Unleashing ChatGPT's Power: A Case Study on Optimizing Information Retrieval in Flipped Classrooms via Prompt Engineering

Mo Wang[†,&], Minjuan Wang[‡,&], Xin Xu[§,*], Lanqing Yang[¶], Dunbo Cai[‖], Minghao Yin[¶,*]
[†]School of Fine Arts, Northeast Normal University, Changchun, Jilin, China, Email: wangm875@nenu.edu.cn
[‡]Learning Design and Technology, San Diego State University, California, USA, Email: mwang@sdsu.edu
[§] School of Mediascience/School of Journalism, Northeast Normal University, Changchun, Jilin, China, Email:xux894@nenu.edu.cn
[¶] School of Information Science and Technology, Northeast Normal University, Changchun, Jilin, China, Email:{yanglq358,ymh}@nenu.edu.cn
[‖] Center for Technology Research & Innovation, China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou, China, Email:caidunbo_yewu@cmss.chinamobile.com

***Abstract*—This research project investigates the impact of Prompt Engineering, a key aspect of ChatGPT, on college students' information retrieval in flipped classrooms. In recent years, an increasing number of students have been using AI-based tools such as ChatGPT rather than traditional research engines to learn and to complete course assignments. Despite this growing trend, previous research has largely overlooked the influence of prompt engineering on students' use of ChatGPT and effective strategies for improving the quality of information retrieval in learning settings. To address this research gap, this study examines the information quality obtained from ChatGPT in a flipped classroom by evaluating its effectiveness in task completion among 26 novice undergraduates from the same major and cohort. The experimental results provide evidence that proficient mastery of prompt engineering improves the quality of information obtained by students using ChatGPT. Consequently, by acquiring proficiency in prompt engineering, students can maximize the positive impact of ChatGPT, obtain high-quality information, and enhance their learning efficiency in flipped classrooms.**

***Index Terms*—ChatGPT, Prompt Engineering, Flipped Classrooms, Information Retrieval.**

I. Introduction

CHATGPT, an artificial intelligence (AI) conversational chatbot developed by OpenAI, since its release on November 30, 2022, has garnered worldwide interest [1]. As commented in the journal Nature, "Conversational AI is a game-changer for science" [2]. Conversational AI tools like ChatGPT [3] and ChatGLM [4] employ Large Language Models (LLMs) to mimic human conversation, comprehending and providing responses to user queries in a manner reminiscent of human interaction. These AI-based chatbots have proven to be invaluable assets in various sectors, such as customer service [5] and healthcare [6], as they can optimize workflow efficiencies, minimize expenses, and elevate the overall user experience.

ChatGPT is an advanced language model with the remarkable capability of generating text that closely resembles human language in response to given prompts [7]. By providing a concise description of a code snippet, along with pertinent details and constraints, ChatGPT can be initiated. To illustrate, suppose we require a function to compute the average of a set of numbers. In this case, the following prompt can be used to initiate ChatGPT: "Write a Python function that takes a list of numbers as an argument and returns their average."

The potential for significant improvement lies in augmenting flipped learning through the use of AI-based chatbots [8]. The flipped classroom pedagogy, also known as the inverted classroom, is an innovative and widely embraced teaching approach. It involves the transformation of traditional in-class activities, such as instructor presentations, into homework assignments, while projects or tasks typically assigned as homework are completed during class time [9]–[11].

Recent meta-analyses have indicated that flipped learning has the potential to enhance student achievement across various subject disciplines [12], [13]. However, it is essential to acknowledge that implementing this approach comes with its fair share of challenges. Akçayır et al. [9] discovered two prominent issues associated with flipped learning. Firstly, students often lack proper guidelines or instructions when studying at home. Secondly, they encounter difficulty in seeking assistance during the pre-class learning phase, which subsequently hinders their active participation in in-class activities. Many studies have shown that AI-based chatbots can play a variety of roles, such as learning chatbots [14]–[16], assistant chatbots [17], [18], and mentor chatbots [19], [20]. AI-based chatbots can address these challenges by providing students with 24/7 assistance and personalized support, thereby enhancing their engagement in pre-class learning activities [8].

Although there are many studies on flipped classrooms and AI-based chatbots [13], [21], [22], few have focused on

&. These authors contributed equally to this work and should be considered co-first authors

*. Corresponding Authors

how to improve students' skills in using chatbots to obtain high-quality responses, thus highlighting the need for further research in this area. ChatGPT has been criticized for relying on biased data and potentially providing inaccurate or false information [23]. There can be potential issues and considerations associated with students using ChatGPT without guidance. Therefore, it is imperative to enhance the capability of using ChatGPT to prevent students from accessing harmful information while incorporating it into a flipped classroom setting.

Prompt engineering, a technique involving strategically designed input prompts, plays a crucial role in obtaining more precise responses from AI-based chatbots generated by Large Language Models (LLMs), which are trained on a substantial amount of content and generally exhibit the ability to generate accurate results based on task descriptions. In this context, the task description is referred to as a prompt. Unlike human communication, which often relies on ambiguous cues, LLMs require clearer guidance or specific phrasing to achieve optimal understanding and response generation. Prompt engineering encompasses various elements, such as questions, keywords, and contextual information, all aimed at enhancing the model's comprehension of user needs and improving response accuracy.

Effective prompts empower users to leverage the powerful capabilities of LLMs, obtaining accurate and relevant responses that enhance work efficiency and problem-solving capabilities. Prompt engineering, through gaining a better understanding of user needs, has the potential to significantly improve user experience, their satisfaction with the tool, and thus to realize the optimal use of LLMs.

Having a strong grasp of prompt engineering techniques is essential for unlocking the full benefits discussed in the previous paragraph. Mastery of these techniques allows users to effectively harness the potential of LLMs. By leveraging prompts in the right way, users gain access to the powerful capabilities of LLMs, resulting in responses that are more accurate and relevant. This, in turn, leads to improved work efficiency and enhanced problem-solving capabilities. Understanding how to effectively use prompts guides users in maximizing the potential of LLMs so as to reap the benefits of prompt engineering.

Despite the significance of prompt engineering highlighted in numerous studies and the exploration of methods to elicit high-quality responses from LLMs [24]–[28], there remains a research gap regarding the impact of prompt engineering on students' participation in ChatGPT-enabled flipped classrooms. Therefore, the primary objective of our study is to investigate the influence of prompt engineering on students' retrieval of information and knowledge in flipped learning. It aims to address two specific questions:

- Q1: Does mastering prompt engineering methods help improve the quality of information students obtain from ChatGPT?
- Q2: How can the content of prompt engineering be effectively arranged in teaching to enhance the quality and efficiency of flipped classroom instruction?

By incorporating generative AI into teaching methods, educators can assist students in developing future-proof skills that enable them to excel in the job market and adeptly adapt to emerging challenges. For students who have grown up in an education system centered on skill acquisition, there can be a significant challenge in reconciling the skills they have acquired during their studies with those demanded by the future job market. For example, the advent of the first industrial revolution brought about disruption in light industry, leading to the replacement of manual labors with machines. Similarly, generative AI technologies such as ChatGPT have the potential to replace a substantial number of jobs that only require fundamental skills. During the era of mechanization, the appropriate response was to acquire knowledge of mechanics and become adept at operating machinery. This same logic seems to apply today with the wide use of generative AI. Instead of shying away from this technology, individuals should strive to learn and become proficient in using generative AI systems. Hence, conducting research on the impact of generative AI on current teaching methods and exploring effective ways to use generative AI to enhance teaching is timely and can make a significant contribution to this field of study.

II. Related Work

Artificial Intelligence Generated Content (AIGC) encompasses content generated through deep learning models like GPT (Generative Pre-trained Transformer). These advanced technologies exhibit remarkable proficiency in processing diverse data formats, including natural language, images, audio, and video. By harnessing a variety of multi-modal data sources, including tutorial videos, academic papers, and other reliable information, AIGC holds the promise of achieving substantial progress in the field of education [1]. The incorporation of these diverse data sources can greatly enhance the personalized educational experience. Google Research introduced Minerva [29], an advanced model that builds upon the PaLM [30] while incorporating a science-and-math-focused dataset. The proposed approach attains state-of-the-art performance in reasoning tasks by employing a combination of innovative techniques, such as few-shot prompting, majority voting, and chain of thought (scratchpad prompting) [1].

A. *ChatGPT*

ChatGPT (Chat Generative Pre-trained Transformer), developed by OpenAI, is an advanced conversational chatbot powered by the GPT-3 language model [31]. It produces text that closely resembles human language in response to prompts and engages in open-ended conversations [7]. Its training uses a 'prompt-response' dialogue structure, incorporating Reinforcement Learning with a human-in-the-loop approach. This approach involves gathering feedback from humans, who rank the model's responses, allowing for fine-tuning through proximal policy optimization. ChatGPT demonstrates exceptional capabilities such as answering follow-up questions, acknowledging errors, challenging incorrect assumptions, and refusing inappropriate queries. Despite receiving subpar grades, ChatGPT demonstrates the potential to acquire a university degree [32]. ChatGPT has attracted an impressive 100 million active users in a mere two months since its inception [33].

The ChatGPT model employs the Transformer architecture, leveraging vast corpora during training to facilitate a thorough understanding and generation of human language. The technical specifications of the ChatGPT model include the following.

- Transformer Architecture: ChatGPT is built upon the powerful foundation of the Transformer architecture, serving as the backbone for numerous state-of-the-art models [1], such as GPT-3 [24], DALL-E-2 [34], Codex [35], and Gopher [36]. The Transformer is a deep neural network architecture that revolutionized ”sequence-to-sequence” tasks by introducing self-attention mechanisms. It comprises an encoder-decoder structure [37], where the encoder transforms the input sequence into intermediate hidden representation vectors. These vectors are then processed by the decoder to generate the target sequence. By leveraging this architecture, ChatGPT gains a deep understanding of the contextual meaning in natural language, allowing it to produce coherent and contextually relevant responses.
- Self-Attention Mechanism: The self-attention mechanism plays a crucial role in the Transformer model [1]. It enables ChatGPT to focus on specific elements of the input during processing and assign weights to them according to their significance. By employing this mechanism, ChatGPT can effectively capture long-range dependencies in language and incorporate contextual information into its generated responses.
- Large-scale Pretraining: Through pretraining on a vast corpus of text, ChatGPT acquires language understanding capabilities. During the pre-training phase, ChatGPT is trained to anticipate the succeeding word or phrase in a given context [38], which helps it grasp statistical patterns and language structures. This contributes to its comprehension of language concepts, grammar, and semantics. Through this pretraining process, ChatGPT gains extensive linguistic knowledge, empowering it to generate responses using flexible language reasoning.
- Fine-tuning and Response Generation: After pre-training, ChatGPT enhances its performance through fine-tuning on specific tasks. For dialogue generation tasks, ChatGPT uses a supervised learning approach, where human experts provide example dialogue data to fine-tune the model. This process enables ChatGPT to generate responses that are contextually relevant and meaningful.
- Interactive Response and Model Iteration: Through user interactions, ChatGPT undergoes continuous improvement. When users provide input, ChatGPT considers it as context and generates an appropriate response. The interactive feedback can be used to refine the model, improving the quality and accuracy of the generated replies. Through this iterative process, ChatGPT can steadily optimize its conversational generation capabilities.

ChatGPT is a powerful automated conversational system based on the Transformer architecture, specifically designed for natural language processing. It achieves dialogue generation through a combination of large-scale pretraining and fine-tuning. By using self-attention mechanisms, ChatGPT effectively captures contextual information and generates meaningful responses based on user input. Through iterative refinement and interactive learning, ChatGPT continuously enhances its dialogue generation capabilities.

Despite being a powerful natural language processing model, ChatGPT does possess certain notable limitations.

- Lack of Common Sense and Deep Understanding [39]: Despite extensive pre-training and possessing a broad range of linguistic knowledge, ChatGPT still lacks an accurate grasp of real-world common sense and deep understanding. In some situations, it may generate responses that seem reasonable but are actually incorrect or absurd. Bian et al. [40] highlight that while GPT exhibits the capacity to effectively generate common knowledge by using ”Prompt” and displays competence in responding to general knowledge inquiries, it faces challenges when it comes to addressing specific types of knowledge and lacks the ability to precisely identify the common knowledge needed to answer specific questions.
- Limitations of Controlling Generated Content: The control over the generated responses of ChatGPT is relatively limited due to its foundation in large-scale unsupervised pre-training and supervised fine-tuning [41]. ChatGPT may generate inappropriate, offensive, or inaccurate replies, particularly when faced with sensitive topics or culturally-diverse scenarios. The challenge persists in ensuring that ChatGPT generates responses that align with user expectations. Dave et al. [42] highlighted the inherent lack of controllability in ChatGPT’s output, which raises concerns regarding its applicability in the healthcare domain. Specifically, they identified potential issues such as copyright violations, inadequate handling of complex medical legalities, and a failure to meet the growing demand for transparency in AI-generated content.
- Contextual Over-sensitivity: ChatGPT’s context handling can be excessively sensitive, leading to conservative and repetitive conversation generation. It often produces replies that resemble previous dialogue, lacking innovation and diversity. Consequently, this diminishes the user experience by resulting in less fluid and varied conversations.
- Easily misleading: The training of ChatGPT involves a vast corpus of internet text, encompassing information from diverse sources that may contain errors, biases, and inaccuracies. Consequently, the model may demonstrate biases, misunderstandings, or disseminate incorrect information in its responses. To prevent potential misinformation or the misleading of users, it is crucial to approach the model’s generated responses with caution and diligently verify them. The use of ChatGPT warrants caution due to its potential to mislead both authors and readers. The model has been observed to produce wrong facts, generate references that do not exist, and exhibit a

tendency to staunchly and persuasively support assertions that may be untrue. Consequently, it is crucial to employ this tool ethically and prioritize its application for the betterment of humanity [43].

- Vulnerabilities in Adversarial Attacks: ChatGPT and other deep learning models exhibit vulnerability to adversarial attacks, wherein malicious users manipulate the model by providing specific inputs that lead to the generation of deceptive, harmful, or inappropriate responses. Liu et al. [44] highlighted that users have the ability to launch adversarial attacks on ChatGPT, leading the model to be influenced by specific conversational contexts. This manipulation involves introducing irrelevant premises into the generated content, potentially resulting in inappropriate responses. Although certain defense measures have been implemented to mitigate these attacks, the challenge of addressing adversarial attacks remains an ongoing research area that requires further investigation and resolution.

In addition to the aforementioned limitations, according to the research conducted by Lo et al. [23], ChatGPT exhibits significant variability in its performance across various domains. This performance spectrum encompasses levels ranging from highly satisfactory to acceptable, mediocre, and even poor. Notably, in domains such as economics [45] and programming [46], ChatGPT's performance has been consistently satisfactory. Moreover, it demonstrates commendable proficiency in English comprehension. However, when applied in fields such as law [32], [47] and medical education [48], [49], its performance is about average. In contrast, ChatGPT's performance in mathematics [50], sports science [51], and software testing [52] is notably poor.

B. Teaching and Learning with ChatGPT

ChatGPT is steadily gaining prominence in the education field. Leveraging its robust natural language processing capabilities and adept conversational generation skills, ChatGPT holds significant potential in education and learning support [23]. ChatGPT engages with students and educators, offering customized educational support, answering inquiries, and facilitating innovative learning experiences.

In teaching and learning, ChatGPT is mainly used in four areas: instructional preparation (e.g., generating course materials, providing suggestions, and conducting language translation), instructional assessment (e.g., generating assessment tasks and evaluating student performance), instructional support (e.g., assisting students with practice), and instructional enhancement (e.g., improving the effectiveness of existing teaching methods).

In terms of instructional preparation, ChatGPT can provide assistance and suggestions for instructors. Study has shown that ChatGPT is a valuable tool for educators, helping them identify essential curriculum content, and providing an outline [53]. Megahed et al. [46] requested ChatGPT to prepare a course outline tailored to an undergraduate statistics course, noting that these instructional suggestions require minimal modifications. According to Zhai [54], ChatGPT provides valuable suggestions in the context of special education, which he described as being particularly advantageous for students with unique learning requirements.

For instructional assessment, instructors can use ChatGPT to craft scenarios for interactive learning, assessments, and student evaluations [55]. As an illustration, Han et al. [56] directed ChatGPT to generate multiple-choice questions for medical topics, including illustrations and experimental values. Nonetheless, Al-Worafi et al. [57] advised only using ChatGPT as a supplementary tool for assessment preparation, cautioning that it may not cover all intended learning objectives and should not replace instructors or human tutors.

For instructional support, Topsakal et al. [58] employed ChatGPT to facilitate English language learning for students through interactive dialogues. Once the accuracy of the materials has been verified, instructors can use ChatGPT to adapt them for tools like Google Dialogflow [1]. This enables the provision of interactive and personalized learning environments for students.

As to instructional enhancement, ChatGPT has the potential to improve active learning strategies. For instance, Rudolph et al. [59] propose the use of a flipped classroom approach, wherein students are expected to study pre-class materials in preparation for the class. Nevertheless, students in traditional flipped classrooms often face challenges in pre-class learning [60], and there is a need to improve classroom engagement [13]. The COVID-19 pandemic has exacerbated this issue, as fully online flipped learning has resulted in diminished classroom engagement and decreased interaction among students [61], [62]. Assuming the role of a virtual tutor, ChatGPT assists students in online independent learning by addressing their queries [63], while also bolstering group dynamics through the provision of discussion structures and real-time feedback [64].

While ChatGPT has achieved considerable success, its implementation in education has brought forth a range of new challenges and potential threats. A particular concern arises from ChatGPT's potential to enable AI-assisted cheating, as it allows students to "substitute" themselves during exams and written assignments. According to Susnjak [65], the analytical thinking prowess of ChatGPT, coupled with its ability to produce compelling text with minimal guidance, raises credibility concerns. This is particularly significant given their prevalence in higher education. Besides, the text generated by ChatGPT presents challenges for conventional plagiarism detection tools. Ventayen [66] conducted a study wherein ChatGPT authored an article by drawing from existing publications. Additionally, Khalil and Er [67] showcased the exceptional content generation capability of ChatGPT by creating 50 articles that yielded average similarity scores of 13.72% and 8.76% when assessed respectively by Turnitin and iThenticate (two popular plagiarism detection applications [23]). The recent study revealed that peer reviewers could only identify 63% of fraudulent abstracts produced by ChatGPT, raising significant concerns about AI-driven text in scientific literature [68].

[1] Available online: https://cloud.google.com/dialogflow.

Furthermore, in a comprehensive review of 60 articles encompassing various academic disciplines, Sallam [69] identified the challenges linked to the use of ChatGPT in education, specifically in relation to accuracy and reliability. The research conducted by Mbakwe et al. [70] highlights that ChatGPT's training on vast data corpora makes it vulnerable to biases and the potential incorporation of inaccurate information. Such biases may originate from research predominantly conducted in high-income countries or from textbooks that lack relevance to diverse regions. Additionally, it should be noted that ChatGPT's knowledge is based on data available only until 2021 [49], [55], [71], rendering its responses on professional topics and current events susceptible to inaccuracy and unreliability. This limitation becomes particularly worrisome when considering ChatGPT's potential to generate incorrect or false information [46], [71], [72], posing a significant risk for students who heavily depend on ChatGPT as a source of information. Due to such concerns, some schools have already banned the use of ChatGPT on campus [23]. However, as the saying goes, we should not throw out the baby with the bathwater. As Mhlanga's research indicates, the impact of using ChatGPT for educational purposes needs immediate attention to ensure maximum use of its advantages while minimizing its disadvantages [73].

Researchers have conducted thorough investigations into the potential issues associated with integrating ChatGPT into education. They have also put forth a range of strategies to tackle these concerns, which can be categorized into three primary aspects: task design, AI writing detection, and institutional policies.

In terms of task design, researchers have explored various approaches. Zhai [74] proposed the exploration of innovative formats to foster students' creative and critical thinking. Choi et al. [32] highlighted the significance of demanding students to analyze cases instead of merely recalling knowledge. Geerling et al. [45] proposed that students should be tasked with applying the concepts learned in the course, including the creation of non-replicable materials for artificial intelligence. According to Stutz et al. [75], future assessments should prioritize higher levels of Bloom's taxonomy, including application, analysis, and creation.

In terms of AI writing detection and institutional policies, Szabo [51] reported that while conventional plagiarism detection tools may not detect text produced by ChatGPT, AI detectors can still identify such content. Additionally, generating accurate reference lists could be a challenge for ChatGPT, and this could serve as a vital indicator for instructors to identify student usage of ChatGPT [31], [72], [76]. In addition to identifying students' plagiarism behavior, researchers highlight the significance of implementing anti-plagiarism guidelines and providing education on academic integrity [31], [59], [67].

While ChatGPT holds immense potential for assisting instructors in tasks like generating course materials, providing suggestions, and serving as a virtual tutor for students through answering questions and facilitating collaboration, it also presents challenges in the form of generating inaccurate or fabricated information and evading plagiarism detectors. Promptly adapting teaching approaches and institutional policies in educational institutions is essential to tackle these challenges. Moreover, instructor training and student education play a vital role in effectively responding to the challenges posed by ChatGPT in the educational landscape [23].

The attitude towards ChatGPT should be to retain its essence while eliminating its shortcomings. Efforts should be made to enhance its ability to accelerate the learning process and improve teaching effectiveness, while also minimizing the harm caused by errors and irrelevant information [77]. The effectiveness of integrating ChatGPT into educational activities hinges on assisting students in obtaining accurate information through its use. In the application of ChatGPT, prompt engineering is extensively employed to guide ChatGPT in avoiding the delivery of irrelevant or erroneous information.

C. Prompt Engineering

Prompt engineering, the art of skillful prompt construction for Large Language Models (LLMs) like ChatGPT, is essential in guiding the generation of desired responses [78]. Prompts primarily facilitate communication between users and ChatGPT. Prompts provide guidance to ensure that ChatGPT generates responses aligned with the user's intent. As a result, well-engineered prompts greatly improve the efficacy and appropriateness of ChatGPT's responses.

Effective prompt engineering is crucial for optimizing models as it involves designing, optimizing, and refining prompts to accurately convey the user's intent to ChatGPT [79]. Prompt engineering plays a vital role in bridging the gap between user intent and the models' understanding, thereby significantly impacting the quality of generated replies. Thus, it becomes essential for users to master prompt engineering in order to fully leverage ChatGPT's potential and achieve optimal results in various applications, considering the direct influence of prompt quality on generated replies.

With the continuous improvement of language models, mastering prompt engineering has become crucial for users to fully unleash the potential of ChatGPT and achieve optimal results in various applications [78]. Numerous studies have explored the impact of prompt engineering on AI generative models across domains, including image generation tasks [80] and classification tasks [81], highlighting the critical role of excellent prompt design for large language models like ChatGPT [82], [83]. To guide models in generating effective natural language prompts, Reynolds et al. [84] introduced the concept of meta-prompts. Additionally, Lo et al. [85] proposed the CLEAR framework, consisting of five fundamental principles that enhance interaction with AI language models and facilitate more effective evaluation and content creation. Furthermore, White et al. presented a prompt engineering catalog, akin to software patterns, offering reusable solutions to challenges faced when interacting with large language models [86]. As suggested by Giray et al., academic writers can adapt to the ever-changing landscape by acquiring expertise in prompt engineering and harnessing the power of large-scale language models [87]. To enhance the effectiveness of using ChatGPT, it is important to incorporate prompt engineering into educational activities, empowering both instructors and students.

III. RESEARCH METHODOLOGY

In this study, we address the research questions (RQ1, RQ2) through a quasi-experimental research design, which incorporates field quasi-experiments and analysis. This design is particularly suitable for educational research due to the impracticality of implementing fully compliant random group assignments, as highlighted by Fraenkel et al. [88]. Although this study focused on 4-year college students who shared the same major and grade level, pre-experimental tests were conducted on different groups to evaluate predetermined characteristics including age, and ChatGPT questioning skills. Gender is culturally complex and is excluded from the study. Age is known to affect students' cognitive abilities and questioning skills, and is considered the main influential factor. Following the research design, researchers randomly assigned a portion of the study participants to the experimental group, which received prompt engineering knowledge, while the remaining individuals were designated as the control group without such knowledge.

A. Research Setting

The study took place during the spring semester of 2023 at a pre-service teacher education university located in a province in northeastern China. The experimental and control groups consisted of freshmen from the same class who were majoring in the same subject. By implementing the program in this particular setting, three significant advantages were observed.

- Firstly, both the experimental and control groups were instructed by the same instructor and use comparable learning materials, ensuring a relatively similar environment for conducting a field quasi-experimental design. This consistency enhances the reliability of the study.
- Secondly, by selecting freshmen from the same major, potential differences arising from varying majors or elective courses were minimized to a significant extent. This minimization of confounding variables strengthens the validity of the findings.
- Additionally, this research design allows for the implementation of different types of tests as necessary. The A/B testing [89], also know as split testing, is used when comparing the impact of a specific factor across different groups, while pre-post testing is conducted when examining before-and-after changes within the same group. This comprehensive approach enables the evaluation of overall effects through small sample sizes and facilitates a more accurate assessment of the observed changes.

B. Sampling Approach

Participants who expressed their willingness to participate and contribute to the study were selected using a simple random sampling method. Subsequently, they were randomly assigned to either the experimental or control group. From an initial sample of 26 freshmen majoring in a specific course, the experimental and control groups were formed, each consisting of 13 students. However, two students withdrew from the experiment due to personal reasons, resulting in a final participation of 24 students. Table I presents the age, and ChatGPT Questioning Skills (CQS) of the students involved in the study, with a particular emphasis on demonstrating that there were no significant differences among the different groups in these three aspects at the initial stage. Further details on the measurement of CQS will be provided in subsequent sections. For ease of interpretation, the scores of CQS in this study are scaled from 0 to 4.

The experimental group has an average age of 18.9 years, while the control group has an average age of 18.6 years. The average CQS score of the experimental group was 0.5167, while the control group's average CQS score was 0.6363. The statistical significance level was established at $p < 0.05$. T-tests showed no significant differences between groups in terms of age (p = 0.2798), or the scores of CQS (p = 0.6025). The experimental results demonstrated no significant differences between the experimental group and the control group, indicating that the conditions of the two groups of students were similar.

TABLE I
THE CHARACTERISTICS OF THE STUDENTS, WHO WERE SIMILAR IN TERMS OF AGE, AND CHATGPT QUESTIONING SKILLS. (P-VALUE < 0.05)

	experimental group	control group	p-value
Age	18.9	18.6	0.2798
ChatGPT Questioning Skills (CQS)	0.5167	0.6363	0.6025

C. Design Phase

The research program on integrating ChatGPT into the curriculum was a collaborative effort between researchers and a computer science instructor who teaches *Fundamentals of Computer Networks*. The objective of this innovative program is to foster students' scientific inquiry skills development, including the ability to identify problems effectively, use ChatGPT for information gathering and organization and complete coding tasks. In this course about computer networks, the instructor introduced Transmission Control Protocol (TCP) sockets programming as a flipped classroom task. This approach was designed to offer students practical coding experience and nurture their scientific inquiry skills.

The experiment followed a specific sequence of steps. Initially, each student had no prior knowledge of TCP socket programming and could only ask ChatGPT up to 5 questions based on their own understanding of the task requirements. Subsequently, both the experimental and control groups received instruction on TCP socket-related knowledge. Following the instruction, students in both groups were required to ask ChatGPT up to 5 questions to obtain answers that would assist them in completing their tasks. Furthermore, the instructor introduced the relevant knowledge of the prompt engineering project specifically to the experimental group. The students in this group were taught popular prompt engineering methods and their questions were recorded. By contrast, the control group only received an introduction to the basic concepts of the prompt engineering project, without any instruction on prompt engineering methods, and their questions were also recorded.

To control ChatGPT's influence on students' learning outcomes, the program designed an isolation questioning method, that is, requiring students to ask ChatGPT questions based on the task. The instructor recorded the questions and entered them into ChatGPT to obtain answers. In practice, students were asked to pose questions twice: once during a pretest when they had no prior knowledge of the topic, and again after gaining an understanding of TCP socket programming. We used basic prompts to improve students' questions, simulating the types of questions that students familiar with basic prompt engineering would ask ChatGPT. Additionally, we used a popular prompt engineering framework (CRISPE) [28] to enhance students' questions, simulating the questions that students proficient in prompt engineering would pose to ChatGPT.

The CRISPE framework proposed by shieh [28] is new but considered an excellent template for writing prompts. CRISPE stands for the following:

- CR: Capacity and Role. What role do you want ChatGPT to play?
- I: Insight, what background information and context do you want ChatGPT to provide?
- S: Statement. What do you want ChatGPT to do?
- P: Personality. In what style or manner do you want ChatGPT to answer you?
- E: Experiment. Ask ChatGPT to provide multiple answers for you.

Prompts generated by following this framework can elicit more complete and in-depth answers comparing to the free-style unstructured prompts.

This framework-guided approach offers the advantage of addressing potential negative impacts of new technology on students' learning. By requiring students to formulate task-specific questions for ChatGPT, it ensures that they actively engage with the technology. In addition, while incorporating this technological aspect, the teaching approach in this experiment followed the flipped classroom model, so as to align with established pedagogical practices.

D. Implementation Phase

Before delving into network layer concepts, the students were assigned the task of learning TCP socket programming with Python. They were required to write code for both the server and client sides, adhering to the client/server model. The task involved the client connecting to the server and sending a string. The server was expected to convert the client's string to uppercase and return it to the client.

The instructor delivered a comprehensive introduction to TCP socket programming and client/server mode communication to all students in this study. In a 45-minute session, the students were provided with a detailed explanation of the assignment they would work on. The assignment required the students to research and apply various programming concepts related to TCP sockets, including the TCP protocol, socket programming, IP, ports, and more. Their objective was to use these concepts to ask questions to ChatGPT.

As mentioned earlier, in order to avoid the potential novelty influence of new technology on students, this experiment proposes an isolated questioning method. The specific implementation details of this method are as follows: after the instructor assigned the task, he introduced ChatGPT to the students and instruct them on how to use it. For example, informing students that they can obtain relevant knowledge needed to complete the task by asking questions to ChatGPT. In order to complete TCP socket programming, students were required to design 5 questions and acquire as much high-quality task-related information as possible through these questions.

Our research team manually annotated the answers obtained from students' questioning. The quality of the answers provided by ChatGPT, as measured by CQS (ChatGPT Questioning Skills) scores, was evaluated by experts in AIGC. The evaluation of the answers focused on four dimensions of general intelligence standards, as outlined by Paul [90]:

- Relevance: Is the response provided by ChatGPT relevant to the task? Does it assess the complexity and comprehensiveness of addressing the task?
- Clarity: Is the response from ChatGPT clear, appropriately organized, and logically coherent? Does it employ suitable terminology and diction for the users?
- Accuracy: Is ChatGPT's response accurate? Does it contain any errors in knowledge?
- Precision: Is the answer provided by ChatGPT specific and detailed enough? Is it precise and unambiguous?

The CQS score of a question varied based on the relevance of the answer to TCP socket programming, increasing with relevance and decreasing with lack of relevance.

In order to mitigate the potential influence of new technology on students, this experiment adopted an isolated questioning method. The implementation involved the instructor assigning the task and introducing ChatGPT to the students, providing instructions on its usage. The students were informed that they could acquire the necessary knowledge to complete the TCP socket programming assignment by asking questions to ChatGPT. Each student was required to design five high-quality, task-related questions to obtain relevant information. To evaluate the relevance of the obtained answers, experts were invited to manually annotate them and assign scores to each question based on its alignment with TCP socket programming.

The instructor employed prompt engineering methods to enhance the quality of students' questions, while AIGC experts from our research team assessed the answers generated by ChatGPT. In an ideal scenario, students acquire the techniques of prompt engineering, enabling them to refine their own questions. Consequently, the quality of the answers they obtain is expected to be comparable to those derived from the instructor's improved questions. The isolated questioning method offers two distinct advantages. First, it eliminates concerns about the potential impact of new technology on the teaching process for students. Second, it facilitates the investigation of prompt engineering's influence on students' use of ChatGPT in flipped classroom tasks.

Following is a detailed breakdown of the aforementioned experimental steps:

- 1. All students in both the experimental group (A1) and the control group (A2), without prior knowledge related to the task, asked ChatGPT 5 questions each to obtain information for completing the task. The research team recorded the answers they obtained and their corresponding scores were recorded.
- 2.The researcher added a simple prompt, restricting the scope of the answer to Python's TCP socket programming, for questions in A1 and A2. ChatGPT's answers and scores were recorded for the experimental group (B1) and the control group (B2).
- 3. After the instructor taught the students the knowledge related to the task, all students in both groups (C1 for the experimental and C2 for the control) asked ChatGPT 5 questions. The questions, ChatGPT's answers and their scores were recorded.
- 4. The research team improved the questions in A1 and A2 using the CRISPE framework. We then input the improved questions into ChatGPT, and recorded the answers and scores of the experimental group as D1, and those of the control group as D2.
- 5. Similarly, the questions in C1 and C2 were improved by using the CRISPE framework. The improved questions were input into ChatGPT, and the answers and scores of the experimental group were recorded as E1, and the control group's answers and scores were recorded as E2.

E. Data Gathering and Analysis Approach

We assessed the impact of prompt engineering on students' ability to find information using ChatGPT by conducting a statistical analysis of the answer qualities. The answers were ranked for relevance on a 5-point scale, with scores ranging from 0 to 4, where higher scores indicated greater helpfulness for students to complete their course project(the task). Subsequently, the mean and variance of scores were calculated for each group. We analyzed the quality of information acquired by the students to serve two purposes: a) to determine the impact of task-related knowledge on the knowledge obtained using ChatGPT, and b) to assess the potential improvement in knowledge quality through the prompt engineering method.

For data analysis, researchers used a t-test to compare the scores of the answers and evaluate the significant impact of prompt engineering on students' use of ChatGPT in this particular scenario of learning. T test is a widely used statistical method that calculates t-values and p-values to determine significant differences between the means of two groups. For this analysis, an independent samples t-test was used to compare the means of the experimental and control groups, whereas a paired samples t-test was employed to examine the difference between pre-test and post-test results within the same group.

IV. Findings and Discussions

Table II presents the means and standard deviations (SD) of ChatGPT Questioning Skills (CQS) across all groups. Table III presents the significance of the differences in CQS means between the group conditions in their CQS. When comparing the scores of the experimental group and the control group at each stage (A1 vs A2, B1 vs B2, C1 vs C2, D1 vs D2, E1 vs E2) through inferential testing, the results revealed that the p-value for each group exceeded the threshold of alpha (level of significance at 0.05). This suggests that there were no significant differences in the quality of answers obtained from questions at the same stage.

TABLE II
Displays the means, standard deviations of CQS, and p-value (bold if less than 0.05) of all groups.

	A1	B1	C1	D1	E1
Mean	0.5166	1.0333	0.75	1.3666	1.833
Std	0.5078	0.4658	0.6274	0.3284	0.4811
	A2	B2	C2	D2	E2
Mean	0.6363	0.9636	0.8	1.3	1.4909
Std	0.5714	0.4272	0.5059	0.4546	0.4229

A. Quantitative Results for RQ1

This section presents the quantitative results aimed at addressing the first research question, RQ1: Does mastering prompt engineering methods help improve the quality of information students obtain from ChatGPT?

1) Quasi-Experiment 1: We used a comprehensive analysis to explore the effectiveness of the CRISPE framework in improving the quality of answers obtained from ChatGPT. Comparing the CQS scores between the C1 and E2 groups, as well as the C2 and E1 groups (please refer to Table IV), we observe no statistically significant difference in the mean CQS scores between the C1 and C2 groups. However, the E2 group shows a remarkable improvement of 0.7409 in the CQS scores compared to the C1 group, with a P-value of 0.0033, much lower than alpha at 0.05. Similarly, the E1 group demonstrates an impressive improvement of 1.033 compared to the C2 group, with a P-value of 7.555e-7.

Furthermore, when comparing the C1 group with the E1 group and the C2 group with the E2 group (please refer to Table IV), we found that the E1 (E2) group exhibits a significant improvement of 1.083 (0.6909) in the average CQS score compared to the C1 (C2) group, with P-values of 0.0007 (2.108e-7).

These findings provide compelling evidence supporting the effectiveness of the CRISPE framework in enhancing the quality of answers obtained from ChatGPT. The evidence is derived from both the pre-test and post-test results of the same students over time, as well as the A/B testing conducted between different student groups.

2) Quasi-Experiment 2: The effectiveness of the simple prompting method in improving the quality of answers from ChatGPT is assessed by comparing the differences in the quality of answers between the A1 and B1 groups, as well as the A2 and B2 groups (please refer to Table V). Upon analyzing the gain scores, the average CQS score obtained from ChatGPT by the B1 group was significantly higher than that of the A1 group, with an improvement of 0.5167 and a

TABLE III
THE SIGNIFICANCE OF THE DIFFERENCES BETWEEN THE GROUPS (P-VALUES) WAS DISPLAYED, WITH VALUES LESS THAN 0.05 HIGHLIGHTED IN BOLD.

P-value	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2
A1	-	0.6025	**0.0003**	**0.0325**	0.1158	0.7282	**0.0002**	**0.0011**	**1.268e-5**	**5.934e-5**
A2		-	0.0848	0.0478	0.6541	0.6446	**0.0019**	**0.0068**	**2.863e-5**	**0.0001**
B1			-	0.7119	0.0757	**0.0178**	**0.0153**	0.191	**0.0002**	**0.0222**
B2				-	0.3482	**0.0068**	**0.0209**	**0.0034**	**0.0002**	**0.0017**
C1					-	0.4405	**0.0074**	**0.0276**	**0.0007**	**0.0033**
C2						-	**3.213e-5**	**0.0005**	**7.555e-7**	**2.108e-7**
D1							-	0.7037	**0.0079**	0.444
D2								-	**0.0149**	0.2585
E1									-	0.0837

TABLE IV
THE P-VALUE (BOLD IF LESS THAN 0.05) OF COMPARING THE CQS SCORES OF GROUPS WITH CRISPE OR NOT.

	C1,E2	C2,E1	C1,E1	C2,E2
Improvement	0.7409	1.033	1.083	0.6909
P-value	**0.0033**	**7.555e-7**	**0.0007**	**2.108e-7**

P-value of 0.0003. However, the improvement in the B2 group compared to the A2 group was not statistically significant, with a P-value of 0.0748.

TABLE V
THE P-VALUE (BOLD IF LESS THAN 0.05) OF COMPARING THE CQS SCORES OF GROUPS WITH SIMPLE PROMPTING METHOD OR NOT.

	A1,B1	A2,B2
Improvement	0.5167	0.3273
P-value	**0.0003**	0.0748

It is important to note that based on the current sample data, there is insufficient evidence to support the conclusion that the observed differences are solely attributable to the applied influencing factors. Therefore, we cannot definitively conclude that the implementation of the simple prompting method significantly improves the CQS scores. Further investigation and analysis with a larger sample size are required to draw more definite conclusions.

3) Quasi-Experiment 3: A comparison between the quality of answers obtained in the B1 and D1 groups, as well as the B2 and D2 groups, highlights the effectiveness of CRISPE over the simple prompting method (please refer to Table VI). Analyzing the gain scores reveals that the information obtained from ChatGPT by the D1 and D2 groups was significantly superior to that obtained by the B1 and B2 groups (with improvements of 0.3333 and 0.3364, respectively). Furthermore, a statistically significant difference was observed when comparing the simple prompting method to CRISPE, with respective P-values of 0.0153 and 0.0033.

TABLE VI
THE P-VALUE (BOLD IF LESS THAN 0.05) OF COMPARING THE CQS SCORES OF GROUPS WITH THE SIMPLE PROMPTING METHOD OR CRISPE.

	B1,D1	B2,D2
Improvement	0.3333	0.3364
P-value	**0.0153**	**0.0034**

As this study reveals, the implementation of a sound framework such as the CRISPE framework, in comparison to unstructured prompts, enabled students to acquire higher-quality information from ChatGPT in flipped classroom tasks. Consequently, the CRISPE method leads to a statistically significant improvement in the quality of answers obtained by students from ChatGPT. The significant impact of the prompting method based on the CRISPE framework suggests that mastering prompt engineering can enhance the efficiency and effectiveness of using ChatGPT to acquire knowledge in certain learning settings, ultimately improving students' learning outcomes.

B. Quantitative Results for RQ2

This section presents quantitative results aimed at answering the second research question, RQ2: How can the content of prompt engineering be arranged in teaching to improve the quality and efficiency of flipped classroom teaching?

1) Quasi-Experiment 4: To investigate whether task-related prior knowledge can enhance the quality of answers obtained from ChatGPT, a comparison was made between the answers obtained in A1 and C1, as well as A2 and C2. The comparison of gain CQS scores revealed that the information quality from ChatGPT in groups C1/C2 was higher compared to groups A1/A2 (0.2334/0.1637). Nonetheless, inferential statistical analysis showed that the difference between the two groups was not significant (P-value: 0.1158/0.6446). The absence of a significant difference suggests that comprehending task-related information does not significantly impact the quality of answers obtained from ChatGPT for students. The findings suggest that task-related knowledge does not significantly enhance the quality of information obtained by students from ChatGPT. Therefore, in order to receive high-quality task-related answers from ChatGPT, it is necessary to explore alternative methods that can improve the efficiency of using the model.

TABLE VII
THE P-VALUE (BOLD IF LESS THAN 0.05) OF COMPARING THE CQS SCORES OF GROUPS WITH TASK-RELATED PRIOR KNOWLEDGE OR NOT.

	A1,C1	A2,C2
Improvement	0.2334	0.1637
P-value	0.1158	0.6446

2) Quasi-Experiment 5: To investigate how to arrange prompt engineering methods during the learning process, we conducted a comparative analysis of the improvement in CQS scores between Groups D and E, as opposed to Group A (please refer to Table VIII). The comparison of gain CQS scores revealed that the information quality from ChatGPT in groups E1/E2 was higher compared to groups D1/D2 (0.4667/0.1909), with respective P-values of 0.0079 and 0.2585. The findings indicate that, in some cases, using the CRISPE framework after understanding task-related knowledge could effectively enhance the quality of information obtained by students when using ChatGPT.

TABLE VIII
THE P-VALUE (BOLD IF LESS THAN 0.05) OF COMPARING THE CQS SCORES OF GROUPS WITH AND WITHOUT THE COMBINATION OF CRISPE AND TASK-RELATED KNOWLEDGE.

	D1,E1	D2,E2
Improvement	0.4667	0.1909
P-value	**0.0079**	0.2585

In order to ensure that students effectively use ChatGPT to accomplish flipped classroom tasks with high quality, instructors should provide students with the necessary foundational knowledge related to the tasks during assignments and ensure that students are proficient in prompt engineering methods.

C. Discussion

1) Interpret the Results: Through a series of experiments, the findings indicate that mastering prompt engineering methods indeed contributes to improving the quality of information that students received from ChatGPT. Additionally, the findings suggest that, within the context of flipped classroom instructional activities, it is essential for educators to ensure that students possess solid foundational knowledge before incorporating prompt engineering elements. This approach leads to enhanced efficiency in using ChatGPT, consequently elevating the overall teaching quality in flipped classrooms.

In line with other research findings [78], [79], [82], [83], which prompt engineering plays a pivotal role in efficiently conveying users' intentions to ChatGPT, our study presents an exploration of this assertion within the realm of education, serving to enlighten educators on the use of prompt engineering in aiding students to harness ChatGPT's capabilities.

2) Limitations: This study exhibits certain limitations. First, it primarily focuses on a specific cohort of students, which may constrain the generalizability of the research findings to other academic disciplines or students with different educational backgrounds. Our choice of first-year students as the target group for this research was deliberate, as they have not undergone extensive specialized training, and their diverse academic backgrounds are still diverse and more representative of undergrad students in other universities. Such a choice of subjects helps mitigate constraints when extending the research findings to student populations in other disciplines and educational contexts. It is advisable for future studies to expand the scope of the research objectives to enhance research effectiveness.

Second, the study employed a method termed 'isolated questioning methods' to simulate students' mastery of prompt engineering. The assumption made in the study was that all participants could ideally master prompt engineering methods and without being distracted by the novelty of technology. However, it is important to acknowledge that this approach did not account for individual differences among students, which is another factor for future studies to explore.

3) Recommendations: The research findings indicate that in learning activities, prompt engineering has a positive impact on students' information retrieval quality when using ChatGPT. In this era of Artificial Intelligence (AI), the integration of AI tools into teaching activities is becoming increasingly inevitable. When it comes to the use of AI tools such as ChatGPT, we recommend that educators prioritize the mastery and use of prompt engineering techniques. We also encourage more educational research institutions and scholars to delve into the application of prompt engineering in education across various disciplines and age groups. This will contribute to a deeper understanding of the practical effects of prompt engineering on students' use of AI tools such as ChatGPT and the identification of best practices. Regarding the use of generative AI tools by students, we suggest that schools and educational institutions consider offering courses to instruct students on effective and standardized AI tool usage, thereby enhancing their learning achievements.

V. CONCLUSION

The primary objective of this study was to answer two research questions. First, the research team investigated the impact of prompt engineering on students' ability to obtain high-quality answers from ChatGPT in flipped classroom settings. The study employed multiple comparative quasi-experiments and the quantitative results indicate a statistically significant and positive influence of mastering prompt engineering on the effectiveness of information retrieval from ChatGPT. Second, we explored the optimal arrangement of prompt engineering content to enhance the quality and efficiency of flipped classroom teaching. The findings indicate that in order to maximize the positive impact of ChatGPT in similar learning settings, students should not only master prompt engineering techniques but also possess the prerequisite knowledge relevant to the assigned task. In conclusion, this study highlights the positive impact of prompt engineering on students' completion of tasks when using ChatGPT. In addition, this study underscores the changing role of instructors in this age of AI, from "sage on stage" to mentors and coaches. We recommended that generative AI be used under the guidance of instructors to effectively harness its positive impact. This study carries some limitations, such as using isolated questioning methods and simulated ideal conditions. It also assumes that students have already mastered prompt engineering. Nevertheless, the results and findings still offer valuable insights for furthering our understanding of the role of generative AI tools in assisting students with information acquisition.

In addition, this study may serve as a launch pad for TLT's upcoming special issue on Education in the World of ChatGPT and other Generative AI, which is scheduled to be published by June 2024. We hope researchers around the world will join this ongoing dialogue and provide their insights on the use of Generative AI in education and training.

REFERENCES

[1] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," *arXiv preprint arXiv:2303.04226*, 2023.

[2] E. A. Van Dis, J. Bollen, W. Zuidema, R. van Rooij, and C. L. Bockting, "Chatgpt: five priorities for research," *Nature*, vol. 614, no. 7947, pp. 224–226, 2023.

[3] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[4] Z. Du, Y. Qian, X. Liu, M. Ding, J. Qiu, Z. Yang, and J. Tang, "Glm: General language model pretraining with autoregressive blank infilling," *arXiv preprint arXiv:2103.10360*, 2021.

[5] L. Nicolescu and M. T. Tudorache, "Human-computer interaction in customer service: the experience with ai chatbots—a systematic literature review," *Electronics*, vol. 11, no. 10, p. 1579, 2022.

[6] L. Xu, L. Sanders, K. Li, J. C. Chow *et al.*, "Chatbot for health care and oncology applications using artificial intelligence and machine learning: systematic review," *JMIR cancer*, vol. 7, no. 4, p. e27850, 2021.

[7] Y. Lu, H. Wang, and W. Wei, "Machine learning for synthetic data generation: a review," *arXiv preprint arXiv:2302.04062*, 2023.

[8] P. Diwanji, K. Hinkelmann, and H. F. Witschel, "Enhance classroom preparation for flipped classroom using ai and analytics." in *ICEIS (1)*, 2018, pp. 477–483.

[9] G. Akçayır and M. Akçayır, "The flipped classroom: A review of its advantages and challenges," *Computers & Education*, vol. 126, pp. 334–345, 2018.

[10] J. Bergmann and A. Sams, *Flip your classroom: Reach every student in every class every day*. International society for technology in education, 2012.

[11] B. Sohrabi and H. Iraj, "Implementing flipped classroom using digital media: A comparison of two demographically different groups perceptions," *Computers in Human Behavior*, vol. 60, pp. 514–524, 2016.

[12] L. Cheng, A. D. Ritzhaupt, and P. Antonenko, "Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis," *Educational Technology Research and Development*, vol. 67, pp. 793–824, 2019.

[13] K. F. Hew, S. Bai, P. Dawson, and C. K. Lo, "Meta-analyses of flipped classroom studies: A review of methodology," *Educational Research Review*, vol. 33, p. 100393, 2021.

[14] D. E. Gonda and B. Chu, "Chatbot as a learning resource? creating conversational bots as a supplement for teaching assistant training course," in *2019 IEEE International Conference on Engineering, Technology and Education (TALE)*. IEEE, 2019, pp. 1–5.

[15] W. Huang, K. F. Hew, and D. E. Gonda, "Designing and evaluating three chatbot-enhanced activities for a flipped graduate course," *International Journal of Mechanical Engineering and Robotics Research*, 2019.

[16] T. Ito, M. S. Tanaka, M. Shin, K. Miyazaki *et al.*, "The online pbl (project-based learning) education system using ai (artificial intelligence)," in *DS 110: Proceedings of the 23rd International Conference on Engineering and Product Design Education (E&PDE 2021), VIA Design, VIA University in Herning, Denmark. 9th-10th September 2021*, 2021.

[17] J. Li, L. Ling, and C. W. Tan, "Blending peer instruction with just-in-time teaching: jointly optimal task scheduling with feedback for classroom flipping," in *Proceedings of the Eighth ACM Conference on Learning@ Scale*, 2021, pp. 117–126.

[18] A. Varnavsky, "Chatbot to increase the effectiveness of the <<flipped classroom >>technology," in *2022 2nd International Conference on Technology Enhanced Learning in Higher Education (TELE)*. IEEE, 2022, pp. 289–293.

[19] K. F. Hew, W. Huang, J. Du, and C. Jia, "Using chatbots in flipped learning online sessions: perceived usefulness and ease of use," in *Blended Learning: Re-thinking and Re-defining the Learning Process. 14th International Conference, ICBL 2021, Nagoya, Japan, August 10–13, 2021, Proceedings 14*. Springer, 2021, pp. 164–175.

[20] ——, "Using chatbots to support student goal setting and social presence in fully online activities: learner engagement and perceptions," *Journal of Computing in Higher Education*, vol. 35, no. 1, pp. 40–68, 2023.

[21] S. Wollny, J. Schneider, D. Di Mitri, J. Weidlich, M. Rittberger, and H. Drachsler, "Are we there yet?-a systematic literature review on chatbots in education," *Frontiers in artificial intelligence*, vol. 4, p. 654924, 2021.

[22] C. K. Lo and K. F. Hew, "A review of integrating ai-based chatbots into flipped learning: new possibilities and challenges," in *Frontiers in Education*, vol. 8. Frontiers, 2023, p. 1175715.

[23] C. K. Lo, "What is the impact of chatgpt on education? a rapid review of the literature," *Education Sciences*, vol. 13, no. 4, p. 410, 2023.

[24] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[25] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.

[26] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig, "Pal: Program-aided language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 764–10 799.

[27] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22 199–22 213, 2022.

[28] J. Shieh, "Best practices for prompt engineering with openai api," *OpenAI, February https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api*, 2023.

[29] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, T. Gutman-Solo *et al.*, "Solving quantitative reasoning problems with language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3843–3857, 2022.

[30] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[31] M. Perkins, "Academic integrity considerations of ai large language models in the post-pandemic era: Chatgpt and beyond," *Journal of University Teaching & Learning Practice*, vol. 20, no. 2, p. 07, 2023.

[32] J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz, "Chatgpt goes to law school," *Available at SSRN 4335905*, 2023.

[33] T. Teubner, C. M. Flath, C. Weinhardt, W. van der Aalst, and O. Hinz, "Welcome to the era of chatgpt et al. the prospects of large language models," *Business & Information Systems Engineering*, vol. 65, no. 2, pp. 95–101, 2023.

[34] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[35] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[36] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young *et al.*, "Scaling language models: Methods, analysis & insights from training gopher," *arXiv preprint arXiv:2112.11446*, 2021.

[37] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[38] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "A survey on large language models: Applications, challenges, limitations, and practical usage," *TechRxiv*, 2023.

[39] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, "A swot analysis of chatgpt: Implications for educational practice and research," *Innovations in Education and Teaching International*, pp. 1–15, 2023.

[40] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, and B. He, "Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models," *arXiv preprint arXiv:2303.16421*, 2023.

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

[42] T. Dave, S. A. Athaluri, and S. Singh, "Chatgpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations," *Frontiers in Artificial Intelligence*, vol. 6, p. 1169595, 2023.

[43] I. Švab, Z. Klemenc-Ketiš, and S. Zupanič, "New challenges in scientific publications: Referencing, artificial intelligence and chatgpt," *Slovenian Journal of Public Health*, vol. 62, no. 3, pp. 109–112, 2023.

[44] B. Liu, B. Xiao, X. Jiang, S. Cen, X. He, W. Dou *et al.*, "Adversarial attacks on large language model-based system and mitigating strategies: A case study on chatgpt," *Security and Communication Networks*, vol. 2023, 2023.

[45] W. Geerling, G. D. Mateer, J. Wooten, and N. Damodaran, "Is chatgpt smarter than a student in principles of economics?" *Available at SSRN 4356034*, 2023.

[46] F. M. Megahed, Y.-J. Chen, J. A. Ferris, S. Knoth, and L. A. Jones-Farmer, "How generative ai models such as chatgpt can be (mis) used in spc practice, education, and research? an exploratory study," *Quality Engineering*, pp. 1–29, 2023.

[47] S. Hargreaves, "'words are flowing out like endless rain into a paper cup': Chatgpt & law school assessments," *The Chinese University of Hong Kong Faculty of Law Research Paper*, no. 2023-03, 2023.

[48] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo *et al.*, "Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models," *PLoS digital health*, vol. 2, no. 2, p. e0000198, 2023.

[49] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash *et al.*, "How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment," *JMIR Medical Education*, vol. 9, no. 1, p. e45312, 2023.

[50] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner, "Mathematical capabilities of chatgpt," *arXiv preprint arXiv:2301.13867*, 2023.

[51] A. Szabo, "Chatgpt a breakthrough in science and education: Can it fail a test?" Feb 2023. [Online]. Available: osf.io/ks365

[52] S. Jalil, S. Rafi, T. D. LaToza, K. Moran, and W. Lam, "Chatgpt and software testing education: Promises & perils," in *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2023, pp. 4130–4137.

[53] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, and B. Agyemang, "What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education," *Smart Learning Environments*, vol. 10, no. 1, p. 15, 2023.

[54] X. Zhai, "Chatgpt for next generation science learning," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 29, no. 3, pp. 42–46, 2023.

[55] R. A. Khan, M. Jawaid, A. R. Khan, and M. Sajjad, "Chatgpt-reshaping medical education and clinical management," *Pakistan Journal of Medical Sciences*, vol. 39, no. 2, p. 605, 2023.

[56] Z. Han, F. Battaglia, A. Udaiyar, A. Fooks, and S. R. Terlecky, "An explorative assessment of chatgpt as an aid in medical education: Use it with caution," *MedRxiv*, pp. 2023–02, 2023.

[57] Y. M. Al-Worafi, A. Hermansyah, K. W. Goh, and L. C. Ming, "Artificial intelligence use in university: Should we ban chatgpt?" *Preprints https://doi.org/10.20944/preprints202302.0400.v1*, 2023.

[58] O. Topsakal and E. Topsakal, "Framework for a foreign language teaching software for children utilizing ar, voicebots and chatgpt (large language models)," *The Journal of Cognitive Systems*, vol. 7, no. 2, pp. 33–38, 2022.

[59] J. Rudolph, S. Tan, and S. Tan, "Chatgpt: Bullshit spewer or the end of traditional assessments in higher education?" *Journal of Applied Learning and Teaching*, vol. 6, no. 1, 2023.

[60] C. K. Lo and K. F. Hew, "A critical review of flipped classroom challenges in k-12 education: Possible solutions and recommendations for future research," *Research and practice in technology enhanced learning*, vol. 12, no. 1, pp. 1–22, 2017.

[61] ——, "Design principles for fully online flipped learning in health professions education: a systematic review of research during the covid-19 pandemic," *BMC Medical Education*, vol. 22, no. 1, p. 720, 2022.

[62] C. K. Lo, "Strategies for enhancing online flipped learning: a systematic review of empirical studies during the covid-19 pandemic," *Interactive Learning Environments*, pp. 1–29, 2023.

[63] S. Nisar and M. S. Aslam, "Is chatgpt a good tool for t&cm students in studying pharmacology?" *Available at SSRN 4324310*, 2023.

[64] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[65] F. Ali *et al.*, "Let the devil speak for itself: Should chatgpt be allowed or banned in hospitality and tourism schools?" *Journal of Global Hospitality and Tourism*, vol. 2, no. 1, pp. 1–6, 2023.

[66] R. J. M. Ventayen, "Openai chatgpt generated results: Similarity index of artificial intelligence-based contents," *Available at SSRN 4332664*, 2023.

[67] M. Khalil and E. Er, "Will chatgpt get you caught? rethinking of plagiarism detection," *arXiv preprint arXiv:2302.04335*, 2023.

[68] H. H. Thorp, "Chatgpt is fun, but not an author," pp. 313–313, 2023.

[69] M. Sallam, "The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations," *medRxiv*, pp. 2023–02, 2023.

[70] A. B. Mbakwe, I. Lourentzou, L. A. Celi, O. J. Mechanic, and A. Dagan, "Chatgpt passing usmle shines a spotlight on the flaws of medical education," p. e0000205, 2023.

[71] D. Baidoo-Anu and L. Owusu Ansah, "Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning," *Available at SSRN 4337484*, 2023.

[72] J. Qadir, "Engineering education in the era of chatgpt: Promise and pitfalls of generative ai for education," in *2023 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 2023, pp. 1–9.

[73] D. Mhlanga, "Open ai in education, the responsible and ethical use of chatgpt towards lifelong learning," *Available at SSRN: 4354422*, 2023.

[74] X. Zhai, "Chatgpt user experience: Implications for education," *Available at SSRN 4312418*, 2022.

[75] P. Stutz, M. Elixhauser, J. Grubinger-Preiner, V. Linner, E. Reibersdorfer-Adelsberger, C. Traun, G. Wallentin, K. Wöhs, and T. Zuberbühler, "Ch (e) atgpt? an anecdotal approach on the impact of chatgpt on teaching and learning giscience," *Preprint https://doi.org/10.35542/osf.io/j3m9b*, 2023.

[76] D. R. Cotton, P. A. Cotton, and J. R. Shipway, "Chatting and cheating: Ensuring academic integrity in the era of chatgpt," *Innovations in Education and Teaching International*, pp. 1–12, 2023.

[77] Y. Al Ahmed and A. Sharo, "On the education effect of chatgpt: Is ai chatgpt to dominate education career profession?" in *2023 International Conference on Intelligent Computing, Communication, Networking and Services (ICCNS)*. IEEE, 2023, pp. 79–84.

[78] S. Ekin, "Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices," *TechRxiv*, 5 2023. [Online]. Available: doi:10.36227/techrxiv.22683919.v2

[79] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[80] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–23.

[81] X. Han, W. Zhao, N. Ding, Z. Liu, and M. Sun, "Ptr: Prompt tuning with rules for text classification," *AI Open*, vol. 3, pp. 182–192, 2022.

[82] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[83] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[84] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[85] L. S. Lo, "The clear path: A framework for enhancing information literacy through prompt engineering," *The Journal of Academic Librarianship*, vol. 49, no. 4, p. 102720, 2023.

[86] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[87] L. Giray, "Prompt engineering with chatgpt: A guide for academic writers," *Annals of Biomedical Engineering*, pp. 1–5, 2023.

[88] J. R. Fraenkel, N. E. Wallen, H. H. Hyun *et al.*, *How to design and evaluate research in education*. McGraw-hill New York, 2012, vol. 7.

[89] R. Kohavi and R. Longbotham, "Online controlled experiments and a/b tests," *Encyclopedia of machine learning and data mining*, pp. 1–11, 2015.

[90] R. Paul, "The state of critical thinking today," *New directions for community colleges*, vol. 2005, no. 130, pp. 27–38, 2005.