Liu, L. (2023). Analyzing the text contents produced by ChatGPT: Prompts, feature-components in responses, and a predictive model. *Journal of Educational Technology Development and Exchange*, *16*(1), 49-70

Analyzing the Text Contents Produced by ChatGPT: Prompts, Feature-Components in Responses, and a Predictive Model

Leping Liu
University of Nevada, Reno, USA
liu@unr.edu

Abstract: *ChatGPT is a large language model that uses deep learning to produce natural language and generate intelligent and relevant responses to user prompts. It comes to the field of education as an inevitable wave. Educators have to deal with it and figure out appropriate ways to use it and produce positive learning. This study explores the use of ChatGPT from the perspective of front-end users, focusing on the text-content that ChatGPT can produce for learners to learn new knowledge (e.g., a concept, a theory, or an application). The sample of this study consists of 253 ChatGPT text responses derived from three types of initial prompts/ questions: general questions, specific questions, and questions with interactive prompts. Six feature components of text-information that can help learners to understand new knowledge are analyzed (concept and definition, procedure, example, comparison or contrast, deductive or inductive argument, summary). The results from Chi-square tests indicate that the presence of each feature component in the responses differs by the types of prompts. The results from a logistic regression analysis reveal that the presence of five (out of the six) feature components are significant to the probability that a response provides accurate and reliable information. The integration of using ChatGPT into learning is discussed. Further research questions are suggested.*

Keywords: ChatGPT, general/specific/interactive prompt, feature component, two-way learning, predictive model

1. Introduction

Artificial intelligence (AI) refers to the science, engineering and technology that create systems to operate tasks associated with intelligent activities such as learning, decision making, and problem solving (Farrokhnia et al., 2023; Xu et al., 2021). AI has been used in the natural language processing, resulting in the development of intelligent chatbots that are able to virtually assist in understanding and producing human language (Caldarini et al., 2022). ChatGPT is such an AI-powered chatbot, a natural language processing model developed by OpenAI (OpenAI, 2022). Generative Pre-trained Transformer (GPT) is an autoregressive architecture that uses deep learning to produce human-like text (Roose, 2022). ChatGPT has been trained on substantial amounts of data to understand natural language and generate intelligent and relevant responses to users' queries (Halaweh, 2023). Since its launch in November 2022, millions of users have started to explore the opportunities of using it in various domains especially in education (Adiguzel et al., 2023; Atlas, 2023), such as higher education assessment (Talian & Kalinkara, 2023), teaching, learning, and student engagement (Dijkstra et al., 2022; Gabajiwala et al., 2022; Kasneci et al., 2023), or professional training and development (Halaweh, 2023).

It is believed that AI tools such as ChatGPT are developed to augment human intelligence (Carter & Nielsen, 2017; Cotton et al., 2023). When integrating ChatGPT into teaching and learning, the first step is to analyze or assess what it can do, and whether or to what extent it can be of help to achieve the learning goals. For example, one strength of ChatGPT is its capability of self-learning that enables a user-machine two-way learning process (Farrokhnia et al., 2023; Liu & Gibson, 2023). Knowing this, the user can have a careful design on the initial prompts to facilitate such two-way learning. Furthermore, two weaknesses of ChatGPT are noticed: (a) lack of deep understanding of the words it processes may result in ambiguous information in the response outputs (Farrokhnia, 2023; Gao et al., 2023; Gupta et al., 2023), and (b) lack of capability to determine the credibility of the data it was trained on may result in the uncertainties for the quality of responses (Farrokhnia et al., 2023; Lecler et al., 2023; Tlili et al., 2023). Therefore, a common issue could be: if a user wants to learn some new knowledge (e.g., concepts, a theory, a new method, or a technology application), to what extent can ChatGPT produce responses that are accurate and reliable?

Same as any new technology innovation, ChatGPT comes to the field of education as an inevitable wave. Like it or not, educators will need to deal with it and find appropriate ways of using it to produce positive learning outcomes. The first step to do so is to explore what text-information ChatGPT can produce for a learner. This study aims to explore: (a) whether necessary feature components of text-information are provided in ChatGPT responses that are derived from different types of prompts, and (b) to what extent the presence of the feature components can influence the quality of response information regarding its accuracy and reliability. Six feature components in response contents (concept and definition, procedure, example, comparison or contrast, deductive or inductive argument, and summary), and three types of initial prompts/questions (general questions, specific questions, and questions with interactive prompt) are analyzed in this study.

2. Background Review

For the purpose of this study, it is not likely to conduct a thorough and overall literature review on the effective use of ChatGPT to improve learning, for at least two reasons. Firstly, "there are still many knowledge gaps and uncertainties when it comes to the successful and responsible integration of large language models into learning and teaching process" (Kasneci et al., 2023, p.4). Secondly, there are not many empirical studies that have been published during the past seven months since ChatGPT's launch in November 2022, even though some are being conducted. This section briefly reviews (a) the feature components of information from ChatGPT responses that are critical for a learner to understand and learn new knowledge, (b) weaknesses and strengths of ChatGPT that educators need to consider when using ChatGPT, and (c) types of prompts that could determine what ChatGPT produces.

2.1. The Feature Components in the Text-Information

To learn new knowledge from text-based information (either printed, digital, web-based, or AI produced), sufficient information is necessary. The following feature components in the text-information are considered the same significant for learners to learn from ChatGPT responses as they learn from other traditional information resources (Choi et al., 2023; Fadel, 2019; Gibson et al., 2023; Schunk, 2004).

Concept and Definition. Concepts are defined as abstract ideas, which are understood to be the fundamental building blocks underlying principles, thoughts and beliefs (Goguen, J. (2005). A knowledge concept is discussed with conceptual knowledge that has been defined as understanding of the principles and relationships that underlie a domain (Gilmore & Cragg, 2018), and it is important for success in knowledge learning (Rittle-Johnson & Schneider, 2015). A clear definition of a knowledge concept would be helpful for the conceptual understanding of a theme, a theory, a subject, or a domain. Therefore, "what it is" is supposed to be defined clearly in the text-information for new knowledge.

Procedures and Examples. Procedures on a certain topic of knowledge (e.g., a statistics test, instructional design, calculation of a function, developing an app, etc.) mostly refer to a series of specific operations, steps, or instructions in a particularly defined order to accomplish some tasks, or reach certain conclusion and solution (Webster, 2023). In the explanations of the procedures, some other feature components can be used to assist learners to understand, such as using examples to demonstrate different conditions, problems, or methods.

Comparison and Contrast. In the demonstration of examples or description of procedures, comparison and contrast can be used to examine or establish similarities and dissimilarities, the difference or degree of difference between what is to be compared. The strategies of learning from comparison and contrast are derived from fundamental learning theories and have been applied in teaching and learning for years (Schunk, 2004).

Deductive and Inductive Argument. A deductive argument starts from a set of premises and reasons to a conclusion that is based on and supported by the premises. The reasoning, if done correctly, will result in a valid deduction: the truth of the premises ensures the truth of the conclusion (Norris, 1975). An inductive argument demonstrates a reasoning process of formulating a general principle from a set of evidence (e.g., observations), and making broad generalizations based on specific observations. Comparatively, the conclusion of a deductive argument is certain, given the premises are correct; the truth of the conclusion from an inductive argument is probable, based upon the evidence given (Copi et al., 2006).

Summary. A brief or coherent summary of the knowledge at the end of the response would be helpful for the learners' reinforcement of learning. For a short ChatGPT response, a summary may not be applicable. For a response that delivers a chunk of knowledge, whether a summary is included may indicate the level of quality of the text-information (Liu & Gibson, 2023).

It is still uncertain whether some of the feature components (e.g., comparison and contrast, or deductive and inductive reasoning) can be produced by ChatGPT, and again to what extent the information a learner receives from ChatGPT is accurate and reliable.

2.2. Weaknesses and Strengths of ChatGPT and Two-Way Learning

The strengths and weaknesses of ChatGPT revealed by experts and educational users may provide some insights to understand the functions of ChatGPT, and to find out the appropriate ways to use it.

Weaknesses. Quality of response is crucial to the success and effective adoption of ChatGPT for education (Tlili et al., 2023), and is likely to be influenced by three weaknesses. One weakness is the *lack of deep understanding of the words it processes.* ChatGPT can identify patterns of the words and generate plausible responses but does not fully comprehend the meanings behind the words (Farrokhnia, 2023; Gao et al., 2023), which may result in responses that lack depth and insight and potentially off-topic (Gupta et al., 2023). Regarding the quality of the feature components discussed in the previous section, this weakness at least influences the quality of *concept and definition.*

Another weakness is the *lack of capability to evaluate and determine the credibility of the data ChatGPT was trained on* (Lecler et al., 2023), which limits its capability to examine the accuracy of the generated information, and results in the uncertainties for the quality of the responses (Farrokhnia et al., 2023; Tlili et al., 2023). Therefore, even though the information on the feature components like *procedures and examples* is provided, the accuracy and reliability of the information are still questionable.

A weakness of ChatGPT that is not likely to be improved very soon is *Lack of higher-order thinking skills* such as critical and analytical thinking (Rudolph et al., 2023). This is mostly "because of the high dependency of AI tools on the data that they are trained without deep understanding of the context" (Farrokhnia et al., 2023, p.7). Therefore, feature components such as *comparison and contrast,* and *deductive and inductive argument* are not likely to occur in responses unless ChatGPT receives sufficient training with specific data generated from the prompts.

Strengths and Two-Way Learning. However, ChatGPT has been improved constantly through training language models to follow instructions with human feedback, and to summarize from human feedback (Ouyang et al., 2023; Stiennon et al., 2023). A unique feature of ChatGPT is its self-improvement or self-learning capability (Farrokhnia et al., 2023). The AI text generator of ChatGPT uses reinforcement learning from the user feedback to inform it language model, which enables ChatGPT to improve its responses based on the input prompts from users (Mann, 2023; Shen et al., 2023; Rudolph et al., 2023).

This strength enables the user-machine two-way learning process in a dynamic cycle (Liu & Gibson, 2023; Ouyang et al., 2023):

1. User's prompt – for GPT to learn,
2. GPT's responses – for user to learn,
3. User's prompt again – for GPT to learn and generate responses again, and
4. GPT's responses again – for user to learn again.

In this user-machine two-way learning, a user learns the information produced by the ChatGPT, and the ChaptGPT uses the human prompt/feedback for machine learning to generate more accurate response. The key point of this dynamic cycle of two-way learning is the *user's prompt.* How the prompt is formulated determines what the ChatGPT learns and processes, and what response is produced, especially what specific feature components are presented in the responses.

2.3. Types of Prompts

Based on the weaknesses, strengths and the user-machine two-way learning process, the design of integrating ChatGPT into education (for student learning) is mainly to design the prompts or questions that direct the ChatGPT language processing (Tlili et al., 2023). The prompts can be general, specific, or interactive (Liu & Gibson, 2023).

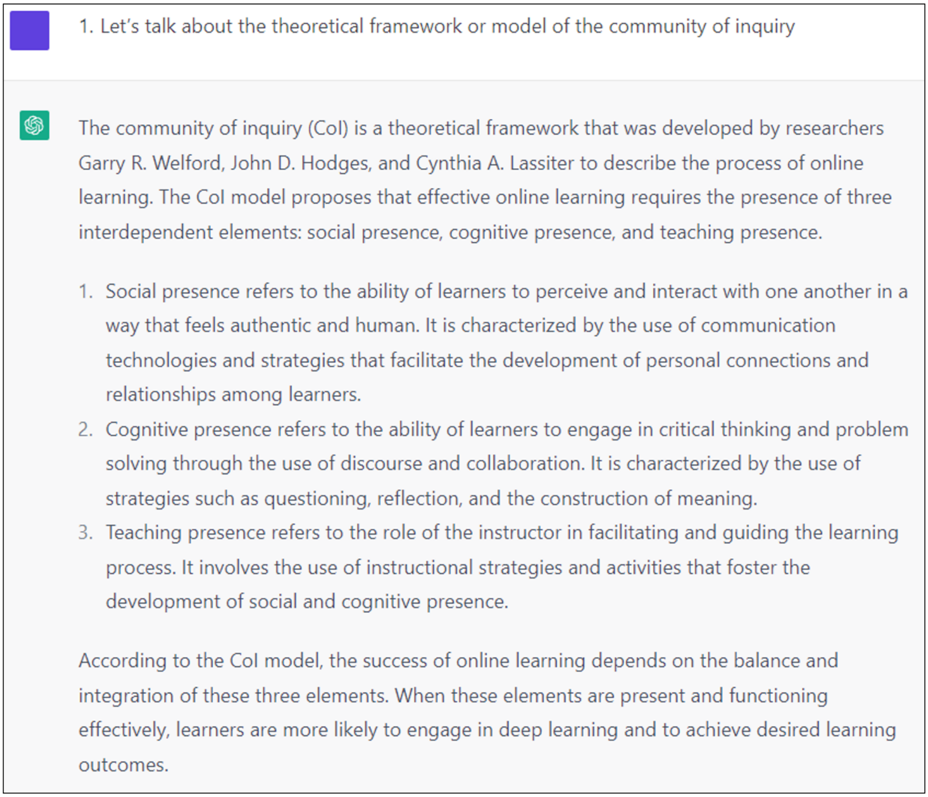
A general prompt, for example, can be: *Let's talk about COI theory (community of inquiry).* The ChatGPT produces a very brief response with general information about COI (Figure 1). A more specific prompt can be: *Let's talk about the theoretical framework or model of the community of inquiry*, or *What are the elements or factors of a learning environment based on community of inquiry?* The ChatGPT can provide some more specific information in the response for the first prompt (Figure 2), and even more definitions and explanations in the response for the second prompt.

0. Let's talk about COI theory (community of inquiry)

The theory of community of inquiry (COI) was developed by researchers to understand and promote effective online learning experiences. According to the COI theory, online learning is most effective when it is supported by three interdependent elements: cognitive presence, social presence, and teaching presence.

Figure 1

Sample ChatGPT response to a general prompt


1. Let's talk about the theoretical framework or model of the community of inquiry

The community of inquiry (CoI) is a theoretical framework that was developed by researchers Garry R. Welford, John D. Hodges, and Cynthia A. Lassiter to describe the process of online learning. The CoI model proposes that effective online learning requires the presence of three interdependent elements: social presence, cognitive presence, and teaching presence.

1. Social presence refers to the ability of learners to perceive and interact with one another in a way that feels authentic and human. It is characterized by the use of communication technologies and strategies that facilitate the development of personal connections and relationships among learners.
2. Cognitive presence refers to the ability of learners to engage in critical thinking and problem solving through the use of discourse and collaboration. It is characterized by the use of strategies such as questioning, reflection, and the construction of meaning.
3. Teaching presence refers to the role of the instructor in facilitating and guiding the learning process. It involves the use of instructional strategies and activities that foster the development of social and cognitive presence.

According to the CoI model, the success of online learning depends on the balance and integration of these three elements. When these elements are present and functioning effectively, learners are more likely to engage in deep learning and to achieve desired learning outcomes.

Figure 2

Sample ChatGPT response to a specific prompt

One weakness of ChatGPT is *lack of higher-order thinking skills* (Rudolph et. Al., 2023). However, if a prompt is specific enough to direct ChatGPT to perform some higher-order thinking skills such as comparison or contract, it will. For example, receiving a prompt like *What are the differences in the design of an online COI-based on learning environment and the design of an in-person COI-based learning environment*, ChatGPT can produce a response that compares several perspectives between the two types of COI designs.

Another type of prompt is interactive prompts. That is, the user and ChatGPT interact with a prompt-response-prompt-response cycle several times to reach a final or combined response content that is more accurate and reliable (Liu & Gibson, 2023). The interactive prompts are the typical two-way learning process, and the prompts mostly follow the contents or questions presented in the responses.

Educators' attentions have focused on what ChatGPT can produce, and to what extent the information it produced is at least correct and useful. In this study, the three types of prompts (general, specific, and interactive) and the responses generated from each type of prompt are examined regarding the presence of the six feature components.

3. Methodology

3.1. Purpose and Research Questions

The purpose of this study is to explore: (a) whether necessary feature components of text-information are provided in ChatGPT responses that are derived from different types of prompts, and (b) to what extent the presence of the feature components can influence the quality of response information regarding its accuracy and reliability. Three research questions were used to guide through the study and the analysis of the ChatGPT response contents:

1. Are the proportions associated with responses derived from general, specific, and interactive prompts the same between those with and without the presence of a certain feature component? – the six feature content components examined in this question are: concept and definition, procedure, example, comparison or contrast, deductive or inductive argument, and summary.

2. Can the probability that a ChatGPT response is accurate and reliable be predicted by the presence of any of the six feature content components — concept and definition, procedure, example, comparison or contrast, deductive or inductive argument, and summary?

3. To what extent do the significant feature components (if any from question 2) influence the probability of a ChatGPT response to be accurate and reliable?

3.2. The Sample and Data Collection

Sample. The sample of this study consisted of 253 ChatGPT responses. They were derived from the questions and prompts described next.

Questions/Prompts. Questions and prompts tested in this study were on the topics of (a) culturally responsive teaching and learning (Liu & David, 2023), (b) community of inquiry, (c) gamification, and (d) instructional design. There were eight initial questions on each topic. For example, one set of questions for gamification were:

1. Explain gamification in simple terms.
2. What are the key concepts of gamification?
3. How can teachers gamify their instruction?
4. Write a checklist for implementing gamification for an in-person classroom.
5. Write a checklist for implementing gamification for an online class.
6. Does gamification improve learning outcomes?
7. How can teachers make their classrooms more engaging through gamification?
8. What are some pitfalls to avoid when implementing gamification?

In these questions, Questions 1, 2, and 3 can be general prompts, which will lead to some overall introduction information. Such information is not necessarily incorrect, but not specific or comprehensive enough for learners to understand the domain. Questions 4, 5 and 7 can be more specific, by which ChatGPT will produce specific information. Questions 6 and 8 can be interactive prompts followed by continual questions of "how," "what," or the human-like interactive prompts like:

User: please provide more explanations (*general prompt*)

ChatGPT: do you mean some examples?

User: yes, an example of social interaction activities in gamification design? (*specific prompt*)

ChatGPT: one example can be …

Sometime, a question can be turned into an interactive question followed by random questions/prompts aiming at a certain content-point presented in the previous response. Then, with a comprehensive combination or edit, the final “responses” can be more accurate and more informative. In this study, besides the eight initial questions on each topic, some impromptu follow-up questions were often used to continue the “interactive conversations” with ChatGPT. Such interactive conversations were then combined into a final response as the result of a series of interactive prompts, which was coded as one case in the data.

The reason to use these very common topics for the exploration was that they were the topics in the courses the author has been teaching. Regarding the accuracy or reliability of the responses, the author was able to quickly evaluate the quality of the responses, and determine whether a response provides correct or useful information for a learner who is not familiar with the topic, and therefore to determine whether a learner can receive useful information, if he/she uses ChatGPT to browse the same content area by himself/herself at another time.

Data Collection Procedures. The author and four graduate students tested the questions/prompts and collected the responses. In a previous pilot exploration, the author found that if ChatGPT was asked a question more than twice at different times, or by different users from different accounts, the response contents were not identical, and some information was presented differently (Liu & Gibson, 2023). Therefore, in this study, the author and the four students used the same questions in each topic area at three different times. After removing invalid responses (e.g., empty response, or with a few words that cannot be coded), 253 responses remained for the data analysis. Among the 253 responses, 85 (33.5%) were from general prompts, 76 (30%) were from specific prompts, and 92 (36.5%) were from a combination of interactive prompts. This study did not sort the responses by the time each prompt was tested and did not examine the difference in responses upon time.

Power Analysis. According to the purpose of the content analysis, and the research questions, a logistic regression analysis should be used for the data analysis. Power analysis for performing a binary logistic regression with independent predictors was conducted. If the odds ratio is expected between 2 and 1.5, a minimum sample size between 225 and 637 should be reasonable (Liu et al., 2019). Therefore, a sample of 253 can be considered a proper sample size for current study.

3.3. Data Coding

The variables used in this study were coded as the following. First, the *quality of response (QR)* was examined. For a response, a value of 1 was coded for “satisfied” when the response included sufficient information that was correct, accurate, and appropriately addressed the prompt(s). Otherwise, a value of zero was coded for an “unsatisfied” response.

Then, the feature component variables were coded by examining whether two criteria were met: (a) the feature component should be presented in the responses, and (b) the feature component should be presented with clear and accurate information related to the prompt. For each of the six feature component variables, *concept and definition (CD), procedure (P), example (E), comparison or contrast (CC), deductive or inductive argument (DI),* and *summary (S)*, a value of 1 (Yes) was given if both criteria were met. Otherwise, a value of zero (No) was given. The definitions and descriptions of the feature components are described in Section 2. Table 1 shows the coding values for the variables.

Table 1

Variable Coding

Variables	Values	
	1	0
(QR) – Quality of Response	Satisfied	Unsatisfied
(CD) – Concept & Definition	Yes	No
(P) – Procedure	Yes	No
(E) – Example	Yes	No
(CC) – Comparison or Contrast	Yes	No
(DI) – Deductive or Inductive	Yes	No
(S) – Summary	Yes	No

3.4. Intrarater Reliability Analysis for the Coding

The coding for the quality of response and the six feature component variables was first completed using all 253 responses. After two months, the author revisited the data to check the reliability of the initial round of coding. According to Landis and Koch, 15% of the total number of the data will be considered as appropriate for the intrarater reliability check (Landis & Koch, 1977). In this study, the author decided to use 25% of the total 253 responses for the intrarater reliability analysis; 63 responses were randomly selected and recoded.

An intrarater reliability analysis using Cohen's Kappa statistic was conducted to determine the agreement of the coding results for the variables between the two coding periods (Cohen, 1960). Table 2 shows the intrarater reliabilities for the coding periods regarding the seven variables. A value of Kappa between .40 and .59 is considered moderate, between .60 and .79 is considered substantial, and above .80 is considered outstanding (Landis & Koch, 1977). Based on this guideline, the levels of agreement between the two coding periods regarding the seven variables were generally very good (as the values of Kappa ranged from .700 to .840).

Table 2

Levels of Agreement between Coding Periods (N = 63)

Variable	Kappa Coefficient	*p*-value
(QR) – Quality of Response	.802	<.001
(CD) – Concept & Definition	.741	<.001
(P) – Procedure	.700	<.001
(E) – Example	.840	<.001
(CC) – Comparison or Contrast	.803	<.001
(DI) – Deductive or Inductive	.804	<.001
(S) – Summary	.738	<.001

4. Data Analysis and Results

Data analyses for this study were undertaken for each research question. For research question 1, Chi-Square tests were conducted, and for research questions 2 and 3, a logistic regression analysis was performed. The data analysis procedures and results are reported as follows.

4.1. Data Analysis and Results for Research Question 1

Question 1. Are the proportions associated with responses derived from general, specific, and interactive prompts the same between those with and without the presence of a certain feature component?

In the data analysis for research question 1, each of the six feature component variables (concept and definition, procedure, example, comparison or contrast, deductive or inductive argument, summary) was examined with Chi-Square ($\chi 2$) tests by the three types of the responses that were derived from general, specific, and interactive prompts. That is, six 2 X 3 ($\chi 2$) tests were conducted, in which

- The Row Variable (A) = each of the feature component variables, with 2 categories (a1 = presented, a2 = not presented)
- The Column Variable (B) = the types of responses, with 3 categories (derived from b1 = general, b2 = specific, and b3 = interactive prompts)

The results from all six ($\chi 2$) tests are described next, including overall ($\chi 2$) test results, and follow up comparison test results.

Concept and Definition by Type of Response. The overall ($\chi 2$) test result was significant: $\chi 2$ (2, N=253) = 15.980, p < .001, and effect size Cramer's V = .231, p < .001, indicating that the proportions of the three types of responses (derived from general, specific and interactive prompts) were significantly different between those with or without the presence of the key concepts and accurate definitions. In 85 responses from general prompts (b1), 40 (47.1%) presented the concept and definition, and 45 (52.9%) did not. In 76 responses from specific prompts (b2), 52 (68.4%) were with the presence of concept and definition, and 24 (31.6%) were without. In 92 responses from interactive prompts (b3), 69 (75.0%) included concept and definitions, and 23 (33.5%) did not.

Among the three types of responses, follow up comparison Chi-square tests were conducted. First, the proportions of responses that presented concept and definition were significantly different between b1 and b2 responses: $\chi 2$ (1, N=161) = 7.477, p = .006, and effect size Phi (φ) = .215, p = .006. Second, the proportions of responses that presented concept and definition were significantly different between b1 and b3 responses: $\chi 2$ (1, N=177) = 14.579, p < .001, and effect size Phi (φ) = .287, p < .001. Third, the proportions of responses that presented concept and definition were NOT significantly different between b2 and b3 responses: $\chi 2$ (1, N=168) = 0.894, p = .344, and effect size Phi (φ) = .073, p = .344 (See Table 3).

Procedures by Type of Response. The overall ($\chi 2$) test result was significant: $\chi 2$ (2, N=253) = 11.966, p = .003, and effect size Cramer's V = .217, p = .003, indicating that the proportions of the three types of responses (derived from general, specific and interactive prompts) were significantly different between those with or without the presence of careful described procedures. In 85 responses from general prompts (b1), 47 (55.3%) presented the procedures, and 38 (44.7%) did not. In 76 responses from specific prompts (b2), 56 (73.7%) were with the presence of careful described procedures, and 20 (26.3%) were without. In 92 responses from interactive prompts (b3), 72 (78.3%) included description of the procedures, and 20 (21.7%) did not.

Among the three types of responses, follow up comparison Chi-square tests were conducted. First, the proportions of responses that presented procedures were significantly different between b1 and b2 responses: $\chi 2(1, N = 1\,61) = 5.888$, $p = .015$, and effect size Phi (φ) = .191, $p = .015$. Second, the proportions of responses that presented the procedures were significantly different between b1 and b3 responses: $\chi 2$ (1, N=177) = 10.578, $p < .001$, and effect size Phi (φ) = .244, $p < .001$. Third, the proportions of responses that presented procedures were NOT significantly different between b2 and b3 responses: $\chi 2$ (1, N=168) = 0.481, $p = .488$, and effect size Phi (φ) = .053, $p = .488$ (See Table 3).

Example by Type of Response. The overall (χ2) test result was significant: $\chi 2$ (2, N=253) = 12.040, $p = .002$, and effect size Cramer's V = .218, $p = .002$, indicating that the proportions of the three types of responses (derived from general, specific and interactive prompts) were significantly different between those with or without the presence of appropriate or sufficient examples. In 85 responses from general prompts (b1), 36 (42.4%) presented the examples, and 49 (57.6%) did not. In 76 responses from specific prompts (b2), 53 (68.4%) were with the presence of careful selected examples, and 24 (31.6%) were without. In 92 responses from interactive prompts (b3), 56 (60.9%) included examples, and 36 (39.1%) did not.

Among the three types of responses, follow up comparison Chi-square tests were conducted. First, the proportions of responses that presented appropriate examples were significantly different between b1 and b2 responses: $\chi 2$ (1, N=161) = 11.002, $p < .001$, and effect size Phi (φ) = .261, $p < .001$. Second, the proportions of responses that presented appropriate examples were significantly different between b1 and b3 responses: $\chi 2$ (1, N=177) = 6.069, $p = .014$, and effect size Phi (φ) = .185, $p = .014$. Third, the proportions of responses that presented examples were NOT significantly different between b2 and b3 responses: $\chi 2$ (1, N=168) = 1.034, $p = .309$, and effect size Phi (φ) = .078, $p = .309$ (See Table 3).

Comparison or Contrast by Type of Response. The overall ($\chi 2$) test result was significant: $\chi 2$ (2, N=253) = 8.087, $p = .018$, and effect size Cramer's V = .179, $p = .018$, indicating that the proportions of the three types of responses (derived from general, specific and interactive prompts) were significantly different between those with or without the presence of comparison or contrast descriptions. In 85 responses from general prompts (b1), 45 (52.9%) presented comparison or contrast descriptions, and 40 (47.1%) did not. In 76 responses from specific prompts (b2), 56 (73.7%) were with the presence of comparison or contrast contents, and 20 (26.3%) were without. In 92 responses from interactive prompts (b3), 62 (67.4%) included comparison or contrast descriptions, and 30 (32.6%) did not.

Among the three types of responses, follow up comparison Chi-square tests were conducted. First, the proportions of responses that presented appropriate comparison or contrast were significantly different between b1 and b2 responses: *χ2* (1, *N*=161) = 7.385, *p* = .007, and effect size Phi (*φ*) = .214, *p* = .007. Second, the proportions of responses that presented appropriate comparison or contrast were significantly different between b1 and b3 responses: *χ2* (1, *N*=177) = 3.859, *p* = .049, and effect size Phi (*φ*) = .148, *p* = .049. Third, the proportions of responses that presented comparison or contrast were NOT significantly different between b2 and b3 responses: *χ2* (1, *N*=168) = 0.788, *p* = .375, and effect size Phi (*φ*) = .069, *p* = .375 (See Table 3).

Deductive or Inductive by Type of Response. The overall (*χ2*) test result was significant: *χ2* (2, *N*=253) = 24.249, *p* < .001, and effect size Cramer's *V* = .310, *p* < .001, indicating that the proportions of the three types of responses (derived from general, specific and interactive prompts) were significantly different between those with or without the presence of deductive or inductive arguments or descriptions. In 85 responses from general prompts (b1), 33 (38.8%) presented deductive or inductive descriptions, and 52(61.2%) did not. In 76 responses from specific prompts (b2), 55 (72.4%) were with the presence of deductive or inductive contents, and 21 (27.6%) were without. In 92 responses from interactive prompts (b3), 64 (69.6%) included deductive or inductive descriptions, and 28 (30.4%) did not.

Among the three types of responses, follow up comparison Chi-square tests were conducted. First, the proportions of responses that presented appropriate deductive or inductive arguments were significantly different between b1 and b2 responses: *χ2* (1, *N*=161) = 18.218, *p* < .001, and effect size Phi (*φ*) = .336, *p* < .001. Second, the proportions of responses that presented appropriate deductive or inductive contents were significantly different between b1 and b3 responses: *χ2* (1, *N*=177) = 18.875, *p* < .001, and effect size Phi (*φ*) = .309, *p* < .001. Third, the proportions of responses that presented deductive or inductive contents were NOT significantly different between b2 and b3 responses: *χ2* (1, *N*=168) = 0.158, *p* = .691, and effect size Phi (*φ*) = .031, *p* = .691 (See Table 3).

Summary by Type of Response. The overall (*χ2*) test result was significant: *χ2* (2, *N*=253) = 16.628, *p* < .001, and effect size Cramer's *V* = .256, *p* < .001, indicating that the proportions of the three types of responses (derived from general, specific and interactive prompts) were significantly different between those with or without the presence of a comprehensive summary. In 85 responses from general prompts (b1), 37 (43.5%) presented a summary, and 48(56.5%) did not. In 76 responses from specific prompts (b2), 34 (44.7%) were with the presence of a well described summary, and 42 (55.3%) were without. In 92 responses from interactive prompts (b3), 65(70.7%) included a summary, and 27 (29.3%) did not.

Among the three types of responses, follow up comparison Chi-square tests were conducted. First, the proportions of responses that presented a summary were NOT significantly different between b1 and b2 responses: *χ2* (1, *N*=161) = 0.024, *p* = .878, and effect size Phi (*φ*) = .012, *p* < .878. Second, the proportions of responses that presented appropriate a well described summary were significantly different between b1 and b3 responses: *χ2* (1, *N*=177) = 13.310, *p* < .001, and effect size Phi (*φ*) = .274, *p* < .001. Third, the proportions of responses that presented deductive or inductive contents were significantly different between b2 and b3 responses: *χ2* (1, *N*=168) = 11.549, *p* < .001, and effect size Phi (*φ*) = .262, *p* < .001 (See Table 3).

Table 3

Follow-up Comparison Chi-Square Results

Variable A Types of Responses (From the Three Types of Prompts)				Chi-Square Results $df = 1$			
Variable Bs	*a1* General	*a2* Specific	*a3* Interactive	$\chi 2$	*N*	*p*	*Phi (φ)**
Concept & Definition				7.474	161	.006	.215
				14.579	177	< .001	.287
				0.894	168	.344	.073
Procedure				5.888	161	.015	.191
				10.578	177	<.001	.244
				0.481	168	.488	.053
Example				11.002	161	< .001	.261
				6.069	177	.014	.185
				1.034	168	.309	.078
Comparison or Contrast				7.385	161	.007	.214
				3.859	177	.049	.148
				0.788	168	.375	.069
Deductive or Inductive				18.218	161	< .001	.336
				18.875	177	< .001	.309
				0.158	168	.691	.031
Summary				0.024	161	.878	.012
				13.310	177	< .001	.274
				11.549	168	< .001	.262

Notes: 1. (*), all the Phi (*φ*) tests had the same significant level of p as in each of the $\chi 2$ tests.

2. Shaded cells indicated the types of prompts in each $\chi 2$ test.

In Chi-square test, effect size Phi (*φ*) is used for 2 by 2 tests, and Cramer's V is used for tests in which either the row variable or column variable (or both) is more than two categories, (e.g., in this study, the 2 by 3 test in this study). For both tests, the values about .10, .30, and .50 were considered a relatively small, medium, and large association respectively (Sprent & Smeeton, 2007) between the row and column variables (e.g., types of responses and each feature component variable in this study).

The overall results showed a pattern that responses derived from specific or interactive prompts tended to present more feature content components that may help users to understand the contents in a certain domain than the responses derived from general prompts.

4.2. Data Analysis and Results for Research Questions 2 and 3

Question 2. Can the probability that a ChatGPT response is accurate and reliable be predicted by the presence of any of the six feature content components — concept and definition, procedure, example, comparison or contrast, deductive or inductive argument, and summary?

Question 3. To what extent do the significant feature components (if any from question 2) influence the probability of a ChatGPT response to be accurate and reliable?

For research questions 2 and 3, logistic regression analyses were conducted. The six feature component variables were used as the explanatory variables, and the quality of response was the response variable. The frequencies for each variable are shown in Table 4.

Table 4

Frequencies

Variables	Values	
	1	**0**
(QR) – Quality of Response (RV)	168	85
(CD) – Concept & Definition (EV)	161	92
(P) – Procedure (EV)	175	78
(E) – Example (EV)	144	109
(CC) – Comparison or Contrast (EV)	163	90
(DI) – Deductive or Inductive (EV)	152	101
(S) – Summary (EV)	136	117

Note: RV—Response Variable, EV—Explanatory Variable

In the first logistic regression analysis, all six feature component variables were included. Results showed that one explanatory variable, Comparison or Contract (*CC*), was not significant (Wald $\chi 2$ = 1.919, p = .166). Therefore, it was removed from the model. The second logistic regression analysis was conducted with the other five explanatory variables as shown in Table 5.

Model summary results showed that the model with these five explanatory variables was significant ($\chi 2$ = 85.588, p < .001) and accounted for about 39.8% of the variation in the response variable (Nagelkerke R^2 = .398), indicating that this model significantly predicts group membership. The Hosmer and Lemeshow Goodness-of-Fit Statistic of 7.298 (p=.505) was not significant, indicating that the hypothesis that the model provides a good fit of data should be accepted. Specifically, 52 out of 85 unsatisfied responses (61.2%), 145 out of 186 satisfied responses (86.3%), and a total of 197 out of 253 responses (77.9%) were correctly predicted by the model.

Table 5

Logistic Regression Outputs

	DF	Parameter Estimate	Standard Error	Wald Chi-Square	*P*	Odds Ratio
(CD)	1	0.868	0.332	6.834	.009	2.381
(P)	1	1.334	0.339	15.455	< .001	3.796
(E)	1	0.898	0.328	7.010	.008	2.383
(DI)	1	1.386	0.327	17.915	<.001	3.998
(S)	1	0.713	0.330	4.674	.031	2.041
Constant	1	-2.364	0.423	31.177	< .001	0.094

Response variable: Quality of Response (QR)

Explanatory variables: Concept & Definition (CD), Procedure (P), Example (E), Deductive or Inductive (DI), and Summary (S)

The logistic regression results were shown in Table 5. A significant Wald chi-square value for a given variable indicates that the variable is significantly related to the response variable. As shown in Table 5, the Wald chi-square values are significant for all five explanatory variables. Therefore, all five explanatory variables are included in the model equation. The Parameter Estimate generates the estimated coefficients of the fitted logistic regression model, and they are used to formulate the following logistic regression equation (1):

$$\text{logit}\,(\hat{p}) = -2.364 + 0.868(CD) + 1.334(P) + 0.898(E) + 1.386(DI) + 0.713(S) \text{ ----------------- } (1)$$

The sign ($\hat{p}$) indicates an estimated probability value (also called *log odds*) for the response variable Quality of Response (*QR*) to be 1, and logit represents *logit transformation* of the event probability.

An estimated coefficient indicates the contribution that explanatory variable makes to the possibility of the response variable being 1. For example, when the variable *Concept and Definition* (*CD*) is 1 (that is, when the concept and definition is clearly presented in the responses), the logit transformation of event probability (that a response presents accurate and reliable information with satisfied quality) increases by 0.868 (see Table 5). The estimated coefficients for the other four explanatory variables can be interpreted similarly.

Odds ratio is another statistic to explain the contribution of an explanatory variable to the model. If the odds ratio for a given explanatory variable is larger than 1, the probability of the response variable being 1 increases because of the presence of that explanatory variable. For example, the odds ratio for variable*Concept and Definition* (*CD*) is 2.381 (see Table 5), indicating that a response would be 2.381 times more likely to present accurate and reliable information with satisfied quality if clear and well formulated concept and definition are presented, compared to responses that do not have concept and definitions presented. If the odds ratio is smaller than 1, the probability of the response variable being 1 decreases (that is, the probability that a response presents accurate and reliable information with quality decreases when that explanatory variable exists). As seen in Table 5, all five odds ratio values are larger than 1 (ranged from 2.041 to 3.998), therefore, all five variables positively contribute to a response with satisfied quality of information.

According to the results, a predictive model can be summarized as in the following model function equation (2) in Figure 3:

P (QR=1) = *f* [CD, P, E, DI, S] -- (2)
Where:
QR= Quality of Response, **P (QR=1) =** Probability that QR provides satisfied quality
***f* […]** indicates "a function of …"
CD = Concept and Definition, **P** = Procedure, **E** = Example,
DI= Deductive or Inductive, **S** = Summary

Figure 3
The predictive model function of quality of response (QR)

The model function reads "the probability that a response provides satisfied quality is a function of the five feature component variables: concept and definition, procedure, example, deductive or inductive argument, and summary."

4.3. Summary of Findings

In summary, first, the three types of prompts (general, specific, and interactive) can lead to responses with different qualities. ChatGPT responses derived from *specific* and *interactive* prompts demonstrate the inclusion of more feature components with more in depth information. Second, five feature components variables (*Concept & Definition, Procedure, Example, Deductive or Inductive, and Summary*) are significant influential variables, which can be used to predict the probability that a response presents accurate and reliable information with satisfied qualities. The question is: what does this mean to educators and researchers?

5. Open-Ended Discussions and Conclusions

ChatGPT has just come to educators for about seven months, and the exploratory practice in the field has just started. Many current publications are position papers or suggested "how-to" papers (Adiguzel et al., 2023; Choi et al., 2023). Even in this present study, although significant results are found, there are still limitations and uncertainties. The author is closing this article with the following open-ended discussions that potentially reach some open-ended conclusions from several perspectives of using ChatGPT: (a) what users ask – prompts, two-way learning process, and a catch-22 dilemma, (b) what users get – quality of the information, (c) effect size and further assessment, and (d) a theoretical framework to integrate ChatGPT into Education.

5.1. Prompts, Two-Way Learning, and a Catch-22

In this study, the results indicate that specific and interactive prompts for the user-ChatGPT conversation have led to responses with more feature components, and more accurate and reliable information. However, specific prompts require an understanding of the knowledge domain, and interactive prompts require even more in depth understanding of two or more steps further regarding the knowledge or conversation. When human feedback constantly provides the data and information based on which the language model is trained, ChatGPT learns and generates more meaningful information for users to learn (Farrokhnia et al., 2023).

Meanwhile, a catch-22 dilemma seems obvious. On one hand, if a user does not have the knowledge, he/she may not be able to ask meaningful questions, or start any specific/interactive prompts, and hence, he/she may not receive appropriate information from ChatGPT. On the other hand, if a user has already had the knowledge to ask the right questions, initiate the appropriate prompts, or develop informative conversations, he/she may not really need to chat with the GPT or learn from what it can provide.

Random or impromptu prompts will not lead to responses with in-depth information. Therefore, careful design on the prompts is a sufficient and necessary condition to the success of using ChatGPT for knowledge learning. An open-ended question for an educator may be: How should the criteria to evaluate the "success" be connected to the design of different types of prompts (e.g., general, specific, interactive, linear-layered, nonlinear layered, or higher order thinking aligned)?

5.2. Quality of Responses

In this study, when coding the quality of ChatGPT responses, the author realized that some responses only provided knowledge contents at the surface level of the domain knowledge, even though the contents are accurate, clear, and appropriately addressed the prompts, and major feature components are included. The author conducted this analysis from the perspective of a teacher to look at what ChatGPT can produce, based on the fact that she is familiar with the knowledge domain, as the four topics are in the teaching contents from her courses. Compared with student learning objectives set in the courses on the four topics or subtopics under each, some responses mostly are like those for your information (FYI) or for your reference (FYR).

The FYI or FYR contents can be a start point of learning, however, at current time ChatGPT responses should not be used as the ONLY resource for student learning, it may work well with additional resources of literature and learning materials. To explore the positive use of ChatGPT for learning, some open-ended questions for an educator may be: What information produced by ChatGPT is important or valuable, and should be included in current teaching materials? What information is not valid for learning, so should not be used by students?

5.3. Effect Size and Further Assessment

In Chi-square tests, the effect size statistics Cramer's V and Phi (φ) explain the degree of association between the row variable (feature component variable) and column variable (response from three types of prompts). In this study, the values of effect size Cramer's V for overall Chi-square tests ranged from .179 to .310. The values of effect size Phi (φ) for all the follow up comparison tests ranged from .148 to .336 for the significant results (Table 3). Both ranges indicate medium degree of the association between the two variables.

In a logistics regression analysis, odds ratio is an effect size statistic to explain the contribution of a predictor to the model. In this study, odds ratios of the five predictors ranged from 2.041 to 3.998 (Table 5), revealing their positive influence on the probability of a ChatGPT response having a desired quality. All the odds ratios are larger than 1, indicating that all explanatory variables positively contribute to the variation of the response variable.

It is not likely to compare the range of effect size from similar studies in recent literature. The author did not find relevant literature on similar ChatGPT related studies that examine the effect size of Cramer's V, Phi (φ), and odds ratio. In further studies, assessment on ChatGPT related learning will be a widely explored area. An open-ended question for researchers may be: to what extent could the consistence of effect size among ChatGPT related studies be likely reached?

5.4. A Brief Theoretical Framework

In the history of technology development, every time a new technology/tool comes to the field, educators are confronted with wonders, concerns, or issues that eventually lead to a series of decisions for them to make (Liu et al., 2019; Liu & Velasques-Bryant, 2003). A summary of three design models may provide a fundamental theoretical framework to guide educators' practice and researchers' further studies.

First, an ITD technology integration model merges the design among three dimensions (Information, Technology, and Design). It was promoted by Liu and Velasques-Bryant (2003), and has been examined, tested, and updated into a dynamic design model (Liu, 2017). Studies have shown that any technology integration case would not be successful if any single one dimension of the model was missing. Second, the ADDIE design model (Gagné et al., 2005) has been applied in the field over decades. It includes five phases of design (Analysis, Design, Development, Implementation, and Evaluation) and all the tasks under each phase. It was promoted into a dynamic design model, which worked effectively as well (Liu, 2017). Third, a new theoretical framework (Gibson et al., 2023) for AI-promoting learning processes at three levels (Micro-level of individual learner, Meso-level of team activity, and Macro-level for larger emergent cultural entities) now can be used as theoretical guidance for the design of integrating ChatGPT at the three levels.

All three approaches of design in technology integration together formulate a comprehensive and practical theoretical framework. This framework provides overall guidance, strategies, and instructions for the integration of ChatGPT into teaching and learning. Findings from the present study also reveal one specific key point of such design. That is, the design of prompts is a crucial key to promote more positive use of ChatGPT in learning.

5.5. Limitations and Further Studies

Limitations. One limitation of this study is that it focuses only on the text-response contents for knowledge learning in the four topics: culturally responsive teaching and learning, community of inquiry, gamification, and instructional design. Other types of using ChatGPT are not included in this study (such as composition writing, calculation, programing, solving problems, critique and review for academic articles, drawing, or video production), so other types of responses are not studied.

Another limitation is that the analysis and results are based on the responses which to certain extent are lacking in-depth knowledge or information (as described in section 5.2.), since that is what ChatGPT can produce at the time. Based on such information, it is not very convincing to make implications about ChatGPT based learning or judgements on the use of ChatGPT in general. Further studies to examine the quality of responses in different ways will need to continue.

Also, this study only serves as an initial pilot, mainly exploring the positive use of ChatGPT. It does not address concerns about some negative use of ChatGPT, such as plagiarism (Kasneci et al., 2023; Qadir, 2022); they are beyond the purpose of present study.

Future Studies. ChatGPT related research is a relatively new area. The open-ended questions raised in previous sections (e.g., in sections 5.1., 5.2., and 5.3.) can be expanded into branches of educational research agenda for future studies. For example, studies can be on: (a) examining the success of using ChatGPT in learning, (b) content analysis on the quality of ChatGPT responses, (c) modeling in design and integration, (d) analyzing effect sizes in ChatGPT related studies, (e) measurement and assessment of AI-based learning – instrument development, (f) methods to promote dynamic user-ChatGPT two-way learning, (g) collaborative learning with ChatGPT, (h) using ChatGPT for interdisciplinary studies, (i) interactive use of AI-tools in education, and more.

Finally, back to the basics, the ideas or research agenda used in technology integration or instructional design and technology over years can still be applicable, except that ChatGPT is another new technology tool with its unique features and functions. This study tends to serve as an initial exploration that hopefully opens more paths for further studies. It is the author's wish that the findings from this study can be of reference to other educators and researchers when they are to explore the effective use of ChatGPT in their work.

References

Adiguzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology, 15*(3), ep429. https://doi.org/10.30935/cedtech/13152

Atlas, S. (2023). ChatGPT for higher education and professional development: A guide to conversational AI. Available: https://digitalcommons.uri.edu/cba_facpubs/548

Caldarini, G., Jaf, S., & McGarry, K. (2022). A literature survey of recent advances in Chatbots. *Information, 13*(1), 41. https://doi.org/10.3390/INFO13010041

Carter, S., & Nielsen, M. (2017). Using artificial intelligence to augment human intelligence. Retrieved May 20, 2023, from: https://distill.pub/2017/aia/ https://doi.org/10.23915/distill.00009

Choi, S., Jang, Y., & Kim, H. (2023). Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational artificial intelligence tools. *International Journal of Human-Computer Interaction, 39*(4), 910–922.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Copi, I. M., Cohen, C., & Flage, D. E. (2006). *Essentials of logic* (2nd ed.). Upper Saddle River, NJ: Pearson Education.

Cotton, D. R., Cotton, P. A., & Shipway, J. (2023). Chatting and cheating. In Ensuring academic integrity in the era of ChatGPT. EdArXiv. Retrieved June 25, 2023, from https://edarxiv.org/mrz8h/

Dijkstra, R., Genç, Z., Kayal, S., & Kamps, J. (2022). Reading comprehension quiz generation using generative pre-trained transformers. https://e.humanities.uva.nl/publications/2022/dijk_read22.pdf.

Fadel, C., Holmes, W., & Bialik, M. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. *The Center for Curriculum Redesign.* Retrieved May 31, 2023 from https://curriculumredesign.org/wp-content/uploads/AIED-Book-Excerpt-CCR.pdf

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

Gabajiwala, E., Mehta, P., Singh, R., & Koshy, R. (2022). Quiz maker: Automatic quiz generation from text using NLP. In *Futuristic trends in networks and computing technologies* (pp. 523–533). Singapore: Springer

Gagné, R. M., Wager, W. W., Golas, K. C., & Keller, J. M. (2005). *Principles of instructional design* (5th ed.). Belmont, CA: Thomson Wadsworth.

Gao, J., Zhao, H., Yu, C., & Xu, R. (2023). Exploring the feasibility of ChatGPT for event extraction. arXiv. https://doi.org/10.48550/arXiv.2303.03836

Gibson, D., Kovanovic, V., Ifenthaler, D., Dexter, S., & Feng, S. (2023). Learning theories for artificial intelligence promoting learning processes. *British Journal of Educational Technology, 2023*, 1-22. https://doi.org/10.1111/bjet.13341

Gilmore, C., & Cragg, L. (2018). The role of executive function skills in the development of children's mathematical competencies. In *Heterogeneity of Function in Numerical Cognition* (pp. 263-286). https://doi.org/10.1016/B978-0-12-811529-9.00014-5

Goguen, J. (2005). What is a Concept? Conceptual Structures: Common Semantics for Sharing Knowledge. *Lecture Notes in Computer Science.* Vol. 3596. pp. 52–77. https://doi.

org/10.1007/11524564_4.

Gupta, P., Raturi, S., & Venkateswarlu, P. (2023). Chatgpt for designing course outlines: A boon or bane to modern technology. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.4386113

Halaweh, M. (2023). ChatGPT in education: Strategies for responsible implementation. *Contemporary Educational Technology, 15*(2), ep421. https://doi.org/10.30935/cedtech/ 12036

Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Gunnemann, S., Jullermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*(2023), 102274, 1-9. https://doi.org/10.1016/j.lindif.2023.102274

Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

Lecler, A., Duron, L., & Soyer, P. (2023). Revolutionizing radiology with GPT-based models: Current applications, future possibilities, and limitations of ChatGPT. *Diagnostic and Interventional Imaging.* https://doi.org/10.1016/j.diii.2023.02.003

Liu, L. (2017). An Initial Model of Dynamic Instructional Design and Learning Analytics: The Power of Dynamic Assessment and Prediction. *Keynote Speaker at The Association for Educational Communications and Technology (AECT)*, November 8-11, Jacksonville, FL.

Liu, L., Chen, L., & Li, W. (2019). Social Media in Dynamic Learning: Logistics and Influential Factors. *Journal of Educational Technology Development and Exchange, 12*(1), 33-57. DOI: 10.18785/jetde.1201.03

Liu, L., & Gibson, D. (2023). Exploring the use of ChatGPT for learning and research: Content data analysis and concerns. *Presentation at Society for Information Technology & Teacher Education International Conference,* March 13-17, New Orleans, Louisiana. https://www.learntechlib.org/p/221924/

Liu, L., & Velasques-Bryant, N. J. (2003). An information technology integration system and its life cycle: What is missing? *Computers in the Schools, 20*(1/2), 93-106.

Mann, D. L. (2023). Artificial Intelligence discusses the role of artificial intelligence in translational medicine. *JACC: Basic to Translational Science, 8*(2), 221–223. https://doi.org/10.1016/j.jacbts.2023.01.001

Norris, S. E. (1975). The intelligibility of practical reasoning. *American Philosophical Quarterly, 12*(1), 77–84.

OpenAI. (2022). *ChatGPT*. https://openai.com/chatgpt

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder,P., Christiano, P., Leike, J., & Lowe, R. (2023). Training language models to follow instructions with human feedback. arXiv:2203.02155. https://doi.org/10.48550/arXiv.2203.02155

Qadir, J. (2022): Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. *TechRxiv*, v1. https://doi.org/10.36227/techrxiv.21789434

Rittle-Johnson, B., & Schneider, M. (2015). Developing conceptual and procedural knowledge of mathematics. In R. C. Kadosh & A. Dowker (Eds.), *The Oxford handbook of numerical cognition* (pp. 1118–1134). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199642342.013.014

Roose, K. (2022). The brilliance and weirdness of ChatGPT. *New York Times,* December 5, 2022. Retrieved December 26, 2022, from https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching, 6*(1). https://doi.org/10.37074/jalt. 2023.6.1.9

Schunk, D. H. (2004). *Learning theories: An educational perspective,* 4th ed. Columbus, OH: Pearson.

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). ChatGPT and other large language models are double-edged words. *Radiology*. https://doi.org/10.1148/radiol.230163

Sprent, P., & Smeeton, N. C. (2007). *Applied nonparametric statistical methods*, 4th Ed. Taylor & Francis Group.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2023). Learning to summarize from human feedback. arXiv:2009.01325. https://doi.org/10.48550/arXiv.2009.01325

Talian, T., & Kalinkara, Y. (2023). The role of artificial intelligence in higher education: ChatGPT assessment for anatomy course. *International Journal of Management Information Systems and Computer Science, 7*(1), 33-40.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkrut, A., Tichey, D. T., Huang, R., & Agyemang B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments,* (2023)10:15. https://doi.org/10.1186/s40561-023-00237-x.

Webster. (2023). Webster Dictionary. Retrieved May 25, 2023 from: https://www.merriam-webster.com/dictionary/procedure

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. *IEEE Internet of Things Journal, 8*(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

Author information:

Leping Liu https://orcid.org/0000-0001-5859-8189

Professor of Information Technology and Statistics, University of Nevada, Reno, USA.