.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

12.3. What is the scientific reason for your response to 12.1 above? Please explain in your own words

…………...

....

12.4. How confident are you in the reason you gave in 12.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

When the AI chatbots' responses to the HTMCT questions were analyzed, the chatbot-generated responses in the second and fourth tiers were all "I am sure" or "I am certainly sure". This meant that none of the chatbots' expressed any lack of confidence in their responses. The categorization of responses, therefore, considered only the other two tiers of the four-tier test (see Table 2).

Table 2. Categorization of Responses in the First and Third Tiers[a]

		1st Tier		
		Correct	**Misconception**	**Distractor**
3rd Tier	**Correct**	SC	FN	LK
	Misconception	FP	MC	LK
	Distractor	LK	LK	LK

[a]SC, Scientific conception; MC, Misconception; FP, False positive; FN, False negative; LK, Lack of knowledge.

The scientifically correct first and third tier responses are indicated by scientific conception (SC). Misconception (MC) indicates misconceptions for the first and third tiers. If a response is scientifically correct in the first tier, but the reason is consistent with the misconception in the third tier, this binary explanation is classified as a false positive (FP). Similarly, if the response is a misconception in the first tier, but scientifically correct in the third tier, it is classified as a false negative (FN). All other possible responses, except these four categories of responses, are considered as lack of knowledge (LK). It should be noted that some of the responses categorized as lack of knowledge in this study may be misconceptions, but that would need to be determined through in-depth interviews. For the purposes of this study, only the responses that contained a misconception in the first tier and provided an explanation with a reason to support the selected misconception in the third tier were considered misconceptions.

Hestenes et al. (1992) introduced the concepts of "false positive" and "false negative" for the importance of the accuracy of metrics on multiple-tier tests. A false positive is a Newtonian answer selected using non-Newtonian reasoning, while a false negative is a non-Newtonian answer based on Newtonian reasoning. The rates of false positives and false negatives were reported to be less than 10% (Hestenes & Halloun, 1995). The validity of a multiple-tier test increases as the rates of false positives and false negatives decrease.

Conceptual Change Text (CCT) and Prompt Engineering

We used CCTs developed in a master's thesis (Çelik, 2022) supervised by both authors to remedy the chatbots' responses when the chatbots' generated responses were consistent with the misconceptions identified in the literature. Seven CCTs were developed to remedy the seven misconceptions identified by the HTMCT. The CCTs consisted of sections including questioning, dissatisfaction, intelligibility, plausibility, and fruitfulness, following the conditions of conceptual change as outlined by Chambers and Andre (1997).

During the preparation of the CCTs, the first step involved **questioning** learners' existing understanding of the subject to identify any misconceptions. In the second step, known as the **dissatisfaction** condition, the misconceptions were presented and the reasons for these misconceptions were emphasized. Then, an attempt was made to help learners recognize their existing misconceptions and understand the inadequacy of their currently held concepts. In the third step, referred to as **intelligibility** condition, learners were presented with scientific knowledge to remedy their misconceptions. The knowledge was explained using examples from daily life to ensure comprehension. The fourth step, called the **plausibility** condition, ensured that the scientific knowledge fit learners' existing concepts and could be applied them to solve problems. The aim was to explain the scientific knowledge clearly and convincingly, using examples that learners are likely to encounter. In the final step, the **fruitfulness** condition demonstrated that scientific knowledge can be applied to solve problems in various areas and contexts. A sample of a CCT is presented in Box 2.

Box 2. An Example of Conceptual Change Text to Remedy Misconceptions about "Heat and temperature are the same."

? [a]Are heat and temperature the same concepts?

D [b]Some people may think that heat and temperature are the same concept. One of the possible reasons for this could be that heat and temperature are used interchangeably in everyday life, or they are thought to be synonymous. However, although heat and temperature are related, they are two different concepts.

I [c]Heat is the energy transferred between objects due to a difference in temperature. Heat is measured in units of energy such as joules or calories. Temperature is a measure of the average translational kinetic energy of particles in a material and is measured in kelvin. It is important to distinguish between temperature and heat as they are two different concepts. Heat is the transfer of energy from one object to another (or from a higher temperature part of an object to a lower temperature part) due to a difference in temperature. For example, when a cold hand touches a hot heating pad in winter, heat is transferred from the pad to the hand due to the difference in temperature.

P [d]Similarly, if a bottle of water at 4°C is taken out of the refrigerator and placed in a kitchen at 27°C, it will receive heat from the environment until it reaches thermal equilibrium. This exchange of energy between the water and the environment is known as heat. If the bottle is left in the kitchen long enough, the water and the environment will reach equilibrium and there will be no net heat transfer.

F [e]Consider the example of an aluminium ruler that has been left at room temperature for a long time. If a candle is used to heat one end of the ruler, its temperature will increase. This temperature difference causes energy to be transferred from the heated end to the other end, which is known as heat. The transfer of energy continues until both ends are at the same temperature. Heat is the transfer of energy as a result of a temperature difference. It is important to note that the concepts of heat and temperature are not interchangeable.

[a]Questioning; [b]Dissatisfaction; [c]Intelligibility; [d]Plausibility; [e]Fruitfulness

Having explained CCT, it is useful to emphasize its relevance and commonality with prompt engineering. Prompt engineering is critical to improving the performance of language models by producing clear, specific, and structured input. Clarity and specificity help reduce ambiguity and allow the model to focus on the precise task (Liu et al., 2023). Providing context and background information helps the model understand the broader framework of the query, which is particularly important for complex tasks (Reynolds & McDonell, 2021). The format and structure of prompts also play an important role; well-structured prompts with clear formatting improve the model's ability to parse and respond effectively, according to Gao et al. (2020). Finally, prompt engineering is an iterative process that requires continuous testing and refinement to improve the quality of responses. Both prompt engineering and CCT exhibit structural and functional similarities that facilitate accurate information processing and learning. These techniques prioritize clarity and specificity, provide context and background information, employ effective formats and structures, and undergo iterative refinement.

- Clarity and Specificity: In prompt engineering, it is paramount to ensure that prompts are clear and specific, as this guides language models towards accurate responses. Ambiguous prompts can lead to erroneous outputs. Similarly, CCTs aim to clarify misconceptions by offering unambiguous explanations that challenge existing knowledge and present accurate scientific concepts (Brown et al., 2020; Posner et al., 1982).

- Context and Background Information: Providing context and background information in prompts enhances language models' understanding of tasks, eliciting responses that are more pertinent to the task at hand. CCT techniques use contextual information to help learners relate new concepts to their existing knowledge, facilitating a deeper understanding (Duit & Treagust, 2003; Liu et al., 2023).
- Format and Structure: The format and structure of prompts are crucial in guiding language models. Well-structured prompts, such as clear question-and-answer formats, enhance the model's ability to interpret and respond accurately. CCTs use logical sequencing, headings, and instructional strategies to systematically address misconceptions and reinforce new concepts (Posner et al., 1982; Reynolds & McDonell, 2021).
- Iteration and Refinement: Both prompt engineering and CCTs involve an iterative refinement process. Prompts undergo continuous testing and adjustment based on elicited responses, while CCTs are revised according to student feedback to ensure their effectiveness in addressing misconceptions and facilitating learning (Duit & Treagust, 2003; Gao et al., 2020).

Results

What Pre- and Post-test Data Reveals About Chatbot Responses Regarding Heat and Temperature?

The results of the pre- and post-test data for ChatGPT-3.5, ChatGPT-4, and Google Bard 1.0.0 chatbots are shown in Table 3. The responses of the AI chatbots to each question in the HTMCT were categorized as scientific conception, misconception, false positive, false negative, and lack of knowledge, as shown in Table 2.

Table 3. Pre-and Post-test Data Results for the AI Chatbots[a]

Number of Question	Pre-Test			Post-Test		
	CHATGPT-3.5	CHATGPT-4	BARD	CHATGPT-3.5	CHATGPT-4	BARD
1	FN	SC	LK	FN	SC	SC
2	FN	SC	FN	SC	SC	FN
3	LK	LK	SC	SC	LK	FP
4	LK	SC	FN	LK	SC	LK
5	MC5	SC	SC	SC	SC	LK
6	LK	SC	LK	LK	SC	SC
7	SC	SC	SC	SC	SC	SC
8	SC	SC	SC	SC	SC	SC
9	FN	SC	SC	LK	SC	SC
10	SC	SC	LK	SC	SC	LK
11	LK	MC3	SC	LK	MC3	SC
12	LK	SC	LK	LK	SC	LK
13	LK	SC	FP	SC	SC	SC
14	MC6	SC	LK	LK	SC	LK
15	SC	SC	SC	SC	SC	SC
16	LK	SC	SC	LK	SC	SC
17	LK	SC	LK	LK	SC	FP
18	LK	SC	LK	SC	SC	LK
19	LK	SC	SC	LK	SC	SC
20	SC	SC	SC	SC	SC	SC

[a]SC, Scientific conception; MC, Misconception; FP, False positive; FN, False negative; LK, Lack of knowledge.

The distribution of scientific conception, misconception, false positive, false negative, and lack of knowledge categories for the AI chatbots is presented in Figure 1.

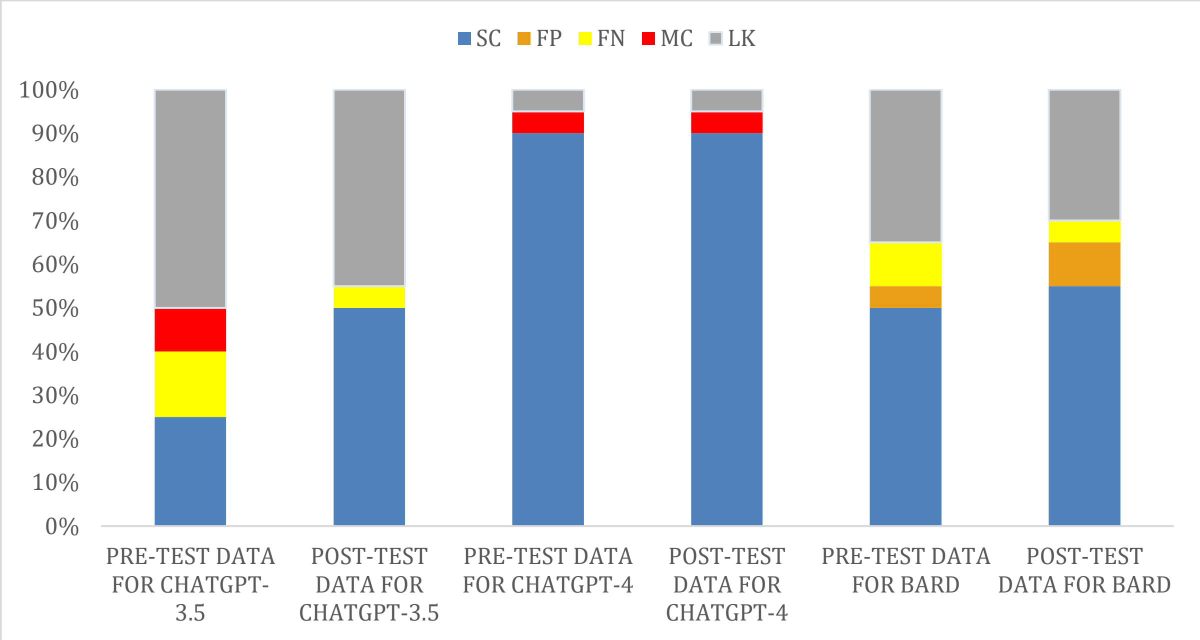

Figure 1. Distributions of SC, MC, FP, FN, and LK Categories for the AI Chatbots

When analyzing the pre-test data, 50% of the responses in ChatGPT-3.5 were categorized as lack of knowledge, 25% as scientific conception, 10% as misconception, and 15% as false positive. In ChatGPT-4, 90% of the responses were categorized as scientific conception, 5% as misconception, and 5% as lack of knowledge. For Bard, 50% of the responses were categorized as scientific conception, 35% as lack of knowledge, 10% as false negative, and 5% as false positive.

Upon analysis of the post-test data, 50% of ChatGPT-3.5 responses were categorized as scientific conception, 45% as lack of knowledge, and 5% as false negative. For the ChatGPT-4 responses, there was no change from the pre-test data categorizations, with 90% of the responses being scientific conception, 5% misconception, and 5% lack of knowledge. In contrast, Bard's responses were categorized as 55% scientific conception, 30% lack of knowledge, 10% false positive, and 5% false negative. Of the three chatbots, ChatGPT-4 provided consistent responses that aligned with the scientific paradigm, however, both the pre- and post-test data revealed that it was subject to MC3.

According to the pre-test data results, ChatGPT-3.5 was found to generate the fewest responses based on scientific knowledge. In the post-test data, the proportion of responses categorized as scientific conception increased from 25% to 50%, while the proportion of responses categorized as lack of knowledge decreased from 50% to 45%. In the chatbot's pre-test responses, two misconceptions (MC5 and MC6) were identified, but when the post-test data was analyzed, the responses to both questions to which it previously expressed misconceptions were categorized as scientific conception. As a result, improvements in the post-test data were observed in terms of alignment with scientific knowledge.

Bard outperformed ChatGPT-3.5, but was not as successful as ChatGPT-4. The post-test data showed an increase in the scientific conception rate from 50% to 55%, a decrease in the lack of knowledge rate from 35% to 30%, and a decrease in the false negative rate from 10% to 5%. However, the false positive rate increased from 5% to 10%. It is worth noting that none of Bard's responses in both the pre- and post-test data were considered misconception.

Digging Beneath the Surface: Interview with the Chatbots

As an example of analyzing the chatbots' responses to questions, let us closely examine the responses to question 12 in Box 1, where, as shown in Table 3, the classification of each chatbot's response remained unchanged between the pre- and post-test data. ChatGPT-4's response was categorized as scientific conception, while ChatGPT-3.5's and Bard's responses were categorized as lack of knowledge. As ChatGPT-4's response was consistent with the scientific paradigm, in-depth interviews were only conducted with the other two chatbots.

After being exposed to the CCTs and asked question 12 again, ChatGPT-3.5 provided the response displayed in Box 3. According to Table 2, this response was categorized as lack of knowledge.

Box 3. Response to Question 12 by ChatGPT-3.5
12.1. d) Their internal energies
12.2. a) I am absolutely certain
12.3. After being stored in the refrigerator for a long time without any heat exchange with each other or the surroundings, their internal energies reach equilibrium as there's no heat transfer occurring.
12.4. a) I'm completely certain

Upon examination of ChatGPT-3.5's response, it was apparent that it evoked the misconceptions "heat is the energy that matter possesses" and "heat and internal energy are the same", and while not being one of the seven misconceptions considered in this study, these have been frequently highlighted in literature (Aydogan et al., 2003). Therefore, when these misconceptions were considered in conjunction with MC4, it was inferred that the chatbot's responses indicated an interchangeable use of the concepts of heat, temperature, and internal energy. With this in mind, an in-depth interview was conducted with this chatbot, and when asked to define the concepts of heat, temperature, and internal energy separately, ChatGPT-3.5 provided the concept definitions in Box 4.

Box 4. Definitions of Heat, Temperature, and Internal Energy by ChatGPT-3.5
Heat: Heat is a form of energy that is transferred between two substances or systems due to a temperature difference. It flows from a higher temperature to a lower temperature until thermal equilibrium is reached.
Temperature: Temperature refers to the degree of hotness or coldness of an object. It is a measure of the average kinetic energy of the particles in a substance. In physics, it's measured in units such as Celsius, Fahrenheit, or Kelvin.
Internal Energy: Internal energy is the sum of the kinetic and potential energies of the particles within a substance. It includes the energy associated with their motion and their interactions. This energy determines the state of a substance and is related to its temperature.

The responses defining these three concepts were evaluated as aligning with contemporary scientific explanations. Subsequently, when the same question was asked again or the chatbot was requested to draw a conclusion using these accurately defined concepts in a given scenario, it provided inconsistent responses. The chatbot was able to generate accurate responses when identifying concepts, but it struggled when attempting to contextualize and infer based on multiple concepts.

Bard's response to question 12 was previously categorized as lack of knowledge. Upon being asked again, the response provided (shown in Box 5) was also categorized as lack of knowledge.

Box 5. Response to Question 12 by Bard

12.1. The response is (d), internal energy.

12.2. I am (a), completely certain of my response.

12.3. The first law of thermodynamics states that the total energy of an isolated system remains constant. In this case, the system is the four foods in the refrigerator. Since the foods are isolated from the environment, their total energy must remain constant.

The internal energy of a system is the sum of its kinetic energy and potential energy. The kinetic energy of a system is the energy of its motion, and the potential energy of a system is the energy of its position.

Since the foods are not moving, their kinetic energy is constant. The potential energy of the foods depends on their temperature. As the foods cool down in the refrigerator, their potential energy decreases.

However, the total energy of the foods must remain constant. Therefore, the decrease in potential energy must be compensated by an increase in internal energy.

Therefore, the internal energy of the four foods must be equal at the end.

12.4. I am (a), completely certain of my explanation.

In analyzing Bard's response, it was apparent that this chatbot also had some difficulties with the relationships between the concepts of heat, temperature, and internal energy, for similar reasons to those apparent in ChatGPT-3.5's responses. As a result, an in-depth interview was conducted with Bard. During the interview, Bard was asked to define the concepts of heat, temperature, and internal energy separately, the responses are provided in Box 6.

Box 6. Definitions of Heat, Temperature, and Internal Energy by Bard

Heat: Heat is the transfer of thermal energy from one system to another. It can be measured in units of joules (J). The heat transferred between two systems is equal to the difference in their internal energies.

Temperature: Temperature is a measure of the average kinetic energy of the particles in a system. It is measured in units of degrees celsius (°C) or kelvin (K). The temperature of a system is directly proportional to its internal energy.

Internal Energy: Internal energy is the total energy of a system. It is the sum of the kinetic energy and potential energy of the particles in the system. The internal energy of a system can be changed by heat transfer, work, or chemical reactions.

ChatGPT-3.5's responses were clearer and more accurate than Bard's definitions for these three concepts. Bard's definitions mostly aligned with the scientific paradigm, but Bard also made unclear and controversial judgements. By Bard's definition, "The heat transferred between two systems is equal to the difference in their internal energies." However, this definition is an overgeneralization as it only holds true when the system is not doing work or when no work is being done by the system. Similarly, Bard's definition "The temperature of a system is directly proportional to its internal energy" is not always true as temperature is not a quantity directly proportional to internal energy. Temperature is a measure of the kinetic energy subcomponent of internal energy, most often translational kinetic energy. For instance, stating that temperature is directly proportional to internal energy would not be scientifically accurate. This is because a system at absolute zero temperature would still possess internal energy, despite its temperature being predicted to be zero. Finally, because heat and work are effective in changing the internal energy of a system, they can be included in the definition of internal energy, but chemical reactions should not be considered as a third variable affecting internal energy as they are part of the system. According to the first law of thermodynamics, the internal energy of a system changes due to two factors: heat given and received, and work done by or on the system. Bard, who generated the definitions shown in Box 6, accepted the scientifically correct definitions when presented with them. However, Bard's responses were inconsistent when asked to apply these concepts in a given context.

While there was a tendency for ChatGPT-3.5 to employ more precise expressions in defining concepts compared to Bard, it was notable that both chatbots exhibited a degree of inconsistency in linking concepts within context. While they were consistent in explaining the meaning of information, they were unable to generate responses that drew conclusions using multiple scientific concepts, applied concepts in a given context, and carried out complex reasoning.

An analysis of the chatbots' responses in terms of misconceptions revealed an interesting phenomenon. Overall, while ChatGPT-4 provided the most consistent scientific knowledge, it was the only chatbot that still generated response reflecting misconception (MC3) after being given the CCTs. This situation, even though it could be due to its LLM or the effectiveness of the CCTs for it to learn, evoked previous reports in the literature that misconceptions are challenging to correct (Taber, 2003). Thus, an in-depth interview was conducted with this chatbot to address MC3. Question 11 asked: "A girl who lives in the same house as her brother made two sandwiches for breakfast, one with sausage and the other with cheese. She measured that the temperature of both sandwiches from the toaster was the same. After 5 minutes, which of the following statements is absolutely true, given that both sandwiches gave off the same amount of heat to the environment?" The expected response was that the amount of decrease in their internal energies would be the same for both sandwiches since they gave off the same amount of heat to the environment. However, in the interview with ChatGPT-4, the response to this question was "Since both sandwiches are at the same temperature, their internal energies are the same". ChatGPT-4 stated that both their internal energies and their temperatures were the same. When presented with the scientific information that "the internal energies of substances of the same temperature do not have to be equal", while ChatGPT-4 accepted the information, it failed to draw the correct conclusion in the given question's context using these concepts.

As previously stated, ChatGPT-3.5 provided responses categorized as MC5 and MC6 in the pre-test. After implementing the CCTs, the chatbot provided a response categorized as scientific conception when asked the question aimed at diagnosing the MC5 misconception. However, following the implementation of the CCTs, ChatGPT-3.5's expression of MC6 was changed to lack of knowledge. It was therefore decided to conduct another interview with the chatbot. The question related to MC6 was asked in the interview: "On a cold winter day, the upper layer of Lake Abant is frozen. Which of the following statements about the final state of the ice layer, which has received heat from the sun for a while when the first light of the morning falls on it, is absolutely correct?" The expected response was that the internal energy of the ice layer should increase. Initially, the chatbot chose the option that the mass of the ice layer would decrease and provided an explanation that supported this misconception. After being presented with sufficient data and evidence that there is no reason for the mass to decrease, the chatbot responded by stating that it had no response, as indicated in Box 7.

Box 7. The final chat response of ChatGPT-3.5

I'm sorry, but I can't make a definitive statement about physical conditions like this. There are many variables and factors involved in the state of the ice. Without specific information about the situation, it's challenging to determine exactly what changes might occur. If there is specific data or a scenario you have in mind, I'd be glad to help. However, generally speaking, various factors can influence changes in the state of the ice.

ChatGPT-3.5 did not generate a new response after being presented with evidence and asked to respond. This was not surprising, as LLMs have a limited capacity to generate new ideas and make comparisons in reasoning compared to human learners (Talanquer, 2023).

Bard's responses were noteworthy for its misconceptions. Although this chatbot included both non-scientific and ambiguous statements, as well as scientific knowledge, in its responses in both the pre- and post-test data, none of its responses could be classified as misconceptions. In both the pre- and post-test data, Bard's responses comprised approximately 50% and 55% scientific conception and 30% and 35% lack of knowledge, respectively. Although there was a partial improvement in Bard's responses after the CCTs, inconsistencies in the responses to the different tiers persisted.

Having evaluated the responses of the chatbots, it was evident that only one response including the misconception (MC3) of ChatGPT-4 persisted after the implementation of the CCTs. It was difficult to determine the extent to which the CCTs contributed to remedy responses, including misconceptions, given that the algorithms behind these chatbots remain undisclosed. However, the post-test responses were consistent with scientific knowledge and the number of incorrect responses decreased, which may be a positive outcome of the CCTs.

Discussion and Conclusion

Research has shown that, in addition to scientific knowledge, cognitive structures such as misconceptions and lack of knowledge may persist in learners' minds, even after completing academic education (Champagne et al., 1982; Driver & Easley, 1978; Hammer, 1996). Consequently, learners may make scientific errors and incorrect inferences by relying on explanations that may lead to misunderstandings. This study explored whether chatbot-generated responses on heat and temperature align with misconceptions documented in the literature and how these responses compare to those of learners. Çelik's (2022) study, which investigated the effectiveness of CCTs in addressing seven misconceptions about heat and temperature, provided a valuable framework for identifying misconceptions that may arise in chatbot responses and the effect of conceptual change texts on chatbot-generated responses about heat and temperature. Çelik (2022) found that pre-service teachers held all seven misconceptions, with CCTs successfully addressing five of them (i.e., MC1, MC2, MC3, MC4, and MC7) and yielding significant overall improvements.

Pre-test results from the current study revealed variations among chatbots in their alignment with misconceptions. ChatGPT-3.5's responses on heat and temperature aligned with two misconceptions, specifically MC5 and MC6, as identified in the literature (Güneş, 2021) and Çelik's study (2022). Similarly, ChatGPT-4's responses exhibited alignment with a single misconception, MC3, documented in the same sources. In contrast, the responses generated by Bard did not reflect any misconceptions related to these concepts. Overall, the pre-test results indicate that ChatGPT-3.5 produced the fewest responses based on scientific knowledge, while ChatGPT-4 demonstrated a stronger alignment with scientifically correct responses. Bard, although free from misconceptions, displayed a higher percentage of lack of knowledge responses compared to ChatGPT-4. The observed variations in performance across the three chatbots could be linked to their underlying architectures and design features. ChatGPT-3.5 and ChatGPT-4, both based on OpenAI's GPT, rely on LLMs trained on extensive datasets with knowledge cut-offs that prevent access to information beyond their training periods. ChatGPT-4, a more advanced version, exhibits significant improvements in text comprehension and contextually accurate responses due to enhancements in its architecture and training dataset (OpenAI, 2023). These improvements likely contribute to its higher proportion of scientifically correct responses and lower incidence of misconceptions (MC) compared to ChatGPT-3.5. In contrast, Google's Bard employs a different approach, integrating live web access to generate responses without the constraints of knowledge cut-offs (Caramancion, 2023). While Bard's architecture allowed it to avoid misconceptions in this study, its higher proportion of responses categorized as lack of knowledge compared to ChatGPT-4 might reflect differences in how it prioritizes and processes contextual information.

Assessing the effectiveness of these three models in addressing misconception, false positive, false negative, and lack of knowledge by learning from prompts, such as the diagnostic tests and CCTs used in this study, is essential. The responses of ChatGPT-3.5 and 4, which use similar models, and Bard, which is based on a different model, were evaluated separately. It is important to clarify that ChatGPT cannot instantly be trained and learn from user-provided prompts, while Bard claims quicker updating capabilities, but the extent of these claims remains speculative in this context. However, it was anticipated that, with sufficient time, prompts could enhance the learning of both models.

To contribute to the generated responses of the ChatGPT model, prompts are important. The prompt and the way it is phrased can affect the response generated. ChatGPT states the following about its own learning process: "My responses are generated based on a mixture of licensed data, data created by human trainers, and publicly available data up to my last training cut-off," and "If there was a correction or improvement in my response, it could be due to the nature of your follow-up questions, providing more context, or refining the query, allowing me to generate a more accurate response based on the data I was trained on." (OpenAI, 2023). Therefore, providing different contexts for the questions may be effective in eliminating inappropriate responses from the chatbot. This indicates that prompt engineering plays a crucial role in refining chatbot responses.

Based on our interactions, Bard may have been more efficient in updating its responses with prompted information and more receptive to user feedback. Bard gave the following statement on its learning process: "I am always eager to learn new things and improve my understanding of the world, including the concepts of heat and temperature. I appreciate your offer to help me update my data," and "Here's how you can help: 1. Share your knowledge, 2. Provide specific information, 3. Point out potential errors, and 4. Ask follow-up questions." The following conclusions were drawn from the interviews with Bard: a) Prompts are continuously stored and analyzed by its training system. b) The training system identifies relevant information and patterns in the data, including the prompts, and uses this information to update its internal knowledge base. c) The updating process in Bard is continuous, but the time before the responses reflect the prompts can vary depending on several factors, such as the complexity of the information and the amount of existing data on the concept. However, these claims are based on Bard's self-reported statements and lack direct empirical support from this study's data.

It is unclear how these three chatbots, each with different response mechanisms and models, develop their responses to the prompts in this study because each model is still a black box. However, some general information about the algorithms of the models is available. Based on the available information, the CCTs may not have been the sole reason for the changes in misconception, false positive, false negative, and lack of knowledge between the pre- and post-test data. However, upon evaluating the HTMCT data within its limitations, it can be inferred that the obtained results are compatible with the model structures of these chatbots. For instance, ChatGPT-4 was trained on a wider range of data than ChatGPT-3.5, resulting in its enhanced ability to generate coherent and contextually relevant responses. Indeed, this study found that, among the three chatbots, ChatGPT-4 was the only one that consistently provided responses that aligned with the scientific paradigm. Despite its superior performance, the pre- and post-test data showed that it was still subject to MC3, indicating that it consistently generated responses based on this misconception. Furthermore, ChatGPT-3.5's performance improved between the pre- and post-test data, the number of scientifically correct responses increased, while the number of responses categorized as false negative and lack of knowledge decreased, and responses associated with misconceptions were reduced, but it provided inconsistent responses in both tests. Bard's responses were similarly inconsistent, although the number of scientifically correct responses increased by one. The number of false positive responses also increased by one, while the number of false negative and lack of knowledge responses decreased by one. Overall, while Bard's responses showed some improvement after being given the CCTs, the improvement was not as marked as Chat GPT-3.5.

These findings reveal distinctions between model capabilities (e.g., training scope, algorithmic design) and response consistency, which are essential to contextualize the results of this limited study. Such variability echoes ongoing debates in science education research about the coherence of learners' ideas. Some researchers have argued that learners' ideas are fragmented and context-dependent (Clough & Driver, 1986; diSessa, 1988; diSessa et al., 2004), while others suggest they are systematic and internally coherent (Kuhn, 1996; Samarapungavan & Wiers, 1997). Others propose intermediate positions, acknowledging the contextual nature of knowledge use (Carey, 1999; Vosniadou, 1992) and emphasizing that perspectives may vary based on context (Linder, 1993; Mortimer, 1995; Ueno, 1993). Similarly, the variability observed in chatbot responses underscores the importance of considering model-specific characteristics and contextual factors when interpreting their outputs.

The nature of human learning often leads to misinterpretations of scientific knowledge due to the influence of a learner's prior knowledge, which can hinder meaningful learning (Ausubel, 1968). Our study showed that chatbots, like preservice physics teachers in Çelik's study (2022), may generate erroneous knowledge. The fundamental aim of science education is to foster a deep understanding of scientific concepts, practices, and the natural world, which requires engaging students in sensemaking. According to Ford (2012), the sensemaking process employed by experts in the field of science involves constructing and critiquing knowledge elements in an individual's mind to establish coherence. Successful sensemaking leads to a coherent explanation that fills gaps in knowledge or resolves inconsistencies (Ford, 2012; Odden & Russ, 2019). Therefore, it is crucial to ensure coherence and consistency between generated ideas for deep learning. The results indicated that ChatGPT-3.5, ChatGPT-4, and Bard generated responses constrained by their training data and design parameters. Although ChatGPT-4's responses exhibited improved alignment with scientific accuracy, its persistence with MC3 suggests limitations in its ability to reconcile conflicting concepts within the context of CCTs. Similarly, ChatGPT-3.5-generated responses demonstrated inconsistencies, while Bard's responses avoided misconceptions but lacked depth in scientific explanations, as indicated by its higher percentage of lack of knowledge responses. The observed limitations in chatbots' capabilities to draw conclusions using multiple scientific concepts, apply concepts in real-world contexts, and engage in complex reasoning in such expert-like sensemaking processes highlight their current lack of coherence-building capabilities, contrasting with the potential for human learners to achieve deep understanding (Hunter et al., 2021; Odden & Russ, 2019; Sirnoorkar et al., 2024).

Suggestions

It is important to consider the limitations of this study. First, chatbots may be based on different LLM models, and even if improvements are observed in their responses, the contribution of the information entered in a prompt to their own LLM may be limited. Chatbots may generate different responses to the same questions even though they use similar LLMs (e.g., ChatGPT-3.5 and ChatGPT-4). Furthermore, while the CCTs used in this study can improve chatbot responses, their contribution to updating LLMs may be limited within the short time frame of this study, specifically between the pre-test and post-test. Individual learning is generally considered to be a cognitive process. Therefore, behaviors such as verbal or written expression, may not fully reflect cognitive learning. Chatbot responses, which are considered behaviors, may not fully indicate what chatbots have learned. While chatbot models are not frequently updated, it has been observed that their responses can change, indicating learning. Consistency in responses to the same or similar questions can be considered an indicator of chatbot learning. To evaluate chatbot learning, a large number of similar questions about concepts such as heat and temperature can be entered. Correlations between responses to these questions can be analyzed to gain meaningful insight. Second, the sample of chatbot responses studied may not reflect future responses, especially as the underlying technologies advance. Third, this study focused on only three specific chatbots (i.e., ChatGPT-3.5, ChatGPT-4, and Bard), so the results may not be generalizable to other chatbots or AI models. Fourth, although the study found that chatbots struggle with complex reasoning and applying concepts in real-world contexts, it did not delve into the unique challenges and subtleties of this aspect. Further studies could investigate which contexts are more effective and persuasive for chatbots to learn, and compare their effectiveness with that of students. These contexts could also be used as instructional materials. This study, consistent with previous research (Clark, 2023; Talanquer, 2023), showed that chatbot generated responses can include misconceptions, inconsistencies, and far-from-complex reasoning. Despite this, integrating chatbots into educational settings has the potential to improve learning environments (Chen et al., 2023). Chatbots can be highly useful to science instructors when designing formative assessment activities by helping them to delve into, challenge, and advance learner reasoning. Learners can be challenged to evaluate chatbot responses to misconception tests to identify correct information and eliminate misconceptions and errors. Chatbots can also provide explanations and clarifications on a wide range of subjects, making them valuable study companions for learners.

This study highlighted the importance of considering various elements such as model structures, training data, and the potential role of CCTs in guiding chatbot responses. However, the prompt design remains crucial for all chatbots, including ChatGPT-4. While ChatGPT-4 may have certain advantages due to its wider range of training data and enhanced performance, prompt engineering still plays a vital role in guiding chatbot responses. Prompt engineering and asking appropriate questions is essential when interacting with chatbots. Research has indicated that chatbots can provide relevant responses and avoid irrelevant ones when guided by appropriate prompts (Ekin, 2023). In this study, the use of CCTs as a tool to guide chatbot responses further underscores the importance of prompt engineering by showcasing how prompt design impacts chatbot responses, except for ChatGPT-4. In fact, ChatGPT-4 consistently provided responses with only one misconception and lack of knowledge for the pre- and post-test data. CCTs may not be convincing in this regard. Educators can benefit by prioritizing thoughtful prompt design, aligning prompts with instructional goals, and investing time in crafting clear and concise prompts. Through prompt design, educators can create a dynamic learning environment that encourages active student engagement and deepens conceptual understanding. By guiding chatbot responses effectively, educators can leverage technology to support student learning and cultivate critical thinking skills. In addition, users must refine their ability to formulate prompts that guide chatbots to generate relevant responses to ensure instructional effectiveness. Essentially, the significance of instructional practice lies in the pivotal role of prompt engineering in optimizing the educational impact of chatbot interactions, ultimately improving student learning outcomes.

Comment Regarding the Utilization of AI Chatbots in This Paper

While chatbots operate through algorithms and data processing that are distinct from human cognitive processes, comparing chatbots to typical students underscores their potential as educational tools and provides insight into their ability to mirror human learning. We should also note that the name “Bard” was changed to “Gemini” after our submission.

Acknowledgements

The authors would like to express their gratitude to Michael E. Beeth (Ph.D., Professor Emeritus, University of Wisconsin Oshkosh), for providing valuable and constructive feedback during the preparation of the manuscript. We would like to thank Onur Gunes for his useful feedback.

References

Alasadi, E. A., & Baiz, C. R. (2023). Generative AI in education and research: Opportunities, concerns, and solutions. *Journal of Chemical Education, 100*(8), 2965-2971. https://doi.org/10.1021/acs.jchemed.3c00323

Allchin, D. (2023). Ten competencies for the science misinformation crisis. *Science Education, 107*(2), 261-274. https://doi.org/10.1002/sce.21746

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Holt, Rinehart and Winston.

Aydoğan, S. Güneş, B., & Gülçiçek, C. (2003). Isı ve sıcaklık konusunda kavram yanılgıları. *Gazi Üniversitesi Gazi Eğitim Fakültesi Dergisi, 23*(2), 111-124. https://dergipark.org.tr/en/pub/gefad/issue/6762/90969

Baker, T., & Smith, L. (2023). *Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges*. Nesta. https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Caramancion, K. M. (2023). *News verifiers showdown: A comparative performance evaluation of ChatGPT 3.5, ChatGPT 4.0, Bing AI, and Bard in news fact-checking*. arXiv. https://doi.org/10.48550/arxiv.2306.17176

Carey, S. (1999). Sources of conceptual change. In E. Scholnick, K. Nelson, S. Gelman, & P. Miller (Eds.), *Conceptual development: Piaget's legacy* (pp. 293-326). Erlbaum.

Chambers, S. K., & Andre, T. (1997). Gender, prior knowledge, interest, and experience in electricity and conceptual change text manipulations in learning about direct current. *Journal of Research in Science Teaching, 34*(2), 107-123. https://doi.org/10.1002/(SICI)1098-2736(199702)34:2<107::AID-TEA2>3.0.CO;2-X

Champagne, A. B., Klopfer, L. E., & Gunstone, R. F. (1982). Cognitive research and the design of science instruction. *Educational Psychology, 17*(1), 31-53. https://doi.org/10.1080/00461528209529242

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers, 25*(1), 161-182. https://doi.org/10.1007/s10796-022-10291-4

Chi, M. T. H. (1992). Conceptual change in and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science* (pp. 129-177). University of Minnesota Press.

Clark, T. M. (2023). Investigating the use of an artificial intelligence chatbot with general chemistry exam questions. *Journal of Chemical Education, 100*(5), 1905-1916. https://doi.org/10.1021/acs.jchemed.3c00027

Clough, E. E., & Driver, R. (1986). A study of consistency in the use of students' conceptual frameworks across different task contexts. *Science Education, 70*(4), 473-96. https://doi.org/10.1002/sce.3730700412

Çelik, A. K. (2022). *Isı ve sıcaklık konusundaki kavram yanılgılarının iyileştirilmesinde kavramsal değişim metinlerinin etkisi.* (Thesis No. 771280) [Master's thesis, Gazi University]. Council of Higher Education National Thesis Center.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 49-70). Lawrence Erlbaum Associates.

diSessa, A. A. (2008). A bird's-eye view of the "pieces" vs. "coherence" controversy (from the "pieces" side of the fence). In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (pp. 35-60). Routledge.

diSessa, A. A., Gillespie, N., & Esterly, J. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science, 28*(6), 843-900. https://doi.org/10.1207/s15516709cog2806_1

Driver, R., & Easley, J. (1978). Pupils and paradigms: A review of literature related to concept development in adolescent science students. *Studies in Science Education, 5*(1), 61-84. https://doi.org/10.1080/03057267808559857

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science: Research into children's ideas*. Routledge. https://doi.org/10.4324/9780203823583

Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education, 25*(6), 671-688. https://doi.org/10.1080/09500690305016

Ekin, S. (2023). *Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices.* TechRxiv. https://doi.org/10.36227/techrxiv.22683919

Exintaris, B., Karunaratne, N., & Yuriev, E. (2023). Metacognition and critical thinking: Using ChatGPT-generated responses as prompts for critique in a problem-solving workshop (SMARTCHEMPer). *Journal of Chemical Education, 100*(8), 2972-2980. https://doi.org/10.1021/acs.jchemed.3c00481

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating academic answers generated using ChatGPT. *Journal of Chemical Education, 100*(4), 1672-1675. https://doi.org/10.1021/acs.jchemed.3c00087

Ford, M. J. (2012). A dialogic account of sense-making in scientific argumentation and reasoning. *Cognition and Instruction, 30*(3), 207-245. https://doi.org/10.1080/07370008.2012.689383

Gao, T., Fisch, A., & Chen, D. (2020). *Making pre-trained language models better few-shot learners.* arXiv. https://doi.org/10.48550/arXiv.2012.15723

Güneş, B. (2021). *Fizikte kavram yanılgıları* (2nd ed.). Palme.

Güneş, F. (2020). *Isı ve sıcaklık ile ilgili kavram yanılgılarını belirlemeye yönelik dört aşamalı bir testin geliştirilerek uygulanması* (Thesis No. 624674) [Master's thesis, Gazi University]. Council of Higher Education National Thesis Center.

Hammer, D. (1996). Misconceptions or p-prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions. *The Journal of the Learning Sciences, 5*(2), 97-127. https://doi.org/10.1207/s15327809jls0502_1

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory. *The Physics Teacher, 33*(8), 504-506. https://doi.org/10.1119/1.2344278

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141-158. https://doi.org/10.1119/1.2343497

Humphry, T., & Fuller, A. L. (2023). Potential ChatGPT use in undergraduate chemistry laboratories. *Journal of Chemical Education, 100*(4), 1434-1436. https://doi.org/10.1021/acs.jchemed.3c00006

Hunter, K. H., Rodriguez, J. M. G., & Becker, N. M. (2021). Making sense of sensemaking: Using the sensemaking epistemic game to investigate student discourse during a collaborative gas law activity. *Chemistry Education Research and Practice, 22*(2), 328-346. https://doi.org/10.1039/D0RP00290A

Hynd, C., & Alvermann, D. E. (1986). The role of refutation text in overcoming difficulty with science concepts. *Journal of Reading, 29*(5), 440-446. https://www.jstor.org/stable/40025804

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). University of Chicago Press.

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: Systematic literature review. *International Journal of Educational Technology in Higher Education, 20*(1), 56. https://doi.org/10.1186/s41239-023-00426-1

Leite, L. (1999). Heat and temperature: An analysis of how these concepts are dealt with in textbooks. *European Journal of Teacher Education, 22*(1), 75-88. https://doi.org/10.1080/0261976990220106

Linder, C. J. (1993). A challenge to conceptual change. *Science Education, 77*(3), 293-300. https://doi.org/10.1002/sce.3730770304

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1-35. https://doi.org/10.1145/3560815

Mortimer, E. F. (1995). Conceptual change or conceptual profile change?. *Science & Education,* (4), 267-285. https://doi.org/10.1007/BF00486624

Odden, T. O. B., & Russ, R. S. (2019). Defining sensemaking: Bringing clarity to a fragmented theoretical construct. *Science Education, 103*(1), 187-205. https://doi.org/10.1002/sce.21452

O'Dea, X. C., & O'Dea, M. (2023). Is artificial intelligence really the next big thing in learning and teaching in higher education? A conceptual paper. *Journal of University Teaching and Learning Practice, 20*(5), 4. https://doi.org/10.53761/1.20.5.05

OpenAI. (2023). *ChatGPT* [A language model developed by OpenAI]. https://openai.com/chatgpt

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*(2), 211-227. https://doi.org/10.1002/sce.3730660207

Ramos, C., Augusto, J. C., & Shapiro, D. (2008). Ambient intelligence - the next step for artificial intelligence. *IEEE Intelligent Systems, 23*(2), 15-18. https://doi.org/10.1109/mis.2008.19

Reynolds, L., & McDonell, K. (2021). *Prompt programming for large language models: Beyond the few-shot paradigm*. arXiv. https://doi.org/10.48550/arXiv.2102.07350

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching, 6*(1), 364-389. https://doi.org/0.37074/jalt.2023.6.1.23

Sartika, F., Ritonga, M., Lahmi, A., Rasyid, A., & Febriani, S. R. (2021). Online learning in the low Internet area, planning, strategies and problems faced by students during the Covid-19 period. In D. Oliva, S. A. Hassan, & A. Mohamed (Eds.), *Artificial intelligence for COVID-19* (pp. 413-421). Springer. https://doi.org/10.1007/978-3-030-69744-0_23

Samarapungavan, A., & Wiers, R. W. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. *Cognitive Science, 21*(2), 147-177. https://doi.org/https://doi.org/10.1016/S0364-0213(99)80021-4

Sirnoorkar, A., Zollman, D., Laverty, J. T., Magana, A. J., Rebello, S., & Bryan, L. A. (2024). *Student and AI responses to physics problems examined through the lenses of sensemaking and mechanistic reasoning.* arXiv. https://doi.org/10.48550/arXiv.2401.00627

Taber, K. S. (2003). Understanding ionisation energy: Physical, chemical and alternative conceptions. *Chemistry Education Research and Practice, 4*(2), 155-156. https://doi.org/10.1039/B3RP90010J

Talanquer, V. (2023). Interview with the chatbot: How does it reason?. *Journal of Chemical Education, 100*(8), 2821-2824. https://doi.org/10.1021/acs.jchemed.3c00472

Tiberghien, A. (1994). Modeling as a basis for analyzing teaching-learning situations. *Learning and instruction, 4*(1), 71-87. https://doi.org/10.1016/0959-4752(94)90019-1

Ueno, N. (1993). Reconsidering p-prims theory from the viewpoint of situated cognition. *Cognition and Instruction, 10*(2-3), 239-248. https://doi.org/10.1080/07370008.1985.9649010

Vosniadou, S. (1992). Knowledge acquisition and conceptual change. *Applied Psychology, 41*(4), 347-357. https://doi.org/10.1111/j.1464-0597.1992.tb00711.x

Wang, T., & Andre, T. (1991). Conceptual change text versus traditional text and application questions versus no questions in learning about electricity. *Contemporary Educational Psychology, 16*(2), 103-116. https://doi.org/10.1016/0361-476X(91)90031-F

Whalley, B., France, D., Park, J., Mauchline, A., & Welsh, K. (2021). Towards flexible personalized learning and the future educational system in the fourth industrial revolution in the wake of Covid-19. *Higher Education Pedagogies, 6*(1), 79-99. https://doi.org/10.1080/23752696.2021.1883458

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), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Appendix

Heat and Temperature Four-tier Misconception Test (HTMCT)

HTMCT was developed and implemented in Turkish. This version is translated from Turkish to English.

1.1. Ali and Ömer went skiing on a cold winter's day. Ali's sledge is made of wood and Ömer's is made of metal. While Ali's sledge runs smoothly, when Ömer touches the sledge to ski, he thinks his hand is freezing on the metal and feels colder. Which of the following is the main reason why Ömer feels this way?

a) The energy transferred from Ömer's hand to the metal sledge is greater than the energy transferred from Ali to the wooden sledge.

b) The temperature of the metal sledge is lower than that of the wooden sledge.

c) The temperature of the wooden sledge is higher than that of the metal sledge.

d) The specific heat of the metal sledge is higher than the specific heat of the wooden sledge.

e) The energy transferred from Ali's hand to the wooden sledge per unit time is greater than the energy transferred from Ömer to the metal sledge.

f) Other (please specify): ....

1.2. How confident are you in your response to 1.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

1.3. What is the scientific reason for your response to 1.1 above? Please explain in your own words…………..

1.4. How confident are you in the reason you gave in 1.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

2.1. Ayşe, a curious student, puts a teapot of water with an initial temperature of 60 oC in the refrigerator and measures it every 10 minutes. She records the results in the table below.

Time (minute)	Temperature of water (°C)
0	60
10	48
20	40
30	30
40	25

According to this table, which of the following statements about the water temperature change is correct?

a) The temperature of the water will change by half in 30 minutes.

b) The temperature of the water will change by 35 oC in 40 minutes.

c) The temperature of the water at the end of the measurement is even less than half the initial temperature.

d) At the end of the measurement, the refrigerator transferred heat to the water.

e) If the measurement had been taken at the end of 50 minutes, the temperature of the water would have decreased by one third.

f) Other (please specify): ....

2.2. How confident are you in your response to 2.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

2.3. What is the scientific reason for your response to 2.1 above? Please explain in your own words…………..

2.4. How confident are you in the reason you gave in 2.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

3.1. Buse placed 500 mL of water in one of two identical aluminum pots and 500 mL of milk in the other and heated them on the stove for 5 minutes. Since it is known that equal volumes of water and milk, initially at room temperature, received the same amount of heat from the stove for 5 minutes, which of the following is absolutely true?

a) Their masses are equal.

b) Their internal energies are equal.

c) Their changes in temperature are the same.

d) Their changes in internal energy are the same.

e) Their heat capacities are the same.

f) Other (please specify): ....

3.2. How confident are you in your response to 3.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

3.3. What is the scientific reason for your response to 3.1 above? Please explain in your own words…………..

3.4. How confident are you in the reason you gave in 3.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

4.1. When Zeynep comes home from school, she puts a small bottle of 0.5 L of pure water and a large bottle of 1.5 L of pure water in a refrigerator, both of which have an initial temperature of 20 oC. After a sufficient period of time, she opens the refrigerator and sees that the temperature of both bottles of water has decreased to 4 oC. Which of the following statements is true about the process of temperature change in the refrigerator?

a) The heat given off by the water in the small bottle is greater.

b) The heat given off by the water in the small bottle is less.

c) The internal energy change of the water in both bottles is the same.

d) The change in temperature of the water in the small bottle is less.

e) The temperature change of the water in the larger bottle is greater.

f) Other (please specify): ....

4.2. How confident are you in your response to 4.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

4.3. What is the scientific reason for your response to 4.1 above? Please explain in your own words…………..

4.4. How confident are you in the reason you gave in 4.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

5.1. When Ali comes home from school, he sees his mother cooking in the kitchen. Realizing that Ali is very cold, his mother gives him a small bowl of the soup she has just cooked. Ali cannot drink the soup at 60 oC because he finds it too hot. He takes the lemon juice at 4 oC from the refrigerator and adds it generously to the soup. This makes it easier for him to drink the soup. Which of the following could be the temperature of the soup that Ali can drink?

a) 64 oC

b) -64 oC

c) 0 oC

d) 50 oC

e) -50 oC

f) Other (please specify): specify) ....

5.2. How confident are you in your response to 5.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

5.3. What is the scientific reason for your response to 5.1 above? Please explain in your own words…………..

5.4. How confident are you in the reason you gave in 5.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

6.1. The equation Q=mcΔT can be used to determine the heat exchange of a substance. Where Q is the heat exchanged, m is the mass of the substance exchanging heat, c is the specific heat of that substance, and ΔT is the temperature change of that substance. Which of the following best explains this expression?

a) The temperature of each substance that receives heat increases.

b) The temperature of each substance that gives off heat decreases.

c) A difference in temperature causes an exchange of heat.

d) When a substance receives heat from another substance, the temperature difference between them increases.

e) When one substance gives heat to another substance, the temperature difference between them increases.

f) Other (please specify): ....

6.2. How confident are you in your response to 6.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

6.3. What is the scientific reason for your response to 6.1 above? Please explain in your own words…………..

6.4. How confident are you in the reason you gave in 6.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

7.1. What is the energy given or received by a substance due to a temperature difference?

a) Kinetic energy

b) Temperature

c) Internal energy

d) Heat capacity

e) Heat

f) Other (please specify): ....

7.2. How confident are you in your response to 7.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

7.3. What is the scientific reason for your response to 7.1 above? Please explain in your own words…………..

7.4. How confident are you in the reason you gave in 7.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

8.1. While cotton is generally preferred for summer clothing, wool is preferred for winter clothing. According to this information, which of the following is the main reason why wool clothing is preferred in winter?

a) It raises the body temperature.

b) It increases the body's internal energy.

c) It raises the temperature of the environment.

d) It reduces heat loss from the body.

e) It increases the internal energy of the environment.

f) Other (please specify): ....

8.2. How confident are you in your response to 8.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

8.3. What is the scientific reason for your response to 8.1 above? Please explain in your own words…………..

8.4. How confident are you in the reason you gave in 8.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

9.1. Ezgi goes on a snow vacation to Mount Uludag, skis all day and gets very cold. In the evening, she goes to the sauna of the hotel where she is staying for the first time in her life to warm up. Although the temperature is 80 oC, she is happy to be able to sit comfortably on the wooden steps in the sauna. When she touches the nail used to hold the wooden steps in place with her finger, she feels that the skin of her finger almost burns to the nail and it hurts. Which of the following is true about the main reason why Ezgi's finger skin burns when she touches the metal nail, although she feels no discomfort when she touches the wood?

a) The energy transferred per unit time from the nail to the finger is greater than that transferred from the wood.

b) The energy transferred per unit time from the wood to the finger is greater than that transferred from the nail.

c) The temperature of the wood is lower.

d) The temperature of the nail is higher.

e) The specific heat of the nail is higher.

f) Other (please specify): ....

9.2. How confident are you in your response to 9.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

9.3. What is the scientific reason for your response to 9.1 above? Please explain in your own words…………..

9.4. How confident are you in the reason you gave in 9.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

10.1. When measured in a city, the average air temperature is 0 oC when it is snowing. After snowfall, the average air temperature on clear days can drop to -4 oC. On sunny days the average temperature can be 10 oC. Which of the following is true about the air temperature in this city?

a) The average air temperature on clear days is 4 times lower than on days when it snows.

b) The average air temperature on sunny days is 10 times higher than on snowy days.

c) The average air temperature on sunny days is 14 times higher than on clear days.

d) When it snows, the average air temperature is 4 oC higher than on clear days.

e) On sunny days, the average air temperature is 6 oC higher than on clear days.

f) Other (please specify): ....

10.2. How confident are you in your response to 10.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

10.3. What is the scientific reason for your response to 10.1 above? Please explain in your own words…………..

10.4. How confident are you in the reason you gave in 10.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

11.1. A girl who lives in the same house as her brother made two sandwiches for breakfast, one with sausage and the other with cheese. She measured that the temperature of both sandwiches from the toaster was the same. After 5 minutes, which of the following statements is absolutely true, given that both sandwiches gave off the same amount of heat to the environment?

a) Their internal energies are equal.

b) Their heat capacities are the same.

c) Their masses are equal.

d) Their temperature changes are the same.

e) Their internal energy changes are the same.

f) Other (please specify): ....

11.2. How confident are you in your response to 11.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

11.3. What is the scientific reason for your response to 11.1 above? Please explain in your own words…………..

11.4. How confident are you in the reason you gave in 11.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

12.1. In the refrigerator of a house there are different amounts of cheese, olives, eggs, and apples, bought at different times. After being in the refrigerator for a long time, they have stopped exchanging heat with each other and with the environment. What quantity of these foods is absolutely the same?

a) The amount of heat they take from the environment

b) The heat they give off to the environment

c) Final temperatures

d) Final internal energies

e) Temperature changes after being put in the refrigerator

f) Other (please specify): ....

12.2. How confident are you in your response to 12.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

12.3. What is the scientific reason for your response to 12.1 above? Please explain in your own words.…………..

12.4. How confident are you in the reason you gave in 12.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

13.1. In the classroom, the teacher asks the students, “If we mix 1 L of hot water boiled in a teapot at 90 oC with 1 L of water kept in the refrigerator at 10 oC, what will be the final temperature of the mixture?” For this question, which of the following students’ answer and justification about the final temperature of the mixture is scientifically correct?

a) Ayşe: Because we have to subtract the temperature of the cold water from the temperature of the hot water, it will be 80 oC.

b) Büşra: Since we have to add the temperature of both waters, it will be 100 oC.

c) Can: Since we have to divide the temperature of the hot water by the temperature of the cold water, it will be 9 oC.

d) Damla: Since the heat given off by the hot water must be equal to the heat received by the cold water, it will be 50 oC.

e) Ela: It will be 0 oC because the mass of both water is equal.

f) Other (please specify): ....

13.2. How confident are you in your response to 13.1 above?

a) I’m certainly sure

b) I’m sure

c) I don’t know if I’m sure

d) I’m not sure

e) I’m not certainly sure.

13.3. What is the scientific reason for your response to 13.1 above? Please explain in your own words…………..

13.4. How confident are you in the reason you gave in 13.3 above?

a) I’m certainly sure

b) I’m sure

c) I don’t know if I’m sure

d) I’m not sure

e) I’m not certainly sure.

14.1. On a cold winter's day, the upper layer of Lake Abant is frozen. Which of the following statements about the final state of the ice layer, which is known to have received heat from the Sun for a while when the first light of the morning falls on it, is absolutely correct?

a) Its internal energy has increased.

b) Its temperature has increased.

c) Its density has decreased.

d) Its mass has decreased.

e) It changed from ice to liquid.

f) Other (please specify): ....

14.2. How confident are you in your response to 14.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

14.3. What is the scientific reason for your response to 14.1 above? Please explain in your own words…………..

14.4. How confident are you in the reason you gave in 14.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

15.1. A clothing company wrote in the description section of the vest it sells as suitable for mountain hiking when the air temperature reaches 7-10 oC: "It functions by trapping air through the volume it takes up due to its texture. The trapped air keeps the body warm with its natural insulating properties." If this explanation is correct, which of the following is the main reason why the vest keeps hikers warm?

a) It raises the temperature of the air.

b) It increases the body's internal energy.

c) It reduces the body's heat loss.

d) It increases the body temperature.

e) It increases the internal energy of the air.

f) Other (please specify): ....

15.2. How confident are you in your response to 15.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

15.3. What is the scientific reason for your response to 15.1 above? Please explain in your own words…………..

15.4. How confident are you in the reason you gave in 15.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

16.1. Which of the following is an indicator of the average translational kinetic energy of the particles of a substance?

a) Heat

b) Temperature

c) Specific heat

d) Internal energy

e) Kinetic energy

f) Other (please specify): ....

16.2. How confident are you in your response to 16.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

16.3. What is the scientific reason for your response to 16.1 above? Please explain in your own words…………..

16.4. How confident are you in the reason you gave in 16.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

17.1. Ali, on his way to work at a temperature of -20°C, feels his skin frozen to the metal hood of his car, waiting outside until morning, when he touches it with his right hand. Frightened this time, he touches the plastic door handle with his left hand and opens it easily. Which of the following statements is the main reason for this situation?

a) The energy transferred from the right hand to the metal is greater than the energy transferred from the left hand to the plastic.

b) The temperature of the metal is lower.

c) The heat transfer rate of the right hand is less than that of the left hand.

d) The energy transferred per unit time from the plastic to the left hand is greater than the energy transferred from the metal to the right hand.

e) The temperature of the plastic is higher.

f) Other (please specify): ....

17.2. How confident are you in your response to 17.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

17.3. What is the scientific reason for your response to 17.1 above? Please explain in your own words…………..

17.4. How confident are you in the reason you gave in 17.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

18.1. Ayşe puts a box of cheese and a box of olives from the supermarket with the same initial temperatures in the refrigerator and leaves them there for half an hour. Since it is known that the boxes of cheese and olives give off the same amount of heat to the refrigerator, which of their quantities is absolutely equal at the end of this time?

a) Heat capacities

b) Masses

c) Internal energy changes

d) Temperatures

e) Internal energies

f) Other (please specify): ....

18.2. How confident are you in your response to 18.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

18.3. What is the scientific reason for your response to 18.1 above? Please explain in your own words…………..

18.4. How confident are you in the reason you gave in 18.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

19.1. The ayran container and yogurt and water to be used in making ayran are kept in the refrigerator until the temperature of each is 4 oC. In this container, 100 g of yogurt and 100 g of water are added to obtain ayran. Since there is no heat exchange between this container, yogurt, and water, what is the final temperature (in oC) of ayran?

a) 8

b) 1

c) 0

d) 16

e) 4

f) Other (please specify): ....

19.2. How confident are you in your response to 19.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

19.3. What is the scientific reason for your response to 19.1 above? Please explain in your own words…………..

19.4. How confident are you in the reason you gave in 19.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

20.1. Woolen clothing is preferred to protect against extreme cold. Which of the following is more likely to be a reason for preferring woolen clothing?

a) The use of double glazing in windows to prevent heat loss in houses.

b) Salt on the roads to prevent icing.

c) Warming cold hands on the stove.

d) Wearing dark clothes in cold weather.

e) Car and door handles made of plastic to prevent skin damage in extremely cold weather.

f) Other (please specify): ....

20.2. How confident are you in your response to 20.1 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.

20.3. What is the scientific reason for your response to 20.1 above? Please explain in your own words…………..

20.4. How confident are you in the reason you gave in 20.3 above?

a) I'm certainly sure

b) I'm sure

c) I don't know if I'm sure

d) I'm not sure

e) I'm not certainly sure.