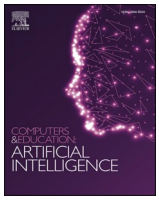
Please quote as: Tolzin, A., Knoth, N. & Janson, A. (2024). Worked Examples to Facilitate the Development of Prompt Engineering Skills. *Thirty-Second European Conference on Information Systems (ECIS 2024)*, Paphos, Cyprus.

Computers and Education: Artificial Intelligence 6 (2024) 100225

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

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



AI literacy and its implications for prompt engineering strategies

Nils Knoth [a], Antonia Tolzin [b], Andreas Janson [c,*], Jan Marco Leimeister [b,c]

[a] Institute for Psychology (IfP), University of Kassel (Germany), Holländische Straße 36-38, 34127, Kassel, Germany
[b] Research Center for Information System Design (ITeG), University of Kassel (Germany), Pfannkuchstraße 1, 34121, Kassel, Germany
[c] Institute of Information Systems and Digital Business, University of St. Gallen (Switzerland), Müller-Friedberg-Strasse 8, 9000, St. Gallen, Switzerland

ARTICLE INFO

Keywords:
Large language model
AI literacy
Prompt engineering
AI interaction
Education

ABSTRACT

Artificial intelligence technologies are rapidly advancing. As part of this development, large language models (LLMs) are increasingly being used when humans interact with systems based on artificial intelligence (AI), posing both new opportunities and challenges. When interacting with LLM-based AI system in a goal-directed manner, prompt engineering has evolved as a skill of formulating precise and well-structured instructions to elicit desired responses or information from the LLM, optimizing the effectiveness of the interaction. However, research on the perspectives of non-experts using LLM-based AI systems through prompt engineering and on how AI literacy affects prompting behavior is lacking. This aspect is particularly important when considering the implications of LLMs in the context of higher education. In this present study, we address this issue, introduce a skill-based approach to prompt engineering, and explicitly consider the role of non-experts' AI literacy (students) in their prompt engineering skills. We also provide qualitative insights into students' intuitive behaviors towards LLM-based AI systems. The results show that higher-quality prompt engineering skills predict the quality of LLM output, suggesting that prompt engineering is indeed a required skill for the goal-directed use of generative AI tools. In addition, the results show that certain aspects of AI literacy can play a role in higher quality prompt engineering and targeted adaptation of LLMs within education. We, therefore, argue for the integration of AI educational content into current curricula to enable a hybrid intelligent society in which students can effectively use generative AI tools such as ChatGPT.

1. Introduction

Artificial intelligence (AI) has developed quickly over the past ten years in a wide range of disciplines, as demonstrated by advancements in areas like computer vision, speech recognition, language modeling, abstract strategic gameplay, and others (Berg, Raj, & Seamans, 2023). Within the different approaches in AI, *large language models* (LLMs) are emerging as a particularly prominent one, constructing human-like language by iteratively anticipating likely next words based on the sequence of preceding words (Bommasani et al., 2021; McCoy, Yao, Friedman, Hardy, & Griffiths, 2023). LLMs are part of the broader category of generative AI, which refers to machine learning algorithms that can learn from different types of content, such as text, images, and audio, to generate new content (Cao et al., 2023). The models used in generative AI are capable of producing a variety of outputs, including audio, video, images, or text, based on user input, which is referred to as a prompt. In terms of text output, LLMs are the most notable development with the introduction of OpenAI's ChatGPT (OpenAI, 2023), which is capable of generating human-like language through a chat-based interface (Schöbel et al., 2024). As conversational user-interfaces present intuitive modes of interaction for various people, LLM-based AI systems and conversational agents are also being used more frequently in human-computer communication (Dwivedi et al., 2023; McLean & Osei-Frimpong, 2019). LLMs enable smooth and effective multi-turn conversations with users, lowering the barriers to developing conversational user experiences (Bommasani et al., 2021). LLMs' outstanding ability to compose high-quality and convincing output has generated excitement among students in higher education because it could be used to write essays and assignments (Dwivedi et al., 2023) and outscore human counterparts in a variety of domains (such as Law, e.g., Choi, Hickman, Monahan, & Schwarcz, 2023).

Furthermore, improvements in LLMs could have a significant impact on the educational field as a whole. For instance, recent studies have emphasized LLM's, such as ChatGPT's, capacity to enrich the

* Corresponding author. University of St.Gallen, Müller-Friedberg-Strasse 8, 9000, St. Gallen, CH, Switzerland.
E-mail addresses: nils.knoth@uni-kassel.de (N. Knoth), antonia.tolzin@uni-kassel.de, leimeister@uni-kassel.de (A. Tolzin), andreas.janson@unisg.ch (A. Janson), janmarco.leimeister@unisg.ch (J.M. Leimeister).

https://doi.org/10.1016/j.caeai.2024.100225
Received 13 December 2023; Received in revised form 27 February 2024; Accepted 16 April 2024
Available online 18 April 2024

educational experience by supporting a wide range of learning methodologies, including adaptive learning, personalized learning, and self-directed learning (Rahman & Watanobe, 2023; Rasul et al., 2023; Ruwe & Mayweg-Paus, 2023; Zhu, Sun, Luo, Li, & Wang, 2023). Additionally, LLMs can offer timely feedback to students, enhances information accessibility, improves student performance and motivation, and refines teaching practices, as evidenced in various recent publications (Alves de Castro, 2023; Crawford, Cowling, & Allen, 2023; Day, 2023; Farrokhnia, Banihashem, Noroozi, & Wals, 2023; Lee, 2023; Rudolph, Tan, & Tan, 2023; Su & Yang, 2023). As a result, LLMs have the potential to significantly advance higher education by enabling tailored learning experiences (Cao et al., 2023), enhancing group discussions (H. Nguyen, 2023), improving educational outcomes and learning strategies, and providing opportunities to be incorporated into several different learning methodologies (Eager & Brunton, 2023; Kikalishvili, 2023). Consequently, if we wish to actively use LLMs in education rather than ignoring them, we must take on this significant technological leap as a major issue for educators (Kohnke, Moorhouse, & Zou, 2023). Furthermore, because the industry needs workers who can use these tools, it is necessary to train students in prompt engineering. Dell'Acqua et al. (2023) showed the positive effects on consultant work outcomes, thus, this technology will be adopted industry-wide and we need to train our students in prompting LLMs.

Though, users still struggle to control the output produced by LLMs (Zamfirescu-Pereira, Wong, Hartmann, & Yang, 2023); thus educators must have a solid understanding of how to teach students how to interact with LLM based AI-systems effectively (Kohnke et al., 2023). *Prompt engineering*, which entails developing and improving specific inputs for generative AI models in order to obtain high-quality outputs from a model, is essential to this interaction (P. Liu et al., 2023). However, user prompt engineering is most often a matter of trial and error (Dang, Benharrak, Lehmann, & Buschek, 2022). It can be difficult to create effective prompts, and interactions based on prompts are frequently brittle. To accomplish effective communication with LLM based chatbots like ChatGPT, however, the ability to engineer effective prompts is becoming more and more important (White et al., 2023).

Despite the significant interest in LLMs, little is known yet how non-experts (i.e., individuals without formal instruction concerning AI and LLMs) create prompts and how effectively they are at doing so. Initial findings suggest that non-expert users may initiate prompting behaviors that are unsystematic and opportunistic, tending to overgeneralize expectations derived from human-to-human interaction (Zamfirescu-Pereira et al., 2023). One aim of the present study is to examine the ability of non-experts to generate prompts for LLMs and how it affects the LLM output in the context of higher education. Identifying scenarios in which LLM errors occur, coming up with ideas to correct them, and assessing the effectiveness of those solutions are necessary for designing effective prompts (Bommasani et al., 2021; P. Liu et al., 2023).

However, research in prompt engineering advancements is largely lacking these perspectives until now (Wang, Yu, & Huang, 2022; Zamfirescu-Pereira et al., 2023). This is crucial because the accessibility and pervasiveness of AI-based technologies raise questions related to the level of AI literacy among non-experts needed to effectively interact with and critically evaluate these technologies (Long & Magerko, 2020). As an emerging form of digital literacy, *AI literacy* includes the skills necessary for the competent and meaningful usage of AI tools. Even though AI literacy is regarded as a future skill (Vuorikari, Kluzer, & Punie, 2022), studies that examine how it may affect a user's behavior when dealing with LLM-based AI systems like ChatGPT are currently lacking (Pinski & Benlian, 2023). As a result, the current study addresses this issue and specifically takes into account the relationship between non-experts' AI literacy and their prompt-engineering skills. As of now, most research on prompt engineering has been conducted from a technology-centric viewpoint (Ding et al., 2021; P. Liu et al., 2023). In this study, we want to introduce a skill-based approach to prompt engineering as a critical element for enabling students to manage LLMs effectively. The guiding research questions (RQ) are as follows: (1) Can prompt engineering be conceptualized as a skill for the goal-directed use of LLMs in the context of higher education? and (2) How is AI literacy related to non-experts' ability to engage in prompt engineering?

2. Theoretical background and hypotheses

2.1. Prompt engineering

Creating input statements (prompts) for generative AI models is called prompt engineering (or prompt design, prompt programming, or prompting) (Oppenlaender, Linder, & Silvennoinen, 2023). For a large language model (LLM) to produce or alter its text output, input text or a set of instructions has to be formulated (White et al., 2023). The resulting interactions with an LLM-based AI system and its output are affected by a prompt's construction, which is accomplished by creating clear guidelines and rules for the LLM's dialogue utilizing a set of predetermined norms. According to White et al. (2023), a high-quality prompt essentially creates the structure for the dialogue and informs the LLM about which information is important as well as about the intended output form and content. Compared to the realm of AI development, prompt engineering is more directly tied to fields that focus on interactions between humans and AI, such as Human-Computer Interaction (HCI), Human-AI Interaction, and conversational AI (Oppenlaender, 2022).

However, LLMs also pose significant challenges (Bommasani et al., 2021) because they require skills to use this technology (Dwivedi et al., 2023; Zamfirescu-Pereira et al., 2023). In addition, LLMs also sometimes produce inaccurate or nonsensical outputs (known as hallucinations), and at the moment often lack common sense and comprehension of reality (Floridi & Chiriatti, 2020; Ji et al., 2023). Prompt engineering is a skill that entails creating and refining specific inputs for LLMs, enabling users to obtain high-quality outputs from a model to overcome these difficulties and take advantage of the capabilities of generative AI (P. Liu et al., 2023). Users utilize text prompts to guide pre-trained models through prompt engineering. This approach differs from adapting these models to downstream tasks via objective engineering, which involves modifying the model with new layers or parameters and training it with labeled data (P. Liu et al., 2023).

As a result, prompt engineering involves bi-directional human and AI interaction. To enhance the output produced by generative models, prompts have to be refined iteratively. Many studies on how humans engage with AI have turned to prompt engineering as a result of the growing use of these models (Dang, Mecke, Lehmann, Goller, & Buschek, 2022; Hou, Dong, Wang, Li, & Che, 2022; P. Liu et al., 2023). Even for natural language processing (NLP) professionals, creating efficient and generalizable prompts is difficult since it takes an extensive amount of trial and error, iterative testing, and rigorous evaluation of different prompt strategies on actual input-output pairs and large datasets (Oppenlaender et al., 2023). However, studies observing the particular prompting process of non-experts and investigating the factors that might support their intuitive prompting strategies are still lacking.

It has been widely acknowledged that developing efficient prompts for LLM-based AI systems like ChatGPT is important for getting a high-quality output (Dang, Benharrak, et al., 2022; Hou et al., 2022; White et al., 2023). Prior studies investigated how prompt keywords affect generative models, such as those that generate and display images (V. Liu & Chilton, 2021). Other research has concentrated on the prompt design for classification tasks and literature queries (Han et al., 2021). Contradictory task instructions within the context have also been discovered, even though the extended context in prompts has been demonstrated to improve text outputs (Wu, Terry, & Cai, 2021). Using the LLM itself to elaborate on problems is another method for improving prompt design (Betz, Richardson, & Voigt, 2021); this method is comparable to the human practice of "thinking aloud".

However, prompt design has not yet been studied broadly and systematically from an HCI perspective and quantitative findings obtained within an empirical study are sparse. An exception is the study by Oppenlaender et al. (2023), investigating the possibilities of prompt engineering in producing art with generative AI. They tested the ability of untrained participants to (1) recognize the quality of prompts, (2) create prompts by themselves, and (3) improve these prompts. The findings indicate that prompt engineering requires practice to become proficient, and that excellent prompt writing requires a working knowledge of important terminologies and phrasing. Zamfirescu-Pereira et al. (2023) examined the potential of prompt engineering by non-experts, using a prototype LLM-based chatbot design tool. They found that non-experts could explore prompt ideas but found it difficult to advance systematically because they had limited awareness of LLM capabilities. Furthermore, non-experts showed a propensity to develop prompts that resemble human-to-human instructions. Dang, Mecke, Lehmann, Goller, and Buschek (2022) assembled an HCI focus group and discovered several problems that occurred while creating prompts. The lack of clear direction in the trial-and-error process, the poor depiction of activities and their outcomes, worries about computing costs, and ethical implications are a few of these problems. Participants reported challenges in formulating efficient prompts, determining their efficacy, and defining their impact. These prompts are optimized to improve the LLM's performance for a specific task. However, our study differs from this technology-focused approaches by examining how non-expert users write and use prompts with an LLM-based AI system. Prompt construction plays a pivotal role in shaping the interaction between users and generative AI models, serving as the blueprint for communication. The iterative process of prompt engineering underscores the importance of refinement and testing to fine-tune model outputs and enhance performance. However, both experts and non-experts encounter challenges in this process, including the necessity for extensive trial and error and the potential limitations of applying human-to-human instruction paradigms to AI systems.

There have been different attempts to categorize and explain prompting methodologies. *Few-shot learning,* which instructs an LLM to learn a new task with few examples, enables task delegation spontaneously and improves model performance (Mialon et al., 2023). For the system's generating process, users can provide a brief text prompt. For users who are not experts in AI, prompts can steer the model in the direction of desired results (Zamfirescu-Pereira et al., 2023). *Zero-shot learning* requires prompting LLMs without any examples, but it can be enhanced by fine-tuning the instructions and reinforced learning via human feedback (Dang, Mecke, et al., 2022). To get better results, few-shot prompting could be paired with *chain-of-thought prompting* (Wei et al., 2022), which entails creating intermediary natural language reasoning steps to lead LLMs through challenging tasks. Furthermore, Eager and Brunton (2023) provided guidance for producing instructional text to direct the development of high-quality outputs from LLMs in higher education. In order to facilitate the process of prompt engineering, they recommend six components that should be included in written prompts: Verb, Focus, Context, Focus and Condition, Alignment, Constraints, and Limitations.

These prompting techniques might have implications for understanding and improving the quality of outputs and interactions with LLMs from an HCI perspective. Thus, we assume for our study that: *Students with higher prompt engineering skills will demonstrate LLM output of higher quality for their given task, due to the construction of better prompts (Hypothesis 1)*.

2.2. AI literacy

With the increasing prevalence of user-facing AI technologies, the concept of AI literacy has garnered significant attention in research (Long & Magerko, 2020). The concept of AI literacy was introduced by Kandlhofer, Steinbauer, Hirschmugl-Gaisch, and Huber (2016) and shaped significantly by Long and Magerko (2020). The authors define AI literacy as "a set of competencies that enables individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI, and use AI as a tool online, at home, and in the workplace" (Long & Magerko, 2020, p. 2). Besides this definition, different perspectives exist regarding the specific definition and skills associated with AI literacy (Laupichler, Aster, Schirch, & Raupach, 2022; Pinski & Benlian, 2023; Wienrich & Carolus, 2021). However, there is a consensus that AI literacy primarily targets non-experts, individuals who are not directly involved with AI in their studies or work and lack formal AI training (Laupichler et al., 2022; Ng, Leung, Chu, & Qiao, 2021).

AI is becoming a part of people's everyday lives, as more technologies and applications rely on AI algorithms and permeate people's decision-making processes and routines (Berg, Raj, & Seamans, 2023). However, there still remains a lack of awareness among individuals regarding their extent of AI usage, its inner workings, and its potential impact on their lives (Ghallab, 2019; Wienrich & Carolus, 2021). Similar to how digital literacy empowered individuals to use digital information and communication technologies, developing AI literacy becomes increasingly important for interacting with the omnipresent AI systems in our personal and professional spheres (Gašević, Siemens, & Sadiq, 2023; Ng et al., 2021; Vuorikari et al., 2022).

In contrast to the opacity of AI usage in previous technology, the launch of OpenAI's ChatGPT in November 2022 has sparked widespread public interest, accompanied by both enthusiasm and apprehension. To move beyond initial impressions and emotions surrounding generative AI, it is essential to consider its potential applications, the tasks it can perform, and areas where human skills remain indispensable. This necessitates a shift in perspective to understand humans' roles in a hybrid human-AI relationship (Baird & Maruping, 2021; Dellermann, Ebel, Söllner, & Leimeister, 2019; Salomon, Perkins, & Globerson, 1991). Addressing the maintenance of such a co-constructive relationship requires at least a basic understanding of AI, enabling informed decision-making aligned with personal goals (Vuorikari et al., 2022). The prominent emergence of generative AI technologies such as LLMs thus creates a momentum that calls for increased research efforts to investigate the impact of such AI-related competencies on the purposeful adoption and use of AI technologies.

As the future of education is expected to undergo significant transformation due to the widespread availability of powerful generative AI systems, it becomes crucial for non-experts to acquire the necessary skills, knowledge, and attitudes toward AI systems (Kasneci et al., 2023; Tarafdar, Page, & Marabelli, 2023). This might have implications not only for academic productivity and future employment opportunities but also for confident, critical, and safe engagement with emerging tools and technologies, building resilience against their vulnerabilities and risks (Long & Magerko, 2020; Tarafdar et al., 2023; Wienrich & Carolus, 2021). Consequently, equipping users with AI literacy might become a key factor in successfully integrating AI into higher education and future learning endeavors, enabling individuals to participate and act autonomously in a AI-infused world (Dignum, 2019). Therefore, AI literacy emerges as a decisive competency for higher education and academic success.

As natural language interfaces and their intuitive designs led to the prominent emergence of generative AI systems like ChatGPT, users also face specific challenges, often stemming from a tendency to anthropomorphize these systems (Krämer & Manzeschke, 2021; Zamfirescu-Pereira et al., 2023). While processes like the Theory of Mind may be useful for interpreting human behavior (Byom & Mutlu, 2013), it proves unreliable when applied to understanding AI, as AI and humans reason differently (Burrell, 2016; Schuetz & Venkatesh, 2020). Consequently, users who rely on their Theory of Mind mental models to interpret natural language AI outputs may develop misconceptions, leading to frustrating interactions and failure to realize the true potential of this technology (Bewersdorff, Zhai, Roberts, & Nerdel, 2023; Fügener, Grahl, Gupta, & Ketter, 2022). Instead, it is more appropriate

to develop a functional understanding of these systems as cognitive tools (Salomon et al., 1991), knowing when and how to use them and when not to, maximizing their educational benefits while minimizing their pitfalls (Lin, Ginns, Wang, & Zhang, 2020). Acquiring AI-related literacies holds the promise of enhancing human-AI interactions constructively, indicating a hybrid-intelligent educational paradigm where students augment their human intelligence with intelligent technologies, enabling them to achieve more collectively (Baird & Maruping, 2021; Dellermann et al., 2019; Salomon et al., 1991).

However, empirical studies investigating the impact of different levels of AI literacy on behaviors that emerge in partnership with technologies supported by AI are still scarce (Pinski & Benlian, 2023). Therefore, the present study aims to fill this research gap by examining the relationship between the individual AI literacy of AI non-experts and their prompt engineering behavior as a potential key factor for higher education to facilitate the reflective and goal-directed use of language models. By tracing the ways in which AI literacy influences real-world human-AI interactions, the findings aim to inform higher education institutions about the role of AI literacy in using language models for learning purposes such as collaboration and problem solving (Joksimovic, Ifenthaler, Marrone, Laat, & Siemens, 2023; Tan, Lee, & Lee, 2022) and advocate for the incorporation of AI literacy modules into higher education curricula. Thus, we assume that *students who load higher on AI literacy will engage in more sophisticated prompt engineering behavior by using more purposive prompting strategies (Hypothesis 2)*. Furthermore, we expect that *higher AI literacy is positively associated with LLM output quality (Hypothesis 3)*. The conceptual model underlying the hypotheses about the assumptions of the effective relationships between AI literacy, prompt engineering, and LLM outputs, can be found in Fig. 1.

3. Method

To investigate how non-experts interact with LLM-based AI systems and engage in prompt engineering, we used a mixed-methods research design to evaluate the hypothesized effects (Venkatesh, Brown, & Sullivan, 2016). Participants were asked to complete two tasks using a General Data Protection Regulation (GDPR) compliant platform that uses Open AI's Application Programming Interface (API) and the OpenAI gpt-3.5-turbo model for conversational interactions.

3.1. Sample

The sample size included $N = 45$ university students, aged between 19 and 35 years, thereof $n = 15$ women, $n = 28$ men, and $n = 2$ non-specified. They studied different subjects (Mechanical Engineering: $n = 15$, Psychology: $n = 6$, Business and Economics: $n = 21$, and $n = 3$ non-specified). Specifically, the classes in which the study was conducted were the following ones: A seminar on scientific writing for mechanical engineering students, an in-depth seminar in developmental psychology and a tutorial for an information science lecture. Based on their study subjects, they were assumed to be AI non-experts. In addition, it is interesting to note that 28 participants had used generative AI systems prior to participating in the study, while 17 participants didn't use generative AI systems prior to participating in the study. Thus, 17 students performed prompt engineering for the first time.

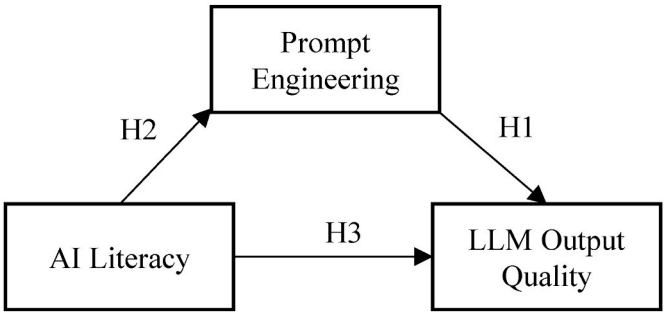


Fig. 1. Conceptual model of AI literacy and LLM interaction qualities.

3.2. Study design and materials

To assess students' prompt engineering, two tasks were designed that had to be solved employing an LLM: creating a comprehensive travel plan to Andorra (Task 1) and planning a scientific project on the topic of automated essay scoring (Task 2). These two tasks were constructed to capture two different usage scenarios. While Task 1 captures a generic prompt engineering scenario for leisure, Task 2 captures a scenario that can be contextualized within higher education requirements. Since we needed behavioral indicators of the sessions conducted with the GPT-based platform, we collected information through written protocols structured to capture the following aspects for each of the two tasks: (1) the prompts generated by the students to gain an output of the LLM (analyzed quantitatively and qualitatively) and (2) the output generated by the LLM (analyzed quantitatively). Therefore, participants had to copy their prompts and the generated outputs from the GPT-based platform into another tab that provided a structured environment to paste this information.

After completing the tasks, students were assigned a short reflection protocol designed to gather additional information concerning their thoughts and feelings when working with the LLM, addressing their (1) perceived ease of writing prompts, (2) perceived task complexity, (3) perceived quality of LLM outputs, and (4) general user experience with the generative AI. Moreover, students were asked about their personal innovativeness (Agarwal & Prasad, 1998), a measure used to capture their general enthusiasm for new technologies, and trust in generative AI (Lankton, McKnight, & Tripp, 2015) for subsequent statistical control.

The study was conducted in May 2023. There was no compensation for participation in the study. The communication of the purpose of the study, that it was about the use of ChatGPT, was intended to serve as an incentive to participate, as there was no official input from the university on the topic at that time, but student interest in the technology might have been strong.

3.2.1. Assessment of students' AI literacy

We assessed students' levels of AI literacy utilizing the AI Literacy Scale (Pinski & Benlian, 2023) to investigate the impact of generic AI-related competencies on the use of generative AI tools. Although the concept of AI literacy was fundamentally shaped by the framework of Long and Magerko (2020), we used an AI literacy instrument that is not directly tied to the competency dimensions proposed by these authors. While we acknowledge the valuable contribution of Long and Magerko (2020), AI literacy may encompass additional aspects (see, e.g., Ng et al., 2021). At the time of the study, Pinski and Benlian (2023) AI literacy scale was one of the first instruments to make AI literacy measurable. They define AI literacy as "humans' socio-technical competence consisting of knowledge regarding human and AI actors in human-AI interaction, knowledge of AI process steps, that is input, processing, and output, and experience in AI interaction." (Pinski & Benlian, 2023, p. 169). Thus, the use of this scale in the present study was motivated by the interactionist and experiential perspectives it captures. While other AI literacy scales that emerged at the time of the study (e.g., Laupichler, Aster, & Raupach, 2023) capture more declarative knowledge-related aspects of AI, the instrument provided by Pinski and Benlian (2023) could potentially provide interesting insights into human-AI interaction qualities, making it particularly suitable for the field of prompt engineering. The original scale consists of 28 items that reflect six subscales. Example items of the scale are: "*I have knowledge of use cases for AI technology*" or "*I have knowledge of the tasks that human actors can assume in human-AI collaboration*". All items were responded to on a 7-point Likert scale (strongly disagree to strongly agree). Since the scale was

translated into German for this study, Cronbach's alpha was determined for each subscale, revealing good to excellent reliability: AI technology knowledge (4 items; $\alpha = 0.80$), human actors in AI knowledge (5 items; $\alpha = 0.86$), AI steps knowledge (12 items; $\alpha = 0.93$), AI usage experience (2 items; $\alpha = 0.80$), AI design experience (2 items; $\alpha = 0.95$), and AI literacy (overall) (3 items; $\alpha = 0.67$). Cronbach's alpha values within this original model exceeded those of an adjusted model with fewer items proposed by Pinski and Benlian (2023). Model testing indicated no multicollinearity problems for any of the dimensions (VIFs <5.00) for this newly translated German version. Therefore, the original 28-item version of the scale was used in this study. The distribution of mean values and standard deviations across all subscales, for the sample collected, can be found in Table 1. Notably AI design experience is the lowest, further indicating that the sample can be characterized as AI non-experts.

3.2.2. Assessing the quality of the LLM output

The quality of the LLM output was evaluated by employing an integrative complexity score (Janson, Söllner, & Leimeister, 2020), addressing differentiation and integration as the two cognitive structural traits (Suedfeld, Tetlock, & Streufert, 1992). *Differentiation* denotes the extent to which a person considers the separate aspects of a problem. *Integration* denotes the extent to which a person creates intricate relationships between diverse features of a problem. Thus, LLM outputs for each task were scored, using a 10-point scale (ranging from 1: minimal or no differentiation and integration, to 10: high differentiation and integration), following the method proposed by Baker-Brown et al. (1992). Each LLM output produced by the participants to solve the given tasks was coded according to this integrative complexity score. The second author trained a graduate student coder, and both coded the data independently from each other. To ensure a comprehensive analysis, the coders carefully reviewed each LLM output multiple times. In addition, both coders were blind to the coding of the respective other. Regarding LLM output coding, inter-rater reliability (IRR; Pearson correlation coefficient; Task 1 (travel task): $r = 0.96$; $n = 42$; $p < 0.001$; Task 2 (project task): $r = 0.96$; $n = 42$; $p < 0.001$) as well as inter-rater agreement (IRA; weighted Cohen's kappa; Task 1: $\kappa_w = 0.81$; $n = 42$; $p < 0.001$, Task 2: $\kappa_w = 0.84$; $n = 42$; $p < 0.001$) showed substantial agreement between raters (LeBreton & Senter, 2008).

3.2.3. Assessing the quality of prompt engineering quantitatively

To gain comprehensive insights into students' actual prompt engineering behaviors, qualitative and quantitative methods were used. For quantitative analysis, a prompt quality score was assigned to each generated prompt, taking into account the prompt components proposed by Eager and Brunton (2023). The components include (1) verb, (2) focus, (3) context, (4) focus and condition, (5) alignment, and (6) constraints and limitations (see Table 2). It is suggested that these aspects affect the quality of the results generated by an LLM. Specific examples of how these components can be applied to the creation of prompts in both industry and higher education settings can be found in Appendix C and Appendix D, respectively. Therefore, each component contained in a prompt was given one point. Thus, a score of 0 indicates that none of the six components were included in the prompt, while a score of 6 indicates that all of the prompt components were included.

Table 1
Mean and standard deviation of AI literacy subscales (Pinski & Benlian, 2023).

AI literacy subscales	*M*	*SD*
AI literacy (overall)	3.28	1.19
AI technology knowledge	4.14	1.22
Human actors in AI knowledge	4.41	1.12
AI steps knowledge	3.98	1.24
AI usage experience	3.61	1.70
AI design experience	2.81	1.90

Note. All items ranged from 0 (strongly disagree) to 7 (strongly agree).

Table 2
Prompt components according to Eager and Brunton (2023).

Component	Purpose
Verb	Indicates a specific action to be performed.
Focus	Provides the process, product, or outcome of the action to be performed (in relation to the 'verb').
Context	Explains the scope or parameters of the task.
Focus and Condition	Provides the focus and condition for the generated output, defining the subject matter and the primary goal. This information can help to narrow down the scope of the task and clarify what the content should include.
Alignment	Instructs the AI model to align content with your desired goal.
Constraints and Limitations	Note any constraints or limitations that the AI model should adhere to.

The second author trained a graduate student coder. The data was coded independently.

To ensure a thorough examination, the coders carefully examined each prompt several times, taking care to identify all elements of the prompt. The coding of prompt engineering quality had largely a good inter-rater reliability (IRR; Pearson correlation coefficient; $r = 0.83$; $n = 42$; $p < 0.001$ & $r = 0.80$; $n = 42$; $p < 0.001$) and inter-rater agreement values (IRA; weighted Cohen's kappa; $\kappa_w = 0.80$; $n = 42$; $p < 0.001$ & $\kappa_w = 0.71$; $n = 39$; $p < 0.001$). Since IRA values were slightly lower for Prompt Engineering coding, both raters resolved any noticeable discrepancies through discussion until agreement on a single consensus score was reached.

3.2.4. Assessing the quality of prompt engineering qualitatively

In addition to the quantitative analyses, which are the main focus of the paper, the prompts used to generate the LLM output and to solve the tasks were also analyzed qualitatively using an inductive approach (Gioia, Corley, & Hamilton, 2013) to identify specific concepts. As there is currently no comprehensive and validated prompt taxonomy, we could not perform a deductive qualitative analysis. By doing so, potential peculiarities or specific prompting behaviors of non-experts in this rather new and emerging topic area can be explored and uncovered. It enables a deeper understanding of prompt engineering from a human-centered perspective, which is relevant for future research.

The second author trained a graduate student coder, and after instruction, the data was coded independently. To ensure a comprehensive analysis, the coders thoroughly reviewed each prompt multiple times, with a keen focus on identifying and categorizing specific features within the prompts. These features were crucial components of the research process, and the coders diligently documented them. These predetermined prompt features were extracted: (1) number of words (total, across all prompts to solve the task), (2) number of prompts used to solve the task, (3) elements of human-like communication/communication style, and (4) syntax type of sentence (declarative, interrogative, imperative, exclamatory). There was strong agreement (i.e., at least 95% agreement) between the two raters.

4. Results

4.1. Quantitative results

4.1.1. Descriptive statistics

As the adoption of generative AI into higher education contexts still poses a novelty, insights into students' evaluations and their perceptions towards their interaction with the LLM-based AI system are provided in the following. For this purpose, several items were collected through a "reflection" protocol assigned after task completion and analyzed descriptively (Table 3). Students perceived the interaction with a GPT-based platform build for education contexts as rather positive, in terms of fulfilled expectations, their quality assessments of outputs, as well as general user experience. They also indicated that they would use

Table 3
Student's perceptions and evaluations of their interactions with generative AI.

Evaluative human-AI interaction items	*M*	*SD*
User Experience while interacting with the generative AI	3.59	0.74
Would you use generative AI again to handle similar tasks?	4.16	1.02
How do you rate the quality of the AI's generated texts in terms of correctness and comprehensibility?	3.91	0.92
Did the generated texts meet your expectations?	4.18	0.83
I found it difficult to write inputs/prompts for generative AI.	1.91	0.92
How comfortable are you with using generative artificial intelligence in general?	3.82	0.98

Note. All items ranged from 0 (strongly disagree) to 5 (strongly agree); User Experience was measured with 3 items.

generative AI again to handle similar tasks. The perceived difficulty to write and design prompts was low, a finding that is interesting to reflect on, taking into account the average of achieved points for their prompting behaviors, which are rather mid (see Table 4). Furthermore, students expressed general interest in using generative AI and reflected a more favorable attitude regarding it. These two findings may be related, as perceived competence is a strong predictor of performance satisfaction when interacting with a chatbot interface (Q. N. Nguyen, Sidorova, & Torres, 2022). This has implications for future AI educational endeavors because interest and attitudes are important success factors for learning (Eccles & Wigfield, 2002). The importance of intrinsic motivation in the adoption of LLMs like ChatGPT for successful learning is also supported by recent research (Lai, Cheung, & Chan, 2023). This effect also appears to be bi-directional, as research suggests that the use of ChatGPT and familiar learning tools can be an enhancement to motivation and self-efficacy (Sikström, Valentini, Sivunen, & Kärkkäinen, 2022; Yilmaz & Karaoglan Yilmaz, 2023).

To control for other variables potentially affecting AI literacy and the quality of prompt engineering, trust in generative AI (Lankton et al., 2015) and personal innovativeness (Agarwal & Prasad, 1998) were assessed additionally, as these typically pose relevant constructs when assessing usage of information technology in learning contexts (Gunness, Matanda, & Rajaguru, 2023). None of these constructs showed any anomalies, as possible outliers were examined using box plots. Since no hypothesis was formulated regarding these constructs, no further calculations were performed.

4.1.2. Regression analysis

To test the postulated hypotheses, a series of regression analyses were performed. The measures of interest taken into account are the AI literacy subscales of Pinski and Benlian (2023) and the sum scores resulting from the quantitative assessment of 1) the prompt engineering performed to solve the given tasks; and 2) the generated outputs derived from the LLMs. These scores were analyzed for each given task (travel plan & project plan) respectively. The analyses presented below are based on different sample sizes because some participants either did not provide prompts for their task performance or did not copy their generated LLM output. Because AI literacy was measured on multiple subscales, the variance inflation factor was examined. The VIF for all subscales in all regression models was less than 10, indicating no significant problem with multicollinearity.

To assess the effect of prompt engineering on the quality of LLM outputs (Hypothesis 1), a linear regression was performed with the rated quality of the generated output as the criterion and the rated quality of the prompt engineering performed as the predictor. In the first model predicting the quality of the generated travel plan (Task 1), the results showed a significant beta coefficient for the quality of the prompt engineering towards the quality of the travel plan output ($b = 1.49$, $t(40) = 6.78$, $p < 0.001$). This model was able to predict approximately 53% of the variance in the quality of the generated travel plan output ($R^2 = 0.535$, $F(1, 40) = 46.01$, $p < 0.001$). The second model, which predicted the quality of the three task solutions in the context of the scientific project planning task (Task 2), also showed a significant beta coefficient (b = 1.376, $t(37) = 11.502$, $p < 0.001$). This model was able to predict about 78% of the variance in the quality of the generated output ($R^2 = 0.782$, $F(1, 37) = 132.3$, $p < 0.001$). Within the two different tasks, the same effect was found, i.e., that higher quality prompt engineering behavior is indeed associated with higher quality LLM output, and that the variance in LLM output quality is largely explained by prompt engineering skills. Therefore, H1 is supported.

As the main focus of this present work, we also analyzed the influence of AI literacy on prompt engineering skills (H2). For this purpose, two multiple regression analyses were performed (see Table 5). The criterion was the quality of prompt engineering for each task. According to the hypothesis, the predictors were the AI literacy subscales of Pinski and Benlian (2023): AI technology knowledge, human actors in AI, AI steps knowledge, AI usage experience, AI design experience, and AI literacy (overall). Due to the explorative nature of this study and its small sample size, effects and trends found within the data should be taken with caution.

The model for the travel plan task (Task 1) yielded a significant effect of AI literacy on prompt engineering behavior and two marginally significant trends. The effect was found on the AI technology knowledge subscale (b = 0.579, $t(36) = 2.244$, $p = 0.031$), suggesting a positive impact of this aspect of AI literacy on prompt engineering skills. Next, the AI usage experience subscale of AI literacy showed a tendency toward better prompt engineering (b = 0.268, $t(36) = 1.791$, $p = 0.082$). Another trend was found in the AI steps knowledge subscale (b = −0.461, $t(36) = -1.810$, $p = 0.079$), suggesting a counterintuitive negative association between this aspect of AI literacy and prompt engineering. The second regression model for the task of planning a scientific project (Task 2) showed neither significant effects nor any tendency. Therefore, H2 is partially rejected. Possible reasons for this and possible implications will be discussed later.

In order to assess the influence of AI literacy on the quality of the generated LLM outputs (H3), an analytical procedure similar to the one used to test H2 was performed (see Table 6).

These models differed only in their criterion, which was the rated quality of the LLM outputs for each respective task. The model for the travel plan task (Task 1) did yield two significant effects and one tendency. One significant effect was again found in AI technology knowledge (b = 1.264, $t(35) = 2.391$, $p = 0.022$), pointing toward a positive influence of this AI literacy component on LLM outputs of higher quality. Another marginal significance was found in AI literacy (overall) (b =

Table 4
Quality of Prompt Engineering and generated LLM outputs.

Variable	*M*	*SD*
Prompt Engineering Quality (Task 1: Travel Task)	3.19	1.16
Prompt Engineering Quality (Task 2: Project Task)	3.36	1.01
LLM Output Quality (Task 1: Travel Task)	4.30	2.45
LLM Output Quality (Task 2: Project Task)	5.06	2.06

Note. Prompt Engineering Quality was rated from 0 to 6; LLM Output Quality was rated from 0 to 10.

Table 5
Regression model for H2 (task 1: Travel task prompts).

Predictor	Estimate	*SE*	CI 95%		*p*
			LL	UL	
AI literacy (overall)	−0.335	0.255	−0.852	0.183	0.198
AI technology knowledge	0.579	0.258	0.056	1.102	0.031*
Human actors in AI knowledge	−0.010	0.215	−0.446	0.426	0.964
AI steps knowledge	−0.461	0.255	−0.978	0.056	0.079.
AI usage experience	0.268	0.150	−0.035	0.571	0.082.
AI design experience	0.056	0.139	−0.226	0.337	0.692

Note. ***$p < .001$, **$p < .01$, *$p < .05$. Unstandardized coefficients and standard errors are reported. R^2 is not significant and is therefore not reported.

Table 6
Regression model for H3 (task 1: Travel task outputs).

Predictor	Estimate	*SE*	CI 95%		*p*
			LL	UL	
AI literacy (overall)	−1.112	0.529	−2.185	−0.038	0.043*
AI technology knowledge	1.264	0.529	0.191	2.337	0.022*
Human actors in AI knowledge	−0.201	0.457	−1.128	0.726	0.663
AI steps knowledge	−0.799	0.521	−1.856	0.259	0.134
AI usage experience	0.521	0.303	−0.095	1.137	0.095.
AI design experience	0.172	0.282	−0.401	0.745	0.546

Note. ***$p < .001$, **$p < .01$, *$p < .05$. Unstandardized coefficients and standard errors are reported. R^2 is not significant and is therefore not reported.

−1.111, $t(35) = -2.103$, $p = 0.043$), indicating a counterintuitive negative influence on LLM outputs. A trend was again found in AI usage experience (b = 0.521, $t(35) = 1.717$, $p = 0.095$) suggesting a potentially positive influence of this aspect of AI literacy towards better LLM outputs. The model for the task of planning a scientific project (Task 2) showed neither significant effects of AI literacy on LLM outputs, nor trends. In light of this, H3 is partially rejected but left open for discussion.

4.2. In-depth analysis of prompts

Next, we will turn to the in-depth analysis of the prompts used to generate the LLM output and to solve the tasks, to shed light on the prompt engineering behaviors of non-experts. By coding the data, we articulated the emergent themes that we discuss below.

4.2.1. Number of words and prompts

On average, students wrote $M = 5.0$ prompts ($SD = 2.7$, range: 1 to 12) and $M = 64.9$ words ($SD = 36.0$, range: 6 to 143) for Task 1 (travel plan). The average number of prompts generated by students for Task 2 (Project Plan) was $M = 4.3$ ($SD = 2.5$, range: 1 to 8). For Task 2 students wrote $M = 57.8$ words ($SD = 35.3$, range: 11 to 203).

4.2.2. Generative AI as a human conversational partner

An analysis of the communication style within the prompts revealed that most students showed signs of a human-to-human conversational structure in their prompting behavior. Students showed a tendency to incorporate polite and socially established elements into their interactions with the generative AI. It included instances of warmth and gratitude, which made their prompts feel more like conversations with a human rather than an AI. For example, Student 14's first prompt began with a friendly greeting and a note of appreciation: *"Hi,"* followed by *"I need to plan a trip to Andorra."*. This politeness continued throughout her queries, such as *"Thank you"* in the ninth prompt.

Student 27 also demonstrated a courteous demeanor, politely asking for recommendations for a trip to Andorra, as seen in its first request: *"I would like to go to Andorra in September, can you please give me some recommendations?"*. This student seems to attribute human-like qualities to the generative AI in its mental model, perhaps not fully understanding the mechanics of how a language model generates responses. In some cases, students went further and asked for the AI's opinion, such as student 28 who asked, *"What do you think would be the best choice, car or plane?"*. This implied a degree of anthropomorphism in their perception of the AI.

In addition, some students approached the generative AI with requests and queries that were in line with the expectations of a human interlocutor. For example, Student 31 asked the AI to *"imagine an automated essay grading"*, while Student 44 explicitly asked for help planning a trip, saying *"Hi, I want to go on a trip to Andorra, but I don't know much about it, can you help me?"*. This behavior suggests that these students saw the AI as more than a tool but as a conversational companion. The extent of this anthropomorphism was most evident in Student 32's prompt. In this case, the student justified their choice of research question by assuming that providing this information would be of interest or benefit to the AI: *"I like research question 5 because it addresses the issue of the objectivity of machines. Can you give me two more similar questions for an undergraduate thesis?"*. This interaction illustrates how some students may have perceived the generative AI as an intelligent, conversational partner with shared interests and skills, rather than as a tool for generating text. Students show a socially oriented communication style, which tends to be informal and focuses on sharing affective and emotional information (Kreijns, Kirschner, & Jochems, 2003).

4.2.3. Prompts as questions

Finally, we analyzed the syntax type of the sentences that the students wrote to generate LLM output. We distinguished four syntax types: Declarative, Interrogative, Imperative, and Exclamatory. Examples of these and their descriptive distribution can be found in Table 7. Across both tasks, most prompts were formulated in the form of questions (interrogative syntax style; Task 1: 131/282; Task 2: 122/214).

Across both tasks, a strikingly substantial portion of the prompts took the form of inquiries. It was apparent that many students failed to provide explicit instructions to guide the generative AI in producing the desired output. For instance, a prime example of this issue can be seen in the approach taken by students 14 and 24. They tackled the tasks solely by posing questions, without furnishing clear directives for the AI. Their sequence of prompts illustrates this pattern: *1*st *prompt: Do you know anything about automated essay scoring? 2*nd *prompt: What research questions can be asked about it? 3*rd *prompt: How long would it take to write a research project on it? 4*th *prompt: What steps need to be completed and when to complete the science project? (student 14)*.

*1*st *prompt: What is the cheapest way to get to Andorra? 2*nd *prompt: What are the sights in Andorra? 3*rd *prompt: What are the best places to stay in Andorra? 4*th *prompt: What is the weather like in Andorra in summer? 5*th *prompt: What language is spoken in Andorra? (student 24)*.

These students frequently approached generative AI as if it were a mere repository of information, similar to traditional search engines. They seemed to overlook the remarkable potential of generative AI to autonomously create novel content. This tendency may stem from a lack of awareness among non-experts regarding the multifaceted capabilities of generative AI. Many people who are unfamiliar with the intricacies of AI may inadvertently default to behaviors they are accustomed to when using other familiar technological tools, such as traditional search engines. This behavior could be due to their limited exposure to the transformative capabilities of generative AI, which go beyond mere data retrieval to include the ability to generate entirely original content. The potential of generative AI to innovate and provide unique insights may not have been fully appreciated by these students, leading them to underutilize this powerful tool.

5. Discussion

Our study aimed at conceptualizing prompt engineering skills in higher education, which is an important prerequisite for conducting

Table 7
Syntax Types used in the prompting process.

Syntax Types	Task 1 (Travel Plan)	Task 2 (Project Plan)
Declarative (e.g., "I need to plan a trip to Andorra.")	99	49
Interrogative (e.g., "What else can I see in the capital?")	131	122
Imperative (e.g., "Create a project plan including a time schedule.")	44	41
Exclamatory (e.g., "Thank you, that sounds great!")	8	2

further research on prompt engineering. The quality of prompt engineering predicted the quality of LLM outputs for their respective tasks to a high degree (see section 4.1.2). Thus, our empirical data supports the notion that more advanced prompt engineering does indeed promote the generation of higher-quality LLM outputs, making users more capable of exploiting the enormous potential of this technology. As there is a lack of empirical studies quantifying prompt engineering and LLM outputs, these findings provide some of the first empirical evidence on this topic.

We also aimed to provide insights into the relationship between generic AI literacy and prompt engineering skills. More specifically, we wanted to shed light on the question of what factors determine whether a user is capable of proper prompt engineering. Stemming from the theoretical line of reasoning based on mental models of AI (Tolzin & Janson, 2023) we investigated AI literacy in more detail. According to Zamfirescu-Pereira et al. (2023), AI non-experts can engage in prompt engineering but often struggle to make systematic progress due to an incomplete understanding of the capabilities of LLMs and a tendency to create prompts that mimic human-to-human instructions. These findings were corroborated and extended by our qualitative analysis of the prompts. The students behaved towards the LLM-based AI system in the same way as towards a human interlocutor, using socially desirable phrases ('hello', 'thank you') and trying to explain their inner lives and motives. Thus, AI non-experts perceive computers, and especially LLM-based AI systems, as social actors (Nass, Steuer, &; Tauber, 1994). Because of the human-like interface and conversational capabilities of LLMs, people attribute human characteristics to them (Bewersdorff et al., 2023). This behavior is known from other conversational interfaces, chatbots (Janson, 2023), voice assistants, and learning tutors, and is based on social response theory (Nass & Moon, 2000). Human-like cues, such as the way language is used and the context in which AI is introduced to students, can impact how people perceive social presence as well as mindful and mindless anthropomorphism (Araujo, 2018; Munnukka, Talvitie-Lamberg, & Maity, 2022). Moreover, identifying the LLM-based AI system as human raises user expectations for interactivity (Go & Sundar, 2019). Hill, Randolph Ford, and Farreras (2015) showed, that people used more and shorter messages with a more restricted vocabulary and more profanity when chatting with a chatbot compared to a human-to-human online conversation. AI non-experts do not know how LLMs generate their output and what information is important to be included in effective prompts. Therefore, as LLM-based AI systems are seen as a teammate and companions for collaboration and task-solving (Nißen et al., 2022; Seeber et al., 2020; Siemon et al., 2022), it could be helpful in higher education to impart knowledge about the functioning of generative AI to leverage the opportunities of AI-based tools, and, at the same time, preventing increasing anthropomorphizing and potentially coming with that, diffusion of responsibility. We hypothesized that this tendency might diminish as people become more AI literate.

In anticipation of answering this question, the current study examined the role of AI literacy in prompt engineering and the quality of LLM outputs. The results are mixed and must be treated with great caution, as the statistical power with a sample size of $N = 45$ is not large enough to detect small to medium effects. Thus, the general regression models were not significant. However, inspecting the data in more detail, AI technology knowledge predicted prompt engineering quality in the travel-plan task (Task 1). This AI literacy subscale is characterized by knowledge of the distinctiveness between AI and non-AI technology, the identification of use cases for AI technology, and the roles that AI technology can play in human-AI interaction. Taking this into perspective, these aspects are also important when interacting with LLM-based AI systems. In particular, knowledge of the roles that AI can play in human-AI interaction is important for building correct mental models of AI behavior and functioning, which has implications for constructing an effective dialogue with LLMs. Aspects of this can also be corroborated by our qualitative results. AI technology knowledge was also a relevant predictor of the quality of the LLM output for Task 1, together with a negative estimate of AI literacy (overall). It should be noted, however, that AI literacy (overall) is not a unique subscale, but a short general AI literacy measure whose Cronbach's alpha of 0.67 raises doubts about its reliability. Nevertheless, this negative effect may point to possible counterintuitive relationships between AI literacy and human-AI interactions. Similar findings were recently reported by Tully, Longoni, and Appel (2023), who showed that higher levels of AI knowledge predicted lower rates of AI receptivity. Nonetheless, the significance of both of these predictors, AI technology knowledge and AI literacy (overall), did not hold for Task 2, thus casting doubt on their robustness.

Another aspect that showed a positive trend within the travel plan task (Task 1) at the prompt engineering and output level was AI usage experience. Although not statistically significant, this trend may indicate some influence of prior experience on interactions with AI for prompt engineering, particularly as these are often characterized by a trial-and-error nature. Within Task 2, however, this trend was again not significant and pointed towards a negative influence. Further research is needed to make a more conclusive statement about the role of this aspect of prior AI usage experience. The remaining negative trend of AI steps knowledge falls into the same category and should be re-observed under conditions of higher statistical power, within a more large-scale study.

Alternatively, if we stay with the null hypothesis, the negative estimate result and the fact that the remaining subscales did not show substantial effects could also lead to the conclusion that AI literacy may not be necessary to use LLMs through targeted prompt engineering strategies. Rather, everyone may be able to generate prompts to some degree, pointing to the democratization and consumerization of AI as well as basic empowerment through the provision of this general-purpose technology per se (Gregory, R. W., Kaganer, E., Henfridsson, O., Ruch, T. J., 2018; Schmitt, Zierau, Janson, & Leimeister, 2023). Nevertheless, the average quality of the prompts examined in this study was of rather low quality, as were the outputs (see Section 4.1.1). Given that some participants were able to produce higher quality prompts, the question remains as to what predicts whether a person is capable of being a good prompt engineer. As such, future research may want to investigate the factors that can support people in their prompt engineering strategies, with AI literacy being a possible cornerstone, but not sufficient to explain the actual use of strategies.

5.1. Limitations

Next to the obvious constraint of the limited sample size of this explorative study, the major limitation within this present study may be its operationalization of prompt engineering behavior, as it solely relied on the prompt components proposed by Eager and Brunton (2023). With this operationalization, other concepts of prompt engineering are not captured that may have more pronounced relationships to AI literacy. In addition, there is no objective measurement option for prompt engineering skills to date, which poses a serious limitation to prompt engineering research as a whole. It is therefore advocated to further explore concurring options to model and measure prompt engineering behaviors. Despite the limitations mentioned, this study provides first insights into the intuitive behaviors of students, while engaging in prompt engineering, rather than capturing data via self-report questionnaires, that may lack validity. Another aspect worth discussing is the choice and construction of the two tasks that may be relatively easy to solve. Future studies could replicate our approach with more attention to task features that require more prompt engineering and investigate how scaffolds such as worked examples facilitate prompt engineering with varying task complexity (Tolzin, Knoth, & Janson, 2024).

5.2. Implications

Despite the limitations, two aspects need to be further discussed concerning their practical implications. An important result of our investigation is that prompt engineering indeed can be conceptualized

as a skill that can potentially be learned and promoted, affecting the quality of outputs one can achieve by using LLMs. This is supported by the finding that prompt engineering predicted LLM output quality in a significant way (Hypothesis 1). Future studies should investigate this aspect, through pre-post experiments that provide interventions that could potentially foster prompt engineering skills. Thus, only experimental research can provide a conclusive statement about the learnability of prompt engineering. Nevertheless, future studies in different contexts, such as industry use cases, investigating prompt engineering can benefit from the present study by its provided novel research design that allows for the systematic investigation and quantification of prompt engineering behaviors.

Furthermore, the mixed results concerning the relationship between AI literacy and prompt engineering on the one hand and AI literacy and the quality of the LLM output on the other hand, suggest that prompt engineering as a skill may be partially independent of an individual's AI literacy, opening up the possibility of teaching prompt engineering even to student populations that have very little to no AI literacy. Nonetheless, and importantly, AI literacy may play another role in the usage and interaction with LLM-based AI systems, namely task delegation. For example, a recent study by Pinski, Adam, and Benlian (2023) showed that empowering people with AI knowledge (increasing their AI literacy) influences their evaluation of tasks that are more appropriate for either humans or AI (human-fit vs. AI-fit task appraisal), as well as their decisions to delegate AI-appropriate tasks to AI tools. Taking this finding into account, AI literacy could provide a general context for the appropriate use of LLMs, such as ChatGPT, in higher education, as identifying the suitability of tasks for such systems is just as important as the prompting behavior itself. In addition, AI literacy may also have a significant impact on the tendency of students to rely on AI outputs, and as such, may contribute to the maintenance of student agency in the context of AI-assisted learning (Darvishi, Khosravi, Sadiq, Gašević, & Siemens, 2023). This may have implications for the responsible, fair, and safe use of LLMs in educational settings and needs to be further explored.

6. Conclusion

The present study provides a first glimpse into the role of non-experts' AI literacy for prompt engineering skills and their intuitive behaviors toward LLM-based AI systems. Although the small sample size was a serious limitation, the basic mixed-methods research design still provided some fruitful insights in this exploratory research area. First, we found empirical evidence that higher-quality prompt engineering indeed predicts LLM output quality. With this finding, we position prompt engineering as a quantifiable skill that differentiates between individuals who are able to use LLMs in a productive manner and those who may have difficulty producing the results they desire. This also points to future research that investigates the trainability of this particular skill. Second, AI literacy of non-experts may play a role in prompt engineering of higher quality, especially knowledge of AI technology and its role in human-AI collaboration may be important. As a result, AI literacy, or certain aspects of it, could serve as a prerequisite for the development of prompt engineering skills. However, AI literacy may also serve other purposes in human-AI interactions with LLM-based AI systems that could not be investigated in this study, such as trusting generated results or dealing with hallucinations. However, it could also be argued that AI literacy is not necessarily required to use LLMs at all, as the remaining subscales besides AI technology knowledge showed few significant associations. Still, there is a quantifiable difference between people who are more and less adept at prompt engineering. This leaves the question of what makes a competent prompt engineer, and AI literacy may still be a relevant, if not sufficient, factor in answering that question. Taken together, more evidence is needed in this area of research. Therefore, future research should build on this work with a more comprehensive prompt taxonomy, larger sample sizes, and tasks that require more prompt engineering to provide more rigorous and nuanced insights into the influences that AI literacy may have on prompt engineering with LLMs in higher education. Such research could also benefit from improved measures of AI literacy that rely on objective knowledge tests, rather than self-assessments of likely biased impressions of one's AI literacy. Getting more valid, real-world indicators of AI literacy might also be conceptually closer to actual prompt engineering behaviors, potentially revealing more about what makes certain users proper prompt engineers. To sum up, we argue for the integration of AI literacy and prompt engineering training into current curricula to enable a hybrid-intelligent society in which students can effectively utilize generative AI tools, such as ChatGPT, to enhance learning processes. While learning how to create powerful instructional prompts for AI models has the potential to enhance the practice of teaching and learning, equipping teachers and learners with AI literacy can provide them with the general competency to address the future challenges and opportunities presented by the rapid development of AI technologies and their increasing integration into our lives.

Acknowledgements and Funding

We thank Mirjam Ebersbach for reading the manuscript and for providing us with valuable feedback. We also want to thank Denise Richberg and Lukas Pieritz for very helpful coding and commenting. The results presented in this article were partially developed in the research projects Komp-HI funded by the German Federal Ministry of Education and Research (BMBF, grant 16DHBKI073) and Managing the Algorithm: Prompt Engineering for AI-based Systems as an Emerging Business Skill by the Swiss National Science Foundation (SNSF, grant number: 221281). We thank the BMBF and SNSF for supporting our research.

CRediT authorship contribution statement

Nils Knoth: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Antonia Tolzin:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andreas Janson:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Jan Marco Leimeister:** Supervision, Funding acquisition.

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

Statement: During the preparation of this work the author(s) used DEEPL Write in order to improve readability and streamline language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

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

APPENDIX A. Final Survey Instrument

Construct Information and Literature Source		Indicator	Statements
Reflection Protocol	Reflection on human-AI interaction	Reflect1a	In your view, to what extent was generative AI appropriate for handling the FIRST task?
Items developed by the authors		Reflect1b	In your view, to what extent was generative AI appropriate for addressing the SECOND task?
		Reflect2	Would you use generative AI again to handle similar tasks?
		Reflect3	Why or why not?
		Reflect4	How do you rate the quality of the AI's generated texts in terms of correctness and comprehensibility?
		Reflect5	Did the generated texts meet your expectations?
		Reflect6	How did you formulate the prompts (requests) to the system? Did you use specific strategies to get better results? Briefly describe your approach.
		Reflect7	Did you have the generative AI run queries multiple times to get different results?
		Reflect8	From your point of view, what are the advantages and disadvantages of using generative AI? Mention here the aspects that are most relevant to you.
		Reflect9	I found it difficult to write inputs/prompts for generative AI.
		Reflect10	Were there particular aspects about generative AI that helped or hindered you in your processing?
		Reflect11	I found the task contents to be complex.
		Reflect12	I have the impression that I have worked well on the tasks.
	User Experience		I found working with generative AI to be …
		UX1	… pleasant
		UX2	… motivating
		UX3	… difficult
	Previous usage of generative AI	Prev_usage1	Have you used any generative artificial intelligence before? (e.g. ChatGPT, DALL-E, Jasper, Whisper or others).
		Prev_usage2	If yes: How often do you use generative AI on average per week? Please estimate the average number of prompts.
		Prev_usage3	How comfortable are you with using generative artificial intelligence in general?
	Personal Innovativeness		How much do you agree with the following statements?
	Agarwal and Prasad (1998)	PI1	When I hear about a new technology, I want to try it out.
		PI2	From my group, I am usually the first person to try a new technology.
		PI3	In general, I tend to shy away from trying out new technologies.
		PI4	I like to test new technologies.
	Trust in generative AI	Eval1	The generative AI had the features I needed for the tasks.
	Items adapted from: Lankton et al. (2015)	Eval2	I trust the explanations and information provided by generative AI.
		Eval3	Generative AI has the capabilities to do what I want.
		Eval4	Generative AI provides competent guidance.
		Eval5	Generative AI will give me all the help I need.
		Eval6	Generative AI is very reliable.
		Eval7	Generative AI will not let me down.
AI literacy			I have knowledge of …
	AI Technology Knowledge	TK1	… of the types of technology that AI is built on.
Scale adapted from:		TK2	… of how AI technology and non-AI technology are distinct.
Pinski and Benlian (2023)		TK3	… of use cases for AI technology.
		TK4	… of the roles that AI technology can have in human-AI interaction.
			I have knowledge of …
	Human actors in AI knowledge	HK1	… of which human actors beyond programmers are involved to enable human-AI collaboration.
		HK2	… of the aspects human actors handle worse than AI.
		HK3	… of the aspects human actors handle better than AI.
		HK4	… of the human actors involved to set up and manage human-AI collaborations.
		HK5	… of the tasks that human actors can assume in human-AI collaboration.
			I have knowledge of …
	AI steps knowledge	SK1	… of the input data requirements for AI.
		SK2	… of how input data is perceived by AI.
		SK3	… of potential impacts that input data has on AI.
		SK4	… of which input data types AI can use.
		SK5	… of AI processing methods and models.
		SK6	… of how information is represented for AI processing.
		SK7	… of the risks AI processing poses.
		SK8	… of why AI processing can be described as a learning process.
		SK9	… of using AI output and interpreting it.
		SK10	… of AI output limitations.
		SK11	… of how to handle AI output.
		SK12	… of which AI outputs are obtainable with current methods.
			I have experience in …
	AI usage experience	UE1	… in interaction with different types of AI, like chatbots, visual recognition agents, etc.
		UE2	… in the usage of AI through frequent interactions in my everyday life.
			I have experience in …
	AI design experience	DE1	… in designing AI models, for example, a neural network.
		DE2	… in development of AI products.

(*continued on next page*)

(*continued*)

Construct Information and Literature Source	Indicator	Statements
AI literacy (overall)	AIL1	In general, I know the unique facets of AI and humans and their potential roles in human-AI collaboration.
	AIL2	I am knowledgeable about the steps involved in AI decision-making.
	AIL3	Considering all my experience, I am relatively proficient in the field of AI.
Demographics	Gender	Please specify your gender.
	Age	Please indicate your age.
	Study subject	Please indicate your course of study and whether you are studying for a Bachelor's or Master's degree.
	Semester count	Please indicate the number of semesters you have been studying.

APPENDIX B. Prompt Engineering Tasks

Assessment Task 1 – Trip to Andorra

Traveling can be a wonderful way to discover new places, relax and learn about new cultures. But planning a trip can often be challenging, especially if you're traveling to a new country or if you're unsure of everything you want to do and perhaps traveling alone for the first time.

In such cases, it can be helpful to turn to the assistance of chatbots. One of the most advanced chatbots is ChatGPT, an artificial intelligence chatbot that is able to have human-like conversations and handle a variety of topics.

We'll now look at whether and how ChatGPT can help you plan trips.

These sample prompts (input or instructions you type for the AI) might help you with your trip planning:

- "You are a tour guide. I'm very interested in theater in Naples, please tell me more about what places and buildings I should visit and in what order."
- "What is the cheapest destination for a 3-day city trip in Europe? My budget is around 1000 euros."
- "List me free museums in Amsterdam. I am primarily interested in modern art."

Your task now is to plan a 4-day trip to Andorra in September. Whether you travel alone or with others, where you stay, whether you travel around, what activities you do, etc., are entirely up to you. Please plan your trip as concretely as possible.

However, avoid "unnecessary" personal contributions in the form of your own formulations. Try to create the itinerary as "automated" as possible using (almost exclusively) the chatbot.

You have 7 min for the task "Travel to Andorra".

At this point, please wait until the experimenters let you know so that everyone can start working on this task together at the same time.

Click here to start the AI.

Assessment Task 2 – Project planning with AI

During your studies you will always be confronted with the challenge of setting up your own research project. At the latest, the bachelor's or master's thesis confronts you with the task of coming up with your own research question and ways to investigate it.

In such cases, it can be helpful to resort to the support of chatbots. One of the most advanced chatbots is ChatGPT, an artificial intelligence chatbot capable of having human-like conversations and covering a variety of topics.

We're now going to look at whether and how ChatGPT can help you plan a science project.

Your task is to plan 3 important aspects of a research project together with Artificial Intelligence. For our fictional example, you'll investigate the topic of "Automated Essay Scoring".

The 3 aspects to work on are:

1. Introduction to the topic and definition: what is meant by "Automated Essay Scoring"?
2. Developing a research question: brainstorming phase - what are the different research questions that could be explored in this area?
3. Creation of a project plan (incl. time schedule): What steps need to be worked on and when to complete the scholarly project?

The research questions you finally decide on and the methods you use to investigate them are entirely up to you. However, please plan your research project as concretely and meaningfully as possible.

However, avoid "unnecessary" personal contributions in the form of your own formulations. Try to create the project plan as "automated" as possible using (almost exclusively) the chatbot.

You have 10 min time for this task "Project plan - scientific work".

At this point, please wait until the investigators let you know so that everyone can start working on this task together at the same time.

Click here to start the AI.

APPENDIX C. Prompt Components (Eager & Brunton, 2023) – Potential Industry Use Case: Customer Support Automation (developed by the authors)

Component	Purpose
Verb: "Resolve"	Initiates the action of finding a solution to customer queries or issues.
Focus: "Customer queries"	Specifies that the action is centered around addressing customer questions or problems.
Context: "Within the online ticketing system"	Sets the boundary that the task is to be performed within a specific digital platform, providing clarity on where the action takes place.
Focus and Condition: "Personalization and Efficiency"	Personalization involves using customer data to provide responses that are relevant to the individual's history, preferences, and specific issues. Efficiency ensures that these personalized responses are delivered promptly, optimizing customer satisfaction and operational productivity.
Alignment: "Brand guidelines and customer satisfaction"	Ensures that the responses are not only accurate and timely but also consistent with the company's brand voice and aimed at enhancing customer satisfaction.
Constraints and Limitations: "Do not disclose personal information"	Sets a boundary on privacy, ensuring the AI does not overstep regulatory or ethical lines.

APPENDIX D. Prompt Components (Eager & Brunton, 2023) – Potential Higher Education Use Case: Research Assistance (developed by the authors)

Component	Purpose
Verb: "Analyze"	Directs the action towards examining or interpreting a specific set of data or information.
Focus: "Scholarly articles"	Identifies the main subject matter to be analyzed, focusing the task on academic content.
Context: "Within the field of renewable energy"	Narrows down the area of study, providing specificity to the research task.
Focus and Condition: "Latest trends and technologies"	Clarifies that the output should not only relate to renewable energy but specifically to the most recent advancements and innovations in the field.
Alignment: "Course objectives and learning outcomes"	Ensures that the analysis contributes to the educational goals of the course, aligning the AI's output with the curriculum.
Constraints and Limitations: "Use only peer-reviewed sources"	Imposes a quality filter on the information to be analyzed, ensuring reliability and academic standards are met.

References

Agarwal, R., & Prasad, J. (1998). A conceptual and operational definition of personal innovativeness in the domain of information technology. *Information Systems Research, 9*(2), 204–215. https://doi.org/10.1287/isre.9.2.204

Alves de Castro, C. (2023). A discussion about the impact of ChatGPT in education: Benefits and concerns. *Journal of Business Theory and Practice, 11*(2). https://doi.org/10.22158/jbtp.v11n2p28

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: A theoretical framework of delegation to and from agentic IS artifacts. *MIS Quarterly, 45*(1), 315–341.

Baker-Brown, G., Ballard, E. J., Bluck, S., De Varies, B., Suedfeld, P., & Tetlock, P. E. (1992). In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic contentThe conceptual/integrative complexity scoring manual* (pp. 401–418). Cambridge University Press.

Berg, J., Raj, M., & Seamans, R. (2023). Capturing value from artificial intelligence. *Academy of Management Discoveries, 9*(4), 424–428.

Betz, G., Richardson, K., & Voigt, C. (2021). *Thinking aloud: Dynamic context generation improves zero-shot reasoning performance of GPT-2*. https://arxiv.org/pdf/2103.13033.

Bewersdorff, A., Zhai, X., Roberts, J., & Nerdel, C. (2023). Myths, mis- and preconceptions of artificial intelligence: A review of the literature. *Computers and Education: Artificial Intelligence, 4*, Article 100143. https://doi.org/10.1016/j.caeai.2023.100143

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, et al. (2021). On the opportunities and risks of foundation models. https://arxiv.org/pdf/2108.07258.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), Article 205395171562251. https://doi.org/10.1177/2053951715622512

Byom, L. J., & Mutlu, B. (2013). Theory of mind: Mechanisms, methods, and new directions. *Frontiers in Human Neuroscience, 7*.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., et al. (2023). *A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT*. https://arxiv.org/pdf/2303.04226.

Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. B. (2023). ChatGPT goes to Law school. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4335905. Advance online publication.

Crawford, J., Cowling, M., & Allen, K.-A. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching and Learning Practice, 20*(3). https://doi.org/10.53761/1.20.3.02

Dang, H., Benharrak, K., Lehmann, F., & Buschek, D. (2022). In M. Agrawala, J. O. Wobbrock, E. Adar, & V. Setlur (Eds.), *ACM symposium on user interface software and technologyBeyond text generation: Supporting writers with continuous automatic text summaries*. ACM.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. https://arxiv.org/pdf/2209.01390.

Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2023). Impact of AI assistance on student agency. *Computers & Education, 104967*. https://doi.org/10.1016/j.compedu.2023.104967

Day, T. (2023). A preliminary investigation of fake peer-reviewed citations and references generated by ChatGPT. *The Professional Geographer*, 1–4. https://doi.org/10.1080/00330124.2023.2190373

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., et al. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4573321. Advance online publication.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *BISE, 61*(5), 637–643.

Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. In *Artificial intelligence: Foundations, theory, and algorithms ser.* Springer.

Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H.-T., et al. (2021). *OpenPrompt: An open-source framework for prompt-learning*. https://arxiv.org/pdf/2111.01998.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., et al. (2023). So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities,

challenges and implications of generative conversational AI for research, practice and policy. *IJIM, 71*.
Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching and Learning Practice, 20*(5). https://doi.org/10.53761/1.20.5.02
Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*, 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153
Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *ISR*, 678–696.
Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education & Teaching International*, 1–15. https://doi.org/10.1080/14703297.2023.2195846
Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines, 30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1
Gašević, D., Siemens, G., & Sadiq, S. (2023). Empowering learners for the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 4*, Article 100130. https://doi.org/10.1016/j.caeai.2023.100130
Ghallab, M. (2019). Responsible AI: Requirements and challenges. *AI Perspectives, 1*(1), 1–7. https://aiperspectives.springeropen.com/articles/10.1186/s42467-019-0003-z.
Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research. *Organizational Research Methods, 16*(1), 15–31. https://doi.org/10.1177/1094428112452151
Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior, 97*, 304–316. https://doi.org/10.1016/j.chb.2019.01.020
Gregory, R. W., Kaganer, E., Henfridsson, O., & Ruch, T. J. (2018). IT consumerization and the transformation of IT governance. *MIS Quarterly, 42*, 1225–1254. Article 4.
Gunness, A., Matanda, M. J., & Rajaguru, R. (2023). Effect of student responsiveness to instructional innovation on student engagement in semi-synchronous online learning environments: The mediating role of personal technological innovativeness and perceived usefulness. *Computers & Education, 205*, Article 104884. https://doi.org/10.1016/j.compedu.2023.104884
Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., et al. (2021). Pre-trained models: Past, present and future. *AI Open, 2*.
Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior, 49*, 245–250. https://doi.org/10.1016/j.chb.2015.02.026
Hou, Y., Dong, H., Wang, X., Li, B., & Che, W. (2022). *MetaPrompting: Learning to learn better prompts*. https://arxiv.org/pdf/2209.11486.
Janson, A. (2023). How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2023.107954
Janson, A., Söllner, M., & Leimeister, J. M. (2020). Ladders for Learning: Is Scaffolding the Key to Teaching Problem-Solving in Technology-Mediated Learning Contexts? *Academy of Management Learning & Education, 19*(4), 439–468. https://doi.org/10.5465/amle.2018.0078
Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38.
Joksimovic, S., Ifenthaler, D., Marrone, R., Laat, M. de, & Siemens, G. (2023). Opportunities of artificial intelligence for supporting complex problem-solving: Findings from a scoping review. *Computers and Education: Artificial Intelligence, 4*, Article 100138. https://doi.org/10.1016/j.caeai.2023.100138
Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., & Huber, P. (2016). Artificial intelligence and computer science in education: From kindergarten to university. In *2016 IEEE frontiers in education conference (FIE)* (pp. 1–9). IEEE. https://doi.org/10.1109/FIE.2016.7757570
Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. https://doi.org/10.35542/osf.io/5er8f.
Kikalishvili, S. (2023). Unlocking the potential of GPT-3 in education: Opportunities, limitations, and recommendations for effective integration. *Interactive Learning Environments*, 1–13. https://doi.org/10.1080/10494820.2023.2220401
Kohnke, L., Moorhouse, B. L., & Zou, Di (2023). Exploring generative artificial intelligence preparedness among university language instructors: A case study. *Computers and Education: Artificial Intelligence, 5*, Article 100156. https://doi.org/10.1016/j.caeai.2023.100156
Krämer, N., & Manzeschke, A. (2021). Social reactions to socially interactive agents and their ethical implications. In *Lugrin et al 2021 Introduction to Socially Interactive Agents*.
Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: A review of the research. *Computers in Human Behavior, 19*(3), 335–353. https://doi.org/10.1016/S0747-5632(02)00057-2
Lai, C. Y., Cheung, K. Y., & Chan, C. S. (2023). Exploring the role of intrinsic motivation in ChatGPT adoption to support active learning: An extension of the technology acceptance model. *Computers and Education: Artificial Intelligence, 5*, Article 100178. https://doi.org/10.1016/j.caeai.2023.100178
Lankton, N., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems, 16*(10), 880–918. https://doi.org/10.17705/1jais.00411
Laupichler, M. C., Aster, A., & Raupach, T. (2023). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence, 4*, Article 100126. https://doi.org/10.1016/j.caeai.2023.100126
Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *CE AI, 3*.
LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852. https://doi.org/10.1177/1094428106296642
Lee, H. (2023). The rise of ChatGPT: Exploring its potential in medical education. In *Anatomical sciences education*. https://doi.org/10.1002/ase.2270. Advance online publication.
Lin, L., Ginns, P., Wang, T., & Zhang, P. (2020). Using a pedagogical agent to deliver conversational style instruction: What benefits can you obtain? *Computers & Education, 143*, Article 103658. https://doi.org/10.1016/j.compedu.2019.103658
Liu, V., & Chilton, L. B. (2021). Design guidelines for prompt engineering text-to-image generative models. https://arxiv.org/pdf/2109.06977.
Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35.
Long, D., & Magerko, B. (2020). In R. Bernhaupt, F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, et al. (Eds.), *CHI 2020 proceedingsWhat is AI literacy? Competencies and design considerations* (pp. 1–16). ACM.
McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. http://arxiv.org/pdf/2309.13638v1.
McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior, 99*, 28–37. https://doi.org/10.1016/j.chb.2019.05.009
Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., et al. (2023). *Augmented Language models: A survey*. https://arxiv.org/pdf/2302.07842.
Munnukka, J., Talvitie-Lamberg, K., & Maity, D. (2022). Anthropomorphism and social presence in Human–Virtual service assistant interactions: The role of dialog length and attitudes. *Computers in Human Behavior, 135*, Article 107343. https://doi.org/10.1016/j.chb.2022.107343
Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153
Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI* (pp. 72–78).
Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). *Conceptualizing AI literacy: An exploratory review*. CE AI, 2.
Nguyen, H. (2023). Role design considerations of conversational agents to facilitate discussion and systems thinking. *Computers & Education, 192*, Article 104661. https://doi.org/10.1016/j.compedu.2022.104661
Nguyen, Q. N., Sidorova, A., & Torres, R. (2022). User interactions with chatbot interfaces vs. Menu-based interfaces: An empirical study. *Computers in Human Behavior, 128*, Article 107093. https://doi.org/10.1016/j.chb.2021.107093
Nißen, M., Selimi, D., Janssen, A., Cardona, D. R., Breitner, M. H., Kowatsch, T., et al. (2022). See you soon again, chatbot? A design taxonomy to characterize user-chatbot relationships with different time horizons. *Computers in Human Behavior, 127*, Article 107043. https://doi.org/10.1016/j.chb.2021.107043
OpenAI. (2023). *ChatGPT (Mar 14 version)*. https://chat.openai.com/chat.
Oppenlaender, J. (2022). A taxonomy of prompt modifiers for text-to-image generation. https://arxiv.org/pdf/2204.13988.
Oppenlaender, J., Linder, R., & Silvennoinen, J. (2023). *Prompting AI art: An investigation into the creative skill of prompt engineering*. https://arxiv.org/pdf/2303.13534.
Pinski, M., Adam, M., & Benlian, A. (2023). In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, et al. (Eds.), *Proceedings of the 2023 CHI conference on human factors in computing systemsAI knowledge: Improving AI delegation through human enablement* (pp. 1–17). ACM. https://doi.org/10.1145/3544548.3580794.
Pinski, M., & Benlian, A. (2023). AI literacy - towards measuring human competency in artificial intelligence. In T. Bui (Chair). *HICSS, 56*, 165–174.
Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences, 13*(9), Article 5783. https://doi.org/10.3390/app13095783
Rasul, T., Nair, S., Kalendra, D., Robin, M., Oliveira Santini, F. de, Ladeira, W. J., et al. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching, 6*(1).
Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning & Teaching, 6*(1). https://doi.org/10.37074/jalt.2023.6.1.23
Ruwe, T., & Mayweg-Paus, E. (2023). "Your argumentation is good", says the AI vs humans – the role of feedback providers and personalised language for feedback effectiveness. In *Computers and education: Artificial intelligence* (Vol. 5), Article 100189. https://doi.org/10.1016/j.caeai.2023.100189
Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher, 20*(3), 2–9.
Schmitt, A., Zierau, N., Janson, A., & Leimeister, J. M. (2023). The role of AI-based artifacts' voice capabilities for agency attribution. *Journal of the Association for Information Systems, 24*(4), 980–1004. https://doi.org/10.17705/1jais.00827
Schöbel, S., Schmitt, A., Benner, D., Saqr, M., Janson, A., & Leimeister, J. M. (2024). Charting the evolution and future of conversational agents: a research agenda along five waves and new frontiers. *Information Systems Frontiers, 26*, 729–754. https://doi.org/10.1007/s10796-023-10375-9 (2024).
Schuetz, S., & Venkatesh, V. (2020). *Research perspectives: The rise of human machines: How cognitive computing systems challenge assumptions of user-system interaction*. JAIS, 460/82.

Seeber, I., Bittner, E., Briggs, R. O., Vreede, T. de, Vreede, G.-J. de, Elkins, A., et al. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management, 57*(2), Article 103174. https://doi.org/10.1016/j.im.2019.103174

Siemon, D., Elshan, E., Vreede, T. de, Oeste-Reiß, S., Vreede, G. J. de, & Ebel, P. (2022). *Examining the antecedents of creative collaboration with an AI teammate*. ICIS.

Sikström, P., Valentini, C., Sivunen, A., & Kärkkäinen, T. (2022). How pedagogical agents communicate with students: A two-phase systematic review. *Computers & Education, 188*, Article 104564. https://doi.org/10.1016/j.compedu.2022.104564

Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. *ECNU Review of Education, 6*(3), 355–366. https://doi.org/10.1177/20965311231168423

Suedfeld, P., Tetlock, P. E., & Streufert, S. (1992). In C. P. Smith (Ed.), *Motivation and personality: Handbook of thematic contentConceptual/integrative complexity* (pp. 393–400). Cambridge University Press.

Tan, S. C., Lee, A. V. Y., & Lee, M. (2022). A systematic review of artificial intelligence techniques for collaborative learning over the past two decades. *Computers and Education: Artificial Intelligence, 3*, Article 100097. https://doi.org/10.1016/j.caeai.2022.100097

Tarafdar, M., Page, X., & Marabelli, M. (2023). Algorithms as co-workers: Human algorithm role interactions in algorithmic work. *Information Systems Journal, 33*(2), 232–267.

Tolzin, A., & Janson, A.[A. (2023). Mechanisms of common ground in human-agent interaction: A systematic review of conversational agent research. In *Hawaii international conference on system sciences (HICSS)*.

Tolzin, A., Knoth, N., & Janson, A. (2024). Worked Examples to Facilitate the Development of Prompt Engineering Skills. In *In: ECIS 2024 Proceedings*.

Tully, S., Longoni, C., & Appel, G. (2023). *Knowledge of artificial intelligence predicts lower AI receptivity*. https://doi.org/10.31234/osf.io/t9u8g

Venkatesh, V., Brown, S., & Sullivan, Y. (2016). Guidelines for conducting mixed-methods research: An extension and illustration. *Journal of the Association for Information Systems, 17*(7), 435–494. https://doi.org/10.17705/1jais.00433

Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *Digcomp 2.2, the Digital Competence framework for citizens: With new examples of knowledge, skills and attitudes*. Publications Office European Union. https://doi.org/10.2760/115376

Wang, S., Yu, M., & Huang, L. (2022). The art of prompting: Event detection based on type specific prompts. https://arxiv.org/pdf/2204.07241.pdf.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-Thought prompting elicits reasoning in large language models. https://arxiv.org/pdf/2201.11903.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., et al. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. http://arxiv.org/pdf/2302.11382v1.

Wienrich, C., & Carolus, A. (2021). Development of an instrument to measure conceptualizations and competencies about conversational agents on the example of smart speakers. *Frontiers of Computer Science*, Article 685277.

Wu, T., Terry, M., & Cai, C. J. (2021). AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. https://arxiv.org/pdf/2110.01691.

Yilmaz, R., & Karaoglan Yilmaz, F. G. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence, 4*, Article 100147. https://doi.org/10.1016/j.caeai.2023.100147

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings CHI 2023* (pp. 1–21).

Zhu, C., Sun, M., Luo, J., Li, T., & Wang, M. (2023). How to harness the potential of ChatGPT in education? *Knowledge Management & E-Learning: International Journal, 15*(2), 133–152. https://doi.org/10.34105/j.kmel.2023.15.008