When Young Scholars Cooperate with LLMs in Academic Tasks: The Influence of Individual Differences and Task Complexities

Jiyao Wang[1], Chunxi Huang[2], Song Yan[1], Weiyin Xie[1], Dengbo He[1,3,4 *]

1. Thrust of Robotics and Autonomous Systems, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
2. Interdisciplinary Programs Office, The Hong Kong University of Science and Technology, Hong Kong SAR, China
3. Thrust of Intelligent Transportation, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
4. Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

Wang J. : jwanggo@connect.ust.hk; ORCID: 0000-0002-0743-0121
Hunag C.: tracy.huang@connect.ust.hk
Yan S.: syan931@connect.hkust-gz.edu.cn
Xie W.: wxie593@connect.hkust-gz.edu.cn
He D.: dengbohe@hkust-gz.edu.cn; ORCID: 0000-0003-4359-4083

*Corresponding author.

Abstract

As a novel AI-powered conversational system, large language models (LLMs) have the potential to be used in various applications. Recent advances in LLMs like ChatGPT have made LLM-based academic tools possible. However, most of the existing studies on the adoption of LLM for academic tasks were based on theoretical or qualitative analyses, which failed to provide empirical evidence on the effects of LLMs on users' behaviors. Additionally, although previous work has investigated users' acceptance of conventional conversational systems, little is known about how scholars evaluate LLMs when they are used for academic tasks. Hence, we conducted an empirical field experiment to assess the performance of 48 early-stage scholars on two core academic activities (paper reading and literature reviews) under varying time constraints. Prior to the tasks, participants underwent different training programs about LLM capabilities and limitations. Then, we built a hierarchy dependency network using the Bayesian network. Statistical regression analyses were further conducted to quantify relationships among influential factors of task performance and users' attitudes toward the LLMs. It was found that young scholars have upheld relatively high academic integrity when using LLMs for academic tasks, and user-LLM performance varied with the task type and time pressure but not with the type of training we used. Further, scholars' traits can also affect their performance in academic tasks and attitudes towards the LLMs. This work can inspire the future development of LLM-related user training and guide the optimization of LLMs.

Keywords: Large language model; Academic tasks; Human-AI collaboration; User attitudes; Empirical study

1. Introduction

Large language models (LLMs), as advanced conversational systems enabled by artificial intelligence (AI), have been promoted worldwide over the past years. The impressive performance of LLMs in complex content understanding and human-like text generation attracted increasing attention from industries and researchers (Mogavi et al., 2023). Some well-known commercial LLM-based products (e.g., ChatGPT[1]; Claude[2]) have been released in recent years. Being different from previous natural language models (Devlin et al., 2018; Yang et al., 2019), as one of the neural network systems, LLMs have a more complex internal structure and training process over massive bodies of text, which empowers them the ability to handle a wide range of topics and enables continuous learning. Thus, in recent years, the LLM has been widely adopted in healthcare (Alberts et al., 2023; Kung et al., 2023), education (Jungherr, 2023; Kasneci et al., 2023), and creative writing (Gero et al., 2023), and has brought in profound changes of the workflow in those domains.

As a field that requires intensive intelligence investment, academic work may also benefit from the LLMs. For example, Matthew (2022) and Dis et al. (2023) pointed out that the ChatGPT and other similar LLMs have already been utilized by researchers for a range of tasks, including generating essays, condensing literature reviews, composing and polishing academic writing, and even identifying areas of research that need attention. However, it should be noted that, the academic tasks are more special compared to tasks in other domains. Specifically, academic tasks usually require substantial training in skills such as information acquisition, evaluation, and synthesis (Luccioni & Viviano, 2021), and the academic community upholds rigorous standards for logical consistency, the accuracy of the information, and originality of ideas (Bommasani et al., 2021) – the LLMs can hardly meet these criteria (Gordijn & Have, 2023). Thus, understanding the limitations of LLMs in academic tasks is urgent to facilitate appropriate usage of the LLMs. For example, although the LLMs were believed to bring benefits to scholars (e.g., alleviating time pressure of users in academic tasks, Dergaa et al., 2023), previous research found that the lack of synthesis and the risk of plagiarism still exist when conducting

[1] https://openai.com/chatgpt
[2] https://www.claude.co.id/

a literature review with LLMs (Aydın & Karaarslan, 2022), and the abstracts done by LLMs are still easier to be distinguished from human works (Gao et al., 2022). Further, as emerging conversational systems, having an interface that meets social norms and user expectations is essential to the successful promotion of LLMs among new users (Brandtzaeg & Følstad, 2018).

Considering that LLMs can be regarded as automation assisting users in performing tasks, the factors influencing users' performance in human-automation cooperation may affect the performance of the user-LLM system. For example, given that drivers with different mental models of driving automation may hold different attitudes to the automation (Huang et al., 2023), LLM users may also take different strategies to work with the LLM in different academic tasks. On the other hand, users' attitudes towards automation (e.g., trust (Hoff & Bashir, 2015), and acceptance of the technology (Davis, 1989; Ghazizadeh et al., 2012)) and their personalities (Wang, Huang, et al., 2023) may also influence users' reliance on the system, which can be further moderated by task complexities (Bailey & Scerbo, 2007; Lyell et al., 2018). Thus, it is necessary to understand, when performing academic tasks with LLMs, how their strategies and performance are affected by task type (TT), users' attitudes towards the LLMs (as moderated by user training (Sauer et al., 2016) or prior knowledge of users (Hergeth et al., 2017)), and the task complexity (as moderated by time pressure (TiP)).

Hence, in this study, an experiment was conducted to quantify the time pressure and training on new users' performance when using LLMs for two types of academic tasks, i.e., paper understanding (PU), in which one needs to extract essential information from a given paper (i.e., retrieving unknown information from a known source); and literature review (LR), which requires one to identify targeted partially known information when the information source is unknown. Given that less comprehensive information needs to be extracted from each literature in the LR task compared to that in PU tasks, we assume that the complexities of the PU and LR task should be comparable but different skills are needed in these two tasks. Further, in the domain of human-automation interaction, the research found that previous experience with the task can affect users' strategies when cooperating with the automation (He et al., 2022) or AI teammates (Zhang et al., 2023), and novice users may exhibit less appropriate reliance on the system (He & Donmez, 2019) and high uncertainty in terms of strategies (He et al., 2022). Thus, this study targeted junior graduate students, given that these young scholars may not have developed

matured strategies when handling academic tasks compared to senior scholars and the young generations are usually earlier adopters of new technologies (Broady et al., 2010).

In general, through an experiment, this study aims to focus on the following three research questions (RQs): **RQ1**: what factors (including task difficulty as moderated by task type and time pressure, and characteristics of the users) can influence young scholars' performance in conducting academic tasks with LLMs; **RQ2**: whether providing training regarding LLM limitations can influence users' behaviors when using LLMs; **RQ3**: what factors can influence young scholars' attitudes towards LLMs. Specifically, for RQ3, we adopted two theoretical frameworks, the Technology Acceptance Model (TAM) (Davis, 1989) and Automation Acceptance Model (AAM) (Ghazizadeh et al., 2012), which can conceptualize one's acceptance of and attitude toward LLMs. Inspired by relevant research in the human-automation interaction domain, we expanded the TAM by considering some external variables, including users' personality traits (Chien et al., 2016; Cho et al., 2016), trust propensity (Merritt et al., 2013), experience with the system (Dishaw & Strong, 1998), domain knowledge (Hoff & Bashir, 2015), users' trust in the system (Ghazizadeh et al., 2012), and workload (Longo, 2018; Schmutz et al., 2009) during the task.

2. Related Work

2.1 Large Language Model in Academic Tasks

Over the previous five years, a notable transformation has been observed in the field of natural language technologies, primarily due to the emergence of LLMs (Dong et al., 2019). LLMs such as ChatGPT have gained increasing interest in academia due to their ability to generate human-like text by learning from internet datasets (Rahman et al., 2023). Researchers have explored the applications of LLMs on academic tasks, including drafting manuscripts, reviewing papers, and making editorial decisions (Matthew, 2022; Stokel-Walker, 2023). Some studies have shown the potential of using LLMs to generate ideas, synthesize literature, and create testing frameworks (Dowling & Lucey, 2023). However, LLMs still fall short of producing publishable scientific articles compared to skilled researchers (Gordijn & Have, 2023). Though the capabilities of LLMs are expected to improve (Liebrenz et al., 2023), limitations still persist, including subpar synthesis skills, risk of plagiarism,

and deficient abstract writing. For example, Zhu et al., (2023) argued that the current LLM is useful but still generates occasionally unprecise electronic encyclopedia. Existing survey- or interview-based studies may not fully reveal how scholars use LLMs and quantify their impact on academic performance. It is also worth noting that most existing research has concentrated on the attitudes and opinions of senior researchers toward the use of LLM technology in academic tasks (e.g., Morris, 2023). However, little research has focused on younger scholars who may be more receptive to new technologies and may lack the necessary expertise to supervise LLMs in academic tasks (Dis et al., 2023). Thus, empirical studies that consider the LLMs and users as an integrated system are urgent to guide the usage and optimization of the LLMs.

2.2 Users' Evaluation of Conversational Systems

Conversational systems represent a distinct category of AI-powered information systems due to their ability to facilitate interactive conversations through written or spoken language (Pfeuffer et al., 2019; Rubin et al., 2010). Conversational systems can identify and address user intentions effectively (Shawar & Atwell, 2005). Previous studies mainly focused on evaluating the algorithms underlying conversational systems. However, as an AI-powered assists, the performance of the conversational system should be assessed from the user-system integrated perspective of view, and the users' perceptions of the system should be considered (Herlocker et al., 2004; Jannach & Bauer, 2020; Shani & Gunawardana, 2011). For example, trust, defined as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*" (J. D. Lee & See, 2004), has been identified as an influential factor in users' sustainable adoption and usage of automation (Mcknight et al., 2011; Turel & Gefen, 2013) and has been found to boost users' intentions to utilize and accept AI-based systems (Glikson & Woolley, 2020). Similarly, users' perceived interaction quality (Cai et al., 2022; Walker et al., 1997), reliability (Belda-Medina & Calvo-Ferrer, 2022) and usability (Guerino et al., 2021; Guerino & Valentim, 2020) of the AI-based system were also related with the adoption of the conversational systems (Almahri et al., 2019; Belda-Medina & Calvo-Ferrer, 2022; Denecke & May, 2022). Some other studies adopted classic theoretical frameworks (e.g., TAM (Davis, 1989) and Software Usability Measurement Inventory (SUMI) (Kiraskowski & Corbett, 1993)) to model users' adoption of the conversational systems (Pu et al.,

2011; Radziwill & Benton, 2017). However, previous research mostly focused on the adoption of conversational systems in non-professional fields (e.g., music recommendation (Cai et al., 2022)) and conversational systems have evolved dramatically in past years. Thus, evaluating the adoption of the most up-to-date LLMs in academic tasks is necessary and urgent.

3. Methodology

3.1 Participants

In total, 48 participants (30 males and 18 females) were recruited for the experiment, with varying academic backgrounds. All participants were required to be below or equal to 30 years old, as the younger generation tends to exhibit greater acceptance of emerging technologies (Broady et al., 2010). We required the participants to have limited exposure to LLMs and should self-report to "*sometimes use LLMs for academic purposes*" or less. This is to ensure that participants were new users of LLMs so that they had not developed maturated strategies in using LLM and could be more easily influenced by our training method. Given that this study targets young scholars, all participants were required to be students or research assistants without a Ph.D. degree and were affiliated with research institutions or universities, where English was the primary language of instruction. Recruitment was carried out through online and on-campus posters. Table 1 summarizes the demographic information of the participants. All participants who completed the experiment received a compensation of 120 Chinese Yuan. The study received ethical approval from the Human and Artefacts Research Ethics Committee [HREP-2023-0159] at the Hong Kong University of Science and Technology (Guangzhou).

3.2 Academic Tasks and Training

Two types of academic tasks were used in this study, including paper understanding (PU) and literature review (LR). In the PU task, participants were given a scientific paper and were required to answer five questions related to the provided paper. In the LR task, participants were provided with a topic and instructed to complete a literature review of approximately 500 words on the given topic. Considering the diversity of participants' academic backgrounds, we selected publications and reviewed topics from a field that did not overlap any of the participants' research interests, i.e., human factors in transportation. Another rationale behind this choice was that the human factors domain has

long been considered “common sense” (Stevens, Horrock, 2019). Though this may not be the fact, it indicates that it should be relatively easy for laymen to understand the research in this domain. Given that task complexity can moderate the relationship between users' trust in and reliance on the system (Parasuraman & Riley, 1997), and considering the prevalence of time pressure in academia, we implemented two levels of time constraints for the tasks: 10 minutes (ST) and 20 minutes (LT). These time limits were determined based on users' feedback in pilot tests.

In order to avoid learning effects, different articles and topics were used in different trials of the same user. Specifically, in PU tasks, two articles were selected, i.e., P1 (Choy et al., 2022) and P2 (Hickerson & Lee, 2022) and the corresponding questions are provided in Appendix. In the LR tasks, the two topics were ‘Novice driver training’ (T1) and ‘Hazard perception in driving’ (T2). Given that each article or topic can be assigned to ST condition or LT condition, we ended up having 8 combinations of the experimental conditions (i.e., for PU task: P1-ST, P2-ST, P1-LT, P2-LT; for LR task:, T1- ST, T2-ST, T1-LT, T2-LT).

In addition, given that training has been found to be an effective approach to optimizing user-automation interaction, we also controlled the level of training (LeT) a participant received. Specifically, all participants received basic training on how to use LLMs, such as operating the interface. At the same time, a limitation-based training was provided to half of the participants, which emphasized the limitations and potential errors associated with the LLMs (see Appendix); while the other half of the participants did not receive this limitation-based training but only the basic training. All training materials were delivered through pre-recorded videos.

3.1 Experiment Design

A mixed design was adopted for this study, with Task Type (Paper understanding vs. Literature review) and Time Pressure (LT vs. ST) as within-subject factors, and Level of Training (With vs. without limitation-based training) as the between-subjects factor. In other words, each participant needed to complete four academic tasks (two PU tasks and two LR tasks). Given that we have 8 combinations of the experimental conditions, we pre-selected the experimental conditions to make sure that each participant would complete two PU tasks (one with ST and one with LT) and two LR tasks (one with ST and one with LT), and each article (i.e., P1 or P2) and topic (i.e., T1 and T2) were

equally used among all participants. The within-subject factors (i.e., Task Type and Time Pressure) were counterbalanced, leading to 24 orders (A_4^4) and 48 participants (2 levels of training * 24 orders).

Table 1. Demographic information of the participants in different conditions.

Background	**Type**	**Level of training**	
		With limitation-based training	Without limitation-based training
Gender	Male	15	15
	Female	9	9
Age	Years	Mean: 25.5 (SD: 1.5, min: 22, max: 28)	Mean: 25.8 (SD: 1.8, min: 22, max: 27)
Experience with LLM (ExLLM)	Never used	3	3
	Rarely used	9	9
	Sometimes used	12	12
Academic Role	Ph.D. student	11	11
	MPhil student	8	9
	Research assistant	5	4
Number of publications (N.Pub)	0	3	7
	1-3	19	16
	4-6	2	1

3.2 Apparatus

The ChatGPT, a commonly used LLM tool that utilizes advanced language technology, was adopted in the experiment. The ChatGPT was selected as it was widely known, and to the best of our knowledge, there was no other LLM tool that is available to the general public and is with comparable performance to ChatGPT at the time of the study (early 2023). In order to ensure fairness, the use of other LLMs was restricted, and only the official ChatGPT interface was allowed. To provide access to ChatGPT, a virtual machine (VM) was set up on Microsoft Azure. This VM was equipped with pre-installed Google Chrome[3] and Microsoft Office Packages. To simulate real-world scenarios, participants were allowed to use Chrome and ChatGPT during the tasks on a voluntary basis, whenever they deemed it necessary for the tasks. For the paper reading task, the users could copy the content in the provided paper to the ChatGPT to obtain the key information from the paper. For the literature review task, the users could use prompts to ask the ChatGPT to search the information online. All experiments took place in the same meeting room with minimal external disruptions.

[3] https://www.google.cn/chrome/index.html

3.1 Procedures

As shown in Fig.1, upon arrival, the participant's informed consent was obtained. Then, all participants received basic training (around 10 minutes) regarding how to use the LLM. Next, a pre-study questionnaire was issued to collect data on user-related factors, including users' personality based on the Ten Item Personality Inventory (TIPI) (Gosling et al., 2003), trust propensity based on Propensity to Trust scale developed by (Merritt et al., 2013) and users' domain knowledge of LLMs based on a self-designed questionnaire consisted of five multiple-choice questions (see Appendix). The personality was measured given that they have been found to inherently affect users' trust and intention of using automation (Cai et al., 2022); the trust propensity can influence users' trust in the system from the dispositional trust perspective of view (Merritt et al., 2013) and the domain knowledge has been found to be associated with users' trust in automation (Hoff & Bashir, 2015). Next, participants who were assigned to limitation-based training received additional training regarding the limitation of the LLM. Following that, all participants finished four academic tasks in the order they were assigned to. All participants were allowed to use the LLM to assist in their task and it ended up that all participants used the LLM for all tasks in the experiment. All participants were told that their compensation would be decided by their performance in the tasks.

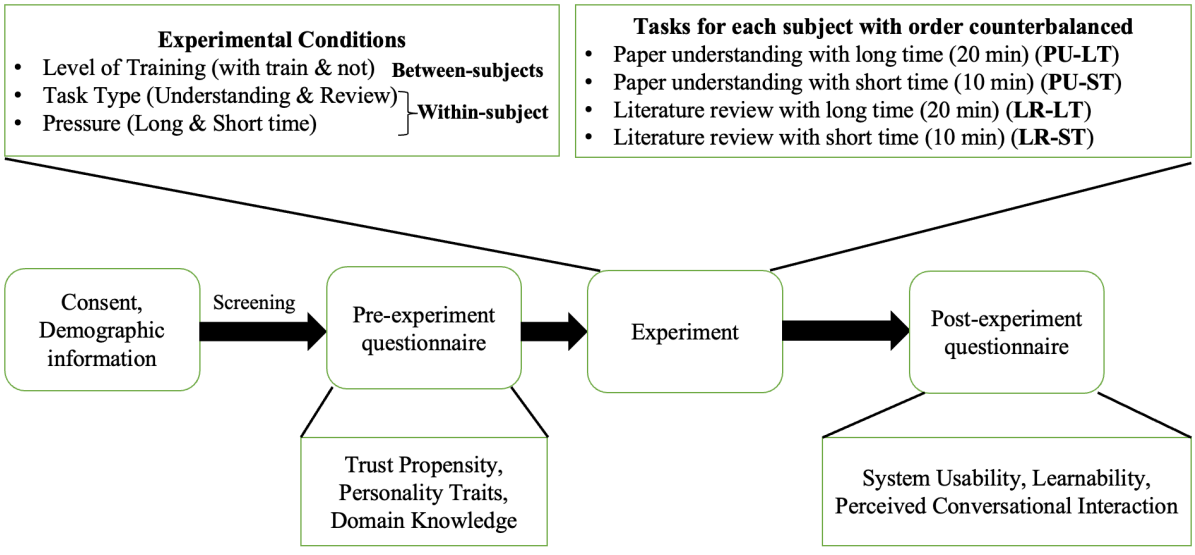

Fig.1. The overall experiment framework.

After participants finished each task, a within-experiment questionnaire was provided. In the within-experiment questionnaire, users' trust in the LLM was measured using the five-item facets of

trustworthiness scale (FIFT) (Franke et al., 2015), and users' perceived workload in the task was measured using the NASA task load index (NASA TLX) (Hart & Staveland, 1988). After the participant finished all tasks, the post-experiment questionnaire was issued to collect the participant's evaluation of and attitudes to the LLM tool. Specifically, we adopted the System Usability Scale (SUS) (Brooke, 2013) that measures the perceived usability and learnability of the system, as they are both considered to be the predictors of the user's technology acceptance (Holden & Rada, 2011; Scholtz et al., 2016). Further, considering that LLM is a conversation-based interaction agent, the Perceived Conversational Interaction score was obtained (Cai et al., 2022; Walker et al., 1997), which assesses how well the LLMs communicate with the users in the tasks. At the end of the experiment, all participants received the full compensation regardless of their task performance. The entire experiment took around two hours.

3.4 Variable Extraction

3.4.1 Task-performance-based variable

The users' time spent on each task (i.e., Time Spent) was recorded directly in the experiment, ranging from 0 to the time limits of each task (e.g., 10 minutes in LR-ST tasks). To better quantize this factor, we transformed the Time Spent to the percentage of the time limit (i.e., Time Percent). For example, if one finished the PU-LT task in 15 minutes, the Time Percent would be 75% (i.e., 15/20). At the same time, the quality of the participants' answers was quantified as the Task Score graded by two senior Ph.D. students in the field of human factors. Both raters had published at least two peer-reviewed journal articles in the field of human factors in transportation. Prior to the evaluation, the two raters agreed upon a scoring standard. Then, they evaluated the answers independently. To measure the consistency and inter-rater reliability of the scores, an intra-class correlation coefficient (ICC) analysis was conducted. The ratings from the two raters reached a high ICC of 0.94, with 95% confidence interval (95%CI) of [0.91, 0.97], $p < .0001$, indicating strong agreement and consistency between the two raters. In addition, we also recorded whether a participant adopted the responses provided by the LLM in a task. A response was marked as fully-adopted if the participant fully used the answers provided by the LLM; otherwise, a response was marked as un/partially-adopted. The adoption rates

(the number of fully-adopted cases over the number of un/partially-adopted cases) in each experimental condition are provided in Table 2.

3.4.2 Questionnaire-based variable

In addition to the task-performance-related factors (i.e., Time Percent and Task Score), variables were also extracted from pre-experiment, within-experiment, and post-experiment questionnaires. Table 3 summarizes all questionnaire-based variables, their distributions, the rationale for choosing these variables, and the calculation methods. Table 4 further illustrates the reliability and validity assessment of the metrics extracted from standard questionnaires, where all metrics reached satisfactory levels.

Table 2. Discreprtive Statistics of Adoption Rates of LLM Answers in Different Experimental Conditions.

With limitation-based training				Without limitation-based training			
PU		LR		PU		LR	
LT	ST	LT	ST	LT	ST	LT	ST
3/21	4/20	0/24	0/24	7/17	5/19	0/24	0/24

Table 3. Descriptive Statistics and Details of Questionnaire-Based Variables.

Source	**Variable (abbreviation)**	**Distribution**	**Description**	**Calculation processes**
Screening questionnaire	Number of Publications (N.Pub)	• 0 (n=10, 20.8%) • 1~3 (n=35, 72.9%) • 4~6 (n=3, 6.3%)	The number of publications in English (i.e., journals, conference proceedings or books).	Participant self-reported directly.
	Experience with LLM (ExLLM)	• Never used (n=6, 12.5%) • Rarely used (n=18, 37.5%) • Sometimes used (n=24, 50.0%)	Frequency of using LLMs (e.g., ChatGPT) for academic tasks.	Participant self-reported directly.
Pre-experiment questionnaire	Personality - Openness to Experiences (O)	Mean: 5.1 (SD: 0.9, min: 2.5, max: 6.5)	From creative and imaginative (high O) to practical and conventional (low O).	The scores from two questions for each personality trait were averaged (Gosling et al., 2003), ranging from 1 to 7 for each trait.
	Personality - Conscientiousness (C)	Mean: 4.7 (SD: 0.9, min: 3, max: 7)	From cautious and prudent (high C) to impulsive (low C).	
	Personality - Extroversion (E)	Mean: 4.1 (SD: 1.2, min: 2, max: 6)	From sociable and outgoing (high E) to reserved and quiet (low E).	
	Personality - Agreeableness (A)	Mean: 4.5 (SD: 0.8, min: 2, max: 6)	From cooperative and sympathetic (high A) to critical and tough (low A).	
	Personality - Emotional Stability (ES)	Mean: 4.4 (SD: 1.1, min: 2.5, max: 7)	From sensitive and easily upset (low ES) to calm and composed (high ES).	
	Trust Propensity (TP)	Mean: 16 (SD: 2.1, min: 12, max: 21)	From a natural inclination to trust others (high TP) to hesitation or reluctance to trust (low TP).	The sum of the scores of the five positive items with the the scores of the remaining negative questions being subtracted (Merritt et al., 2013), ranging from 0 to 24.
	Domain Knowledge (DK)	Mean: 61.7 (SD: 28.0, min: 20, max: 100)	The higher the DK, the better understanding of LLMs.	The sum of the score for each correct answer (20 each), ranging from 0 to 100.
Within-experiment questionnaire	Workload (WL)	Mean: 48.1 (SD: 11.9, min: 22.3, max: 92.9)	The higher the WL, the more workload an individual experiences during a task.	The weighted score of mental demand, physical demand, temporal demand, performance, effort, and frustration, ranging from 1 to 100 (Hart & Staveland, 1988).

	Trust in LLM (TR)	Mean: 4.1 (SD: 0.9, min: 1.6, max: 6.0)	Higher PT indicates higher trust towards LLM.	The average score of reliable, dependable, precise, trustable, and traceable, ranging from 1 to 6 (Franke et al., 2015).
Post-experiment questionnaire	System Usability (SU)	Mean: 76.6 (SD: 12.2, min: 56.3, max: 100.0)	The higher the SU, the more extent that the LLM can be used to accomplish tasks effectively, efficiently, and with satisfaction.	SL and SU are calculated from ten questions in SUS (Brooke, 2013), ranging from 0 to 100.
	System Learnability (SL)	Mean: 81.5 (SD: 17.5, min: 37.5, max: 100.0)	The higher the SL, the more easily that the LLM can be learned to operate or interact with.	
	Perceived Conversational Interaction (PI)	Mean: 24.4 (SD: 4.9, min: 10.0, max: 35.0)	The higher the PI, the higher overall evaluation of the LLM from conversational interaction perspective of view.	The sum of the scores of five questions (Walker et al., 1997), ranging from 5 to 35.

Table 4. Reliability and Validity Assessment over Factors from Standard Questionnaire.

	Cronbach α	KMO	*p* of Bartlett's Sphericity test
Openness to Experiences	0.778	0.743	<.0001
Conscientiousness	0.752	0.765	<.0001
Extroversion	0.781	0.798	<.0001
Agreeableness	0.737	0.782	<.0001
Emotional Stability	0.765	0.786	<.0001
Trust Propensity	0.765	0.730	<.0001
Workload	0.722	0.730	<.0001
Trust in LLM	0.861	0.843	<.0001
System Usability	0.722	0.763	<.0001
System Learnability	0.795	0.797	<.0001
Perceived Conversational Interaction	0.834	0.766	<.0001

3.5 Data analysis

To answer the three research questions, we explored: 1) how users' performance in academic tasks (i.e., Time Percent (TP), Task Score (TS), Workload (WL), and Trust in LLM (TR)) can be affected by experimental conditions (i.e., task difficulty, time pressure and level of training) and individual differences (i.e., Number of Publications (N.Pub), Experience with LLM (ExLLM), Trust Propensity (TP), Domain Knowledge (DK), and Personality). This part of the analysis will answer RQ1 and RQ2 ; 2) how users' attitudes towards the LLM can be moderated by the experimental conditions and individual differences, which will answer RQ3. A hierarchical structure inspired by a variant (Ghazizadeh et al., 2012) of the TAM (Davis, 1989) was adopted.

Traditionally, hierarchy relationships among variables can be identified by statistical tools (e.g., nested linear regression (Seber & Lee, 2003), and structural equation model (Ullman & Bentler, 2012)). However, conventional statistical methods require well-structured data formats and pre-assumptions regarding the relationships between latent variables, which make these approaches unsuitable when factors cannot be measured using Likert scale questions (e.g., performance, domain knowledge). Further, hierarchy relationships between the factors can hardly be explored efficiently in these approaches. Thus, in this study, a mixed approach combining Bayesian network (BN) and regression analyses (referred to as the BN-regression mixed approach (Wang, Tu, et al., 2023)) was adopted. The concept of BN was initially introduced by Judea Pearl (Friedman et al., 1997) and has been applied in various domains, including human-computer interaction and psychology (Sun et al., 2022; Wu et al., 2005).

In our study, a BN model was constructed to examine the structured dependency relationships among the influential factors impacting scholars' acceptance of LLMs and attitude towards LLMs. Then, to further validate if factors in BN with significant dependencies are significantly linearly correlated, and specify the statistical effects, regression analyses were conducted for each identified sub-structure in the BN. Through such a mixed approach, we were able to capture relationships ensuring linear correlations or causality.

3.5.1 Bayesian Network Construction

Bayesian Network (BN) is a graphical model that utilizes Bayes' theorem to showcase the conditional dependency relationships between variables effectively (Heckerman, 2008). BN is represented by a

directed acyclic graph (DAG), where each variable is depicted as a node, and the connections between nodes are represented as edges. These edges demonstrate conditional dependencies, which can either be determined through the data-based method or specified based on prior knowledge (Sun & Erath, 2015). The data-based approach is known to yield informative structures and achieve good prediction performance. However, limitations exist due to the quality and quantity of available data (Khakzad et al., 2011). On the other hand, the prior-knowledge-based approach may struggle to identify the dependency structure accurately. Therefore, our study adopted a hybrid approach, which combines both the data-based and prior-knowledge-based approaches to construct the BN structure.

In our study, we incorporated prior domain knowledge based on the TAM framework proposed by Davis (1989), along with relevant literature (Ghazizadeh et al., 2012; Longo, 2018; Merritt et al., 2013), to identify potential variables and structures. Since the construction of a BN relies on the estimation of conditional probabilities, we first discretized the continuous variables using quartiles equal-frequency discretization (Maslove et al., 2013). To ensure the balance between model fitting performance (ensuring sufficient data in each level of the variables) and avoid information loss (resulting from discretizing continuous variables), we discretized continuous variables into four categories based on their 25%, 50%, and 75% quantile values. In case of equal quartiles, neighboring categories were combined or aggregated.

After the discretization, variables extracted from the data that was collected at the same stage of the study (e.g., Workload and Task Score were both assessed during a task) were put into the same layer of the BN. Specifically, user-trait-related (Number of publications, Experience with LLM, Personality, Trust Propensity, and Domain Knowledge) and experimental factors (Task Type, Time Pressure, and Level of Training) were in the first layer, given that they cannot be influenced by other factors during the experiment and were assessed or decided before conducting the experiment; factors related with mental state (Workload and Trust in LLM) and task performance (Time Percent and Task Score) were in the second layer, as they are the states or outcome obtained during the experiment; and factors related with system evaluation (System Usability, System Learnability, and Perceived Conversational Interaction) were in the last layer, as they can be influenced by factors in all other layers and were collected in post-experiment stage.

Subsequently, we established a fully connected network by linking all factors in one layer to all other factors in other layers. The initial network was then pruned using an automated constraint conditional dependency search approach driven by the data (Schulte et al., 2009). Only edges exhibiting significant conditional dependencies in Chi-squared tests ($p < .05$) were retained in the Bayesian Network (BN). We utilized the "pgmpy" package (Ankan & Panda, 2015) in Python 3.8 for BN structure construction. It should be noted that we did not model the adoption rates of LLM answers into the BN network, given the highly unbalanced data (see Table 2).

3.5.2 Regression Analysis

In order to quantify the associations among influential variables, using "SAS OnDemand for Academics", regression analyses were performed for all hierarchical sub-structures within the BN. Mixed linear regression models (using Proc MIXED) were built for continuous dependent variables, and the generalized linear regression models (using Proc GENMOD) were built for discrete dependent variables. Repeated measures were accounted for through a generalized estimating equation, which can be used to model multiple responses from a single subject. In particular, for each sub-structure in the BN, regression models were developed with the node as the dependent variable, and its parental nodes, as well as their two-way interactions as independent variables. Backward stepwise selection procedures were employed based on model fitting criteria and the Variance Inflation Factor (VIF) was used to mitigate the issue of multicollinearity. To examine the significance of variables within each sub-structure, Tukey-Kramer post-hoc tests (Kramer, 1956) were conducted. Variables with a $p < .05$ were considered statistically significant in the analyses. Power analysis indicates that, the statistical models can reach at least a power of 0.871 when it comes to the most complex model with 9 predictors (as shown in Table 6), with the effect size of 0.1 and the significance level of 0.05, exceeding the general standard of 0.8 (Grosse et al., 2023).

4. Results

4.1 BN Results

Fig.2 visualizes the DAG of constructed BN. Referring to the experiment design (Fig.1), a three-layered structure was observed. Specifically, as the first layer, the blue box consists of user-trait-

related factors, which were assessed before conducting the experiment. In the second layer, the red boxes contain factors related to users' mental state and their task performance, and the orange boxes refer to three experimental conditions. The information collected in the post-experiment stage is in the green box. It should be noted that, we intentionally kept the edges from the experimental factors to task performance and mental states, given that their relationships are of interest in this study.

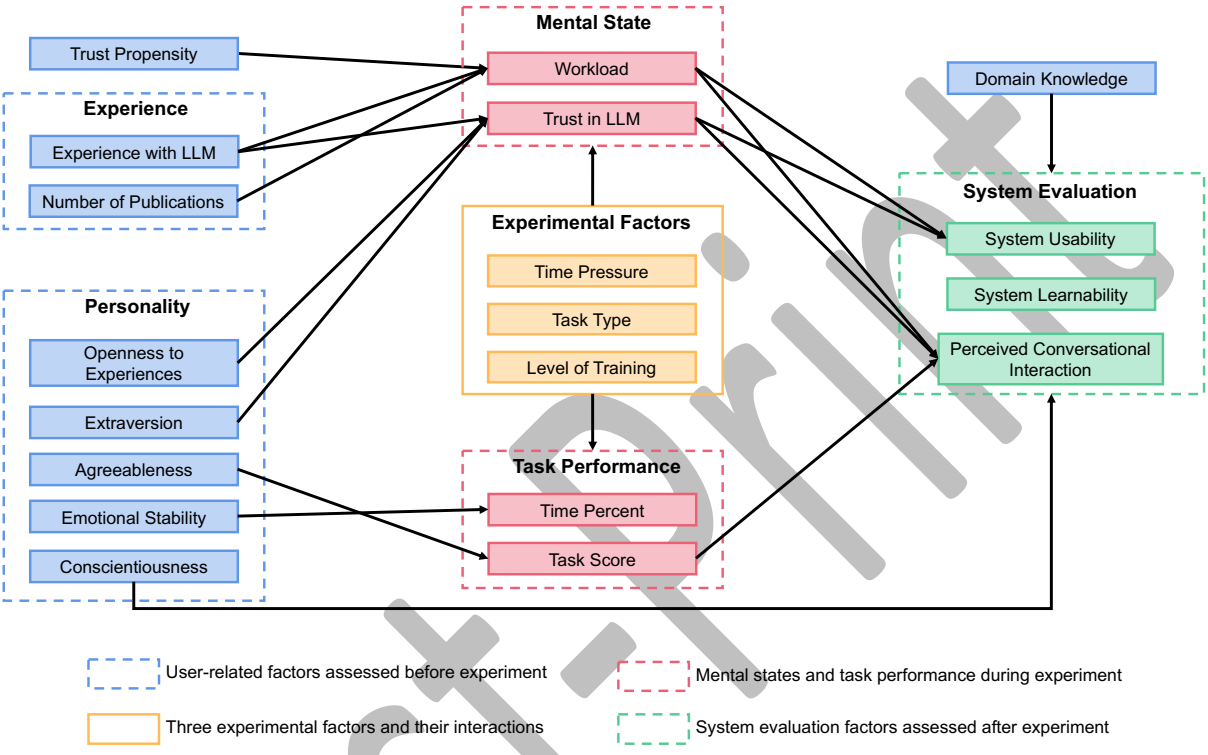

Fig.2. The DAG of the developed BN model. The arrow pointing to a box with dashed border indicates significant dependencies with all factors within the box.

4.2 Regression Analyses Results

We first evaluated the correlations among all factors (Table 5) to help avoid potential multicollinearity issues in the following linear regression modeling process. Based on the statistical analysis of each sub-structure in the BN, we kept statistically significant relationships in BN (as shown in Fig. 3). All statistical results and the corresponding significant ($p < .05$) post-hoc comparison results for significant variables are presented in Table 6 and Table 7. The significant post-hoc comparisons for categorical independent variables are further visualized in Fig. 4.

Table 5. Spearman correlations of all variables.

	EXLLM	N.PUB	O	C	E	A	ES	TP	DK	WL	TR	TS	TP	SU	SL	PI	LeT	TT	TiP
ExLLM	-		**	*	**	**	**	**	**	**	**				**	**			
N.Pub	0.11	-		**	*	*						**			*	*	**		
O	0.37	0.03	-	**		**	**	**						**	**	**			
C	0.14	-0.24	0.36	-	**		**	**	*										
E	0.32	-0.12	0.01	0.27	-	**	**		**	**	*			**			**		
A	0.28	0.13	0.24	0.10	0.43	-	**	**						**		**	**		
ES	0.18	-0.03	0.44	0.24	0.28	0.19	-	**	**		**			**	**	**	*		
TP	0.20	0.03	0.24	0.22	0.08	0.35	0.47	-	*	*			**	**		**	**		
DK	0.22	0.08	0.10	0.12	0.23	0.09	0.23	0.13	-		**	**			**		**		
WL	0.30	-0.01	-0.03	0.02	0.19	0.08	-0.10	-0.13	0.09	-	**					**			**
TR	0.24	0.03	0.03	0.04	0.14	0.05	-0.17	0.03	0.14	0.15	-	**				**		**	
TS	0.06	0.16	0.02	-0.05	-0.06	0.07	-0.09	0.10	0.17	0.07	0.25	-	**			**		**	*
TP	0.03	-0.04	-0.03	-0.05	-0.01	-0.08	-0.11	-0.17	-0.07	0.03	-0.04	-0.23	-						**
SU	0.12	0.00	0.21	0.07	0.16	0.21	0.19	0.34	0.10	0.08	0.02	0.03	-0.05	-	**	**	*		
SL	-0.17	-0.13	-0.20	0.02	0.07	-0.09	-0.21	-0.11	0.22	0.02	0.08	0.06	0.12	0.34	-	**	*		
PI	0.32	0.12	0.27	0.01	-0.02	0.17	-0.19	0.16	0.07	0.23	0.38	0.14	-0.01	0.22	0.24	-			
LeT	0.00	0.21	-0.02	-0.04	0.20	0.27	-0.14	-0.29	-0.15	0.02	0.05	0.06	0.08	-0.12	-0.13	-0.02	-		
TT	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.04	-0.16	-0.55	-0.10	0.00	0.00	0.00	0.00	-	
TiP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.18	-0.11	-0.14	0.27	0.00	0.00	0.00	0.00	0.00	-

*Notes: In this and following tables, * indicates the marginal significant results ($p < .1$), and ** marks significant results ($p < .05$). In the table,* ***ExLLM*** *is Experience with LLM;* ***N.Pub*** *is Number of Publications;* ***O*** *is Openness to Experiences;* ***C*** *is Conscientiousness;* ***E*** *is Extroversion;* ***A*** *is Agreeableness;* ***ES*** *is Emotional Stability;* ***TP*** *is Trust Propensity;* ***DK*** *is Domain Knowledge;* ***WL*** *is Workload;* ***TR*** *is Trust in LLM;* ***TS*** *is Task Score;* ***TP*** *is Time Percent;* ***SU*** *is System Usability;* ***SL*** *is System Learnability;* ***PI*** *is Perceived Conversational Interaction;* ***LeT*** *is Level of Training;* ***TT*** *is Task Type;* ***TiP*** *is Time Pressure.*

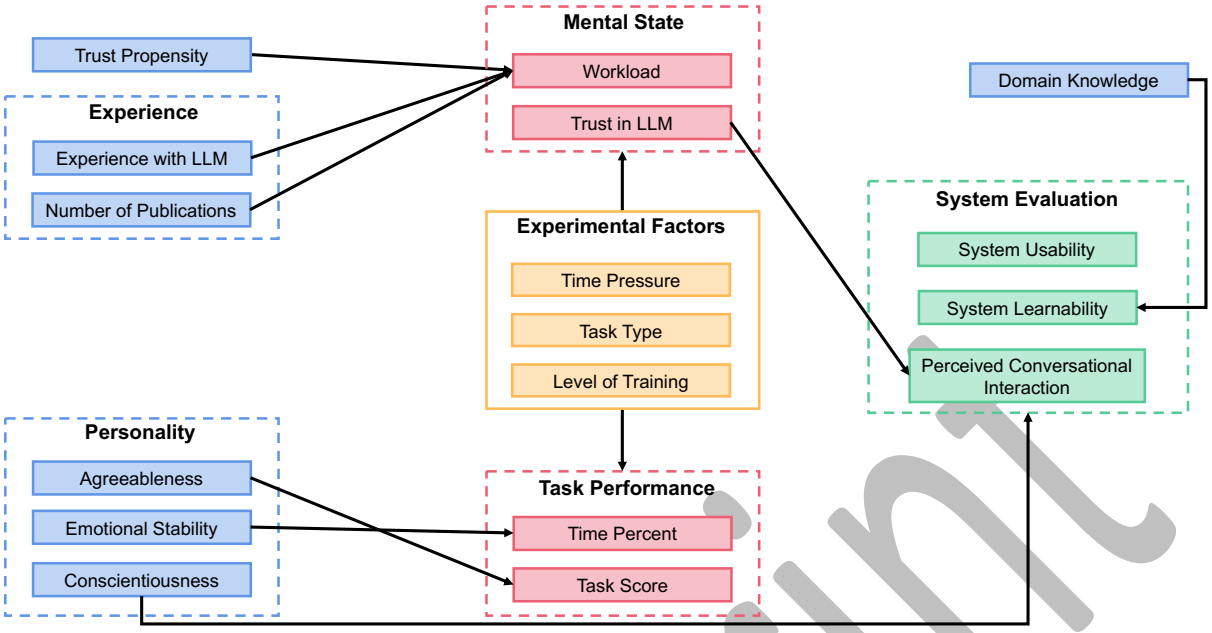

Fig.3. The final DAG after the regression analysis.

Table 6. Summary of Inferential Statistical Results.

Dependent Variable (DV)	**Independent Variable (IV)**	**F-value**	**Estimate (95% CI)**	***p*-value**
Workload	Experience with LLM	F(2, 41) = 3.87	-	.03 **
	Number of Publications	F(2, 41) = 3.82	-	.03 **
	Trust Propensity	F(1, 41) = 4.17	-1.09 [-2.18, -0.01]	.048 **
	Training	F(1, 41) = 0.52	n.s.	.5
	Task Type	F(1, 46) = 0.77	n.s.	.4
	Training * Task Type	F(1, 46) = 4.58	-	.04 **
	Time Pressure	F(1, 46) = 16.08	-	.0003 **
	Training * Time Pressure	F(1, 46) = 0.68	n.s.	.4
	Task Type * Time Pressure	F(1, 47) = 0.24	n.s.	.6
Trust in LLM	Experience with LLM	F(2, 42) = 2.46	-	.09 *
	Openness to Experiences	F(1, 42) = 0.23	n.s.	.6
	Extraversion	F(1, 42) = 0.26	n.s.	.6
	Training	F(1, 42) = 0.15	n.s.	.7
	Task Type	F(1, 46) = 10.13	-	.003 **
	Training * Task Type	F(1, 46) = 0.19	n.s.	.7
	Time Pressure	F(1, 46) = 5.00	-	.03 **
	Training * Time Pressure	F(1, 46) = 0.60	n.s.	.4
	Task Type * Time Pressure	F(1, 47) = 1.25	n.s.	.3
Time Percent	Emotional Stability	F(1, 45) = 5.94	0.51 [0.17, 0.85]	.02 **
	Training	F(1, 45) = 1.56	n.s.	.2
	Task Type	F(1, 46) = 4.56	-	.04 **
	Training * Task Type	F(1, 46) = 0.01	n.s.	.9
	Time Pressure	F(1, 46) = 5.42	-	.02 **
	Training * Time Pressure	F(1, 46) = 0.05	n.s.	.8
	Task Type * Time Pressure	F(1, 47) = 0.64	n.s.	.4
Task Score	Agreeableness	F(1, 45) = 6.26	5.67 [1.11, 10.23]	.02 **
	Training	F(1, 45) = 0.59	n.s.	.5

	Task Type	F(1, 46) = 124.38	-	<.0001 **
	Training * Task Type	F(1, 46) = 0.06	n.s.	.8
	Time Pressure	F(1, 46) = 8.44	-	.006 **
	Training * Time Pressure	F(1, 46) = 0.00	n.s.	.9
	Task Type * Time Pressure	F(1, 47) = 0.04	n.s.	.9
System Usability	Trust in LLM	F(1, 187) = 0.46	n.s.	.5
	Workload	F(1, 187) = 5.21	n.s.	.3
	Domain Knowledge	F(1, 187) = 0.00	n.s.	.9
	Conscientiousness	F(1, 187) = 35.58	4.56 [3.25, 6.47]	<.0001 **
System Learnability	Domain Knowledge	F(1, 189) = 5.45	0.09 [0.01, 0.17]	.02 **
	Conscientiousness	F(1, 189) = 44.79	8.28 [5.84, 10.72]	<.0001 **
Perceived Conversational Interaction	Trust in LLM	F(1, 186) = 46.40	2.35 [1.67, 3.03]	<.0001 **
	Workload	F(1, 186) = 3.39	-	.051 *
	Task Score	F(1, 186) = 0.00	n.s.	.9
	Domain Knowledge	F(1, 186) = 0.33	n.s.	.6
	Conscientiousness	F(1, 186) = 9.97	1.07 [0.40, 1.74]	.002 **

Note:'n.s.' means effect is not significant; '-' means the post-hoc comparisons are provided in Table 7. The Estimate (95% CI) indicates that with every one unit increase of the IV, the changes in the DV.

Table 7. Significant Post-hoc Comparisons for Discrete Independent Variables

DV	**IV**	**IV Level**	**IV Level being compared to**	**Δ (95% CI)**	**t value**	***p*-value**
Workload	Experience with LLM	Rarely use	Sometimes use	-6.15 [-11.80, -0.51]	t(41) = -2.65	.03 **
	Number of Publications	1-3	4-6	-11.15 [-21.80, -0.49]	t(41) = -2.54	.04 **
	Time Pressure	LT	ST	4.97 [2.47, 7.48]	t(41) = 4.01	.0003 **
Trust in LLM	Time Pressure	LT	ST	0.22 [0.02, 0.41]	t(46) = 2.24	.03 **
	Task Type	PU	LR	0.31[0.11, 0.50]	t(46) = 3.18	.003 **
Time Percent	Task Type	PU	LR	3.49 [0.20, 6.78]	t(46) = 2.14	.04 **
	Time Pressure	LT	ST	-3.80 [-7.09, -0.51]	t(46) = -2.33	.02 **
Task Score	Task Type	PU	LR	24.03 [19.69, 28.37]	t(46) = 11.15	<.0001 **
	Time Pressure	LT	ST	6.26 [1.92, 10.60]	t(46) = 2.91	.006 **

Note: Δ = IV Level - IV Level being compared to: when it is positive, it means IV Level > IV Level compared to.

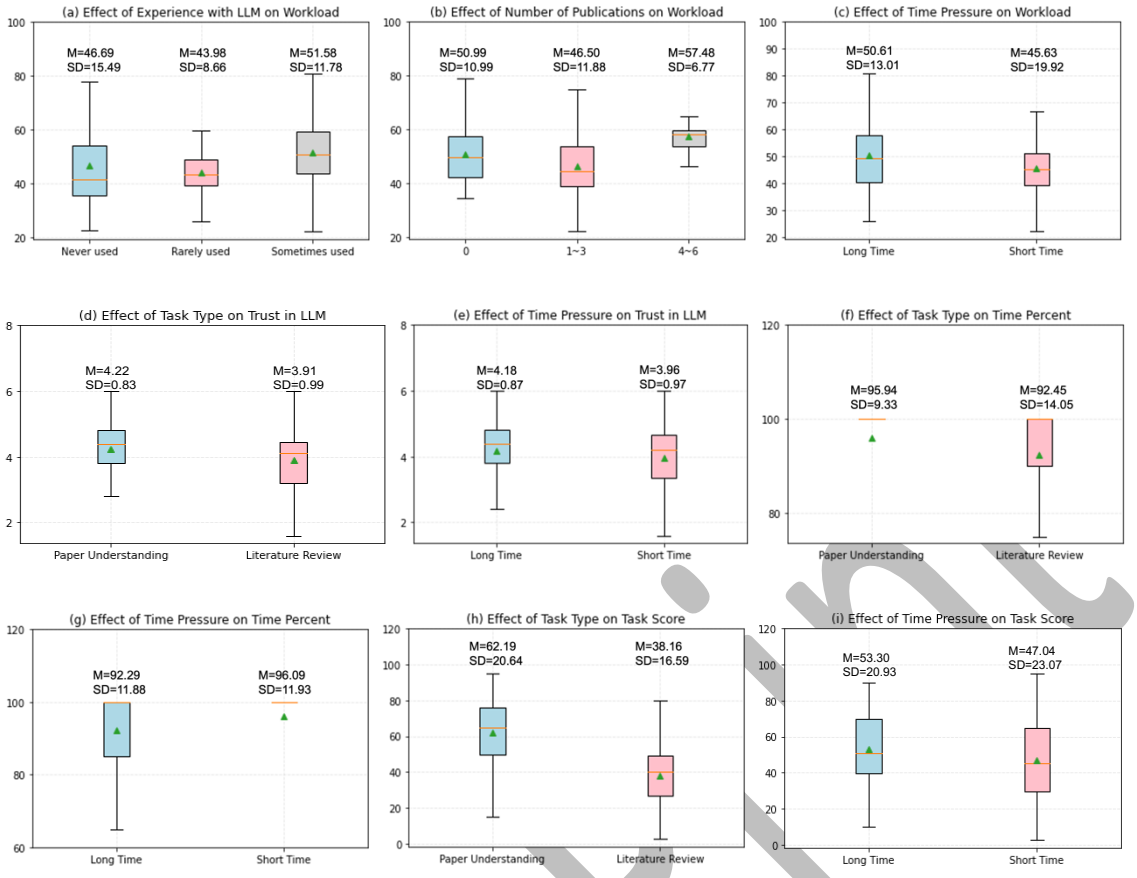

Fig.4. Boxplots for significant post-hoc comparisons. Boxplots show the five numbers summary as well as the averages indicated with green triangles. In the figure, M stands for mean and SD stands for standard deviation.

5. Discussion

In this section, following the hierarchy influence structure proposed in previous theoretical TAM-based models (Davis, 1989; Ghazizadeh et al., 2012), we first discussed factors influencing users' states and performance during the task. Then, we discussed how users' states and performance during the task, along with users' traits can influence their attitudes towards the LLM they used. All discussions are based on Table 6, Table 7, and Fig. 4 in the Results section.

To draw meaningful conclusions, we first examined the effectiveness of the experimental controls. First, the Task Type and Time Pressure have influenced the subjective and objective task complexities in an expected manner. Specifically, with higher Time Pressure, users obtained lower task scores, which indicates that controlling the allowed time for the task has successfully increased

the task complexity (ALQahtani et al., 2016; Maule & Edland, 2002). At the same time, the Task Type did not influence the perceived workload of the users, indicating that though discrepancies existed in the Task Scores of the two types of tasks due to different grading strategies, users did not perceive the two types of tasks as requiring different levels of effort. It should be noted that the influence of Time Pressure on the perceived workload is unexpected but reasonable. Specifically, compared to users with higher time pressure, users with lower time pressure perceived a higher workload. It is likely that academic tasks naturally require high cognitive resources (Omolayo & Omole, 2013), and extra time in the low time pressure condition has been devoted to revising the answers provided by the LLM. This further indicates that the users had upheld a high stand and responsibility when conducting the tasks in the experiment, and the conclusions drawn from this study should reflect early-stage scholars' states and performance when adopting LLMs for academic tasks to some level.

5.1 Influential factors of users' performance in academic tasks

Scholars performed differently when conducting different academic tasks with the help of LLMs. Specifically, we found that users gained higher scores but also spent more time in the paper understanding tasks compared to that when conducting literature review tasks. As mentioned, for human users, paper understanding involves extracting key information from a known source, where correct information can be found; whereas in the literature review task, the targeted information is vague but the source of the information is unknown. As for the LLM, the ChatGPT can summarize information from known sources with relatively high accuracy (Dis et al., 2023; Zhu et al., 2023), but may provide fake or inaccurate information in literature searching tasks when the information source is not provided (Dis et al., 2023). Thus, it is explainable that the users could perform better in the paper understanding task than in the literature review task with the help of LLM. Further, it is not surprising that users spent more time in paper understanding task compared to that in literature review task, given in the paper understanding task, they had to copy the content from a PDF to the ChatGPT before using prompts in order to get the answers; whereas in the literature review task, they only needed to input the prompts into the box. These findings can partially answer RQ1. However, given the current rapid development of LLM, where more advanced models are being introduced (e.g.,

GPT4[4], which was not used in this work as it was not publicly accessible when our experiment began), the capabilities of LLMs may change and future assessment of how users' behaviors change adaptively with the evolvement of LLMs are needed.

The performance in the academic task with LLM can also be moderated by users' traits in addition to the task complexity. With increased agreeableness (i.e., more cooperative and sympathetic), users gained higher task scores. This is intuitively opposite to some previous findings (Shaw & Choi, 2023; Witt et al., 2002). In our experiment, tasks can be regarded as being accomplished by a team, where the LLM played a powerful but noisy (unreliable answers were inevitable) assistant that cooperated with human users. According to Lim et al., (2023), people with high agreeableness is easier to find the solution when resolving noisy problems with teammates who have strong influence. Thus, our finding revealed that LLM was treated as more of a collaborator than a tool in creative tasks, and reveals that the responses generated by LLM might be helpful to certain groups of users, even they can be noisy. At the same time, users with higher emotional stability (i.e., more calm and composed) spent more time finishing the tasks. This is also easy to understand, those who have higher emotional stability may be more resistant to time pressure and still try to guarantee their answer qualities even if it takes more time to complete the task. This is similar to the case in driving scenarios, i.e., drivers with higher emotional stability usually drive slower (Scott-Parker, 2017). These findings can also partially answer RQ1.

As for RQ2, we found that the limitation-based training, surprisingly, did not affect users' performance in the selected academic tasks. However, when designing the academic tasks used in the experiment, in pilot tests, and in actual experiments, we noticed that LLM still generates inaccurate answers in all tasks. It is possible that users may have kept basic academic standards or responsibility in the experiment and thus nullified the effectiveness of the training, given that very few participants have directly adopted answers from LLMs (see Table 2), and neither time-pressure ($\chi^2(1) = 0.02$, p = .9) nor limitation-based training ($\chi^2(1) = 1.09$, p = .3) had effects on the adoption rates. The influence of users' perceived responsibility in the task has been observed in the driving automation

[4] https://openai.com/gpt-4

domain, in which responsibility-based training has been found to be more effective compared to limitation-based training (DeGuzman & Donmez, 2022). Based on the answer adoption rate, we also noticed that participants relied on LLM more in paper understanding tasks compared to literature review tasks. This finding, combined with the low answer adoption rate in general, indicates that scholars have relied on LLMs appropriately and adaptively. The over-reliance problem due to unawareness of limitations of LLMs may be neutralized by users' high responsibility in the academic task and may be less of a concern for professional users such as scholars .

5.2 Influential factors of workload when conducting academic tasks

The users' self-reported workload while using LLMs for academic tasks can provide insights on RQ1 from another perspective. We found that users' perceived workload in the task was not directly related to users' performance in the task and only significant post-hoc effects of the Time Pressure were observed for users' perceived workload. At the same time, users' Trust Propensity, Experience with LLM, and experience in academic tasks (i.e., Number of Publications) were also associated with users' perceived workload, but not performance in tasks. Given that these user traits did not affect users' performance in academic tasks with LLM, it seems that the variations in workload as a result of the heterogeneity in users' traits were not large enough to affect users' performance in the tasks with LLM. Specifically, with the increase of the propensity to trust in automation, users reported lower workload when conducting academic tasks with LLM. This finding is easy to understand and inline with previous research (Cai et al., 2022), that is, those who trusted more in LLMs might rely more on LLMs and devote less effort to performing academic tasks.

The positive relationships between workload and the experience with LLM, and between workload and experience in academic tasks (i.e., Number of Publications) are surprising, as previous studies illustrated that more familiarity with a domain could reduce attention resources (Sweller, 1994). In the academic tasks, however, users with more experience with LLM might be more familiar with the limitations of LLMs, and users with more academic experience may have higher standards of rigorousness and originality (Bommasani et al., 2021). Thus, they were more likely to pay extra attention to double-check answers provided by LLMs (Dis et al., 2023). This finding indicates that, automation such as LLM may not necessarily reduce the workload of users, if the automation is

imperfect and users are aware of the limitations of the system. It is worth noting, however, the current study was conducted with young early-stage scholars, senior researchers may adopt different strategies as they may hold different levels of academic integrity.

5.3 Influential factors of users' attitudes towards LLMs

As for RQ3, we identified that personal traits (i.e., Personality and Domain Knowledge) and mental states could affect users' evaluation of the LLMs. Specifically, we found that users with more domain knowledge of LLM perceived higher system learnability. It is possible that domain knowledge of LLM enabled them to effectively express their preferences (Jin et al., 2018) to maximize LLMs' generation capabilities and thus enhanced their understanding of the system usage, similar to what has been found in other human-automation interaction domain (Knijnenburg et al., 2011); in contrast, novice users had to rely on system-initiated suggestions and had more difficulty understanding how the LLM worked. At the same time, people with higher Conscientiousness perceived the LLM as more useful, easier to learn, and having better conversational interactions. Conscientious users were often characterized as cautious, responsible individuals (John et al., 1999), and they may actively explore and compare different options to find the optimal choice (Miceli et al., 2018). This trait could contribute to a greater appreciation of system suggestions that aided them in making well-informed decisions with confidence (Cai et al., 2022). However, as Tziner et al., (2002) stated, conscientious raters usually could not give ratings that strongly reflect their true attitudes toward systems. Future studies should further validate this finding.

Furthermore, a positive association was also identified between Trust in LLM and Perceived Conversational Interaction. Trust has been identified as an influential factor towards acceptance of the system (J. Lee & Moray, 1992; Pavlou, 2003), and as a dynamic process, the acceptance of the system and trust interact with each other through a feedback mechanism (Gao et al., 2006; Ghazizadeh et al., 2012). It should be noted that, in our study, the trust was further influenced by task complexity (as moderated by Time Pressure) and Task Type, with higher trust being reported when using LLM for paper reading tasks and when the time pressure was lower. The higher reported trust in paper reading task is as expected, given the worse performance of LLMs in literature review tasks (Dis et al., 2023). The relationship between the time pressure and the reported trust may be explained by the relatively

satisfying performance of the LLMs – as a result, long time of exposure can increase users' trust in a system if the system works in a satisfactory way tool (Jensen et al., 2013; Yuviler-Gavish & Gopher, 2011). This relationship between task complexity and evaluation of the system further reveals that the dynamic process of using the system can influence users' attitudes towards the system through trust in the system, indicating the feasibility of considering TAM- and AAM-related factors to explain the variations in users' trust in LLMs, similar to what has been found in previous research on conversational systems (Pu et al., 2011; Radziwill & Benton, 2017).

6. Limitations

To control the number of needed participants, we only focused on two common academic tasks within a single academic domain and only considered time pressure as a moderating factor of task complexity. Future research is needed to validate our findings in more diverse academic scenarios where LLM can be used (e.g., academic writing, data analysis, and experimental design). Finally, only ChatGPT was used in the study. More LLMs and LLM tools that can get access to the information online (e.g., retrieval-augmented LLMs with ChatGPT Plugins) are becoming available after the experiment was completed. Future research should validate our findings when different LLMs are used and compare the influence of LLM capabilities and users' perception of different LLMs on users' performance and behaviors when using LLM for academic tasks.

7. Conclusions

This study investigated the factors influencing young scholars' performance and mental states in academic tasks when LLM was provided. We further explored how TAM- and AAM-related (Davis, 1989; Ghazizadeh et al., 2012) factors can influence users' attitudes toward LLMs in academic tasks. Based on a BN and regression-based approach, we found that:

- When using LLMs to conduct academic tasks, young scholars in our experiment commonly have upheld relatively high academic integrity and were able to adjust their reliance on the LLMs adaptively. Thus, in addition to limitation-based training, future research can explore the role of

enhancing academic integrity on calibrating academic users' trust and reliance on LLMs in academic tasks.

- Individual heterogeneity has moderated the user-LLM performance in academic tasks. Specific user personality traits (i.e., Emotional Stability and Agreeableness) can affect the performance of users in collaborating with LLMs to accomplish academic tasks. Thus, in order to improve effectiveness of using LLM for academic tasks, future LLMs may consider providing adaptive interfaces (e.g., providing hints on prompts for novice users) with users' traits considered.
- Users' trust in the LLMs and the workload in cooperating LLMs varied with academic task type and time pressure, as the LLMs may bring different levels of benefits to users in different situations, given the limited capabilities of the LLMs at this stage. Thus, to help users make better use of LLMs, in addition to enhancing the capability of the LLMs, future LLMs may also consider increasing the system transparency (Manca et al., 2023; Siepmann & Chatti, 2023), for example, by providing confidence level of the answers, so that the users may make decisions easier (reduce workload) and better calibrate their trust in LLMs in different scenarios.
- The evaluation of the LLMs in academic tasks was a dynamic process that can be moderated by users' states (i.e., perceived trust and workload) when interacting with LLMs, which can further be influenced by task complexities and users' traits. This finding indicates that AAM and TAM-based models may explain users' perception of using LLMs for academic tasks. Future work can further extend the theoretical model to explain users' acceptance of LLMs based on our findings.

Acknowledgment

This work was supported by Guangzhou Municipal Science and Technology Project (No. 2023A03J0011) and Guangzhou Science and Technology Program City-University Joint Funding Project (No. 2023A03J0001).

Reference

Alberts, I. L., Mercolli, L., Pyka, T., Prenosil, G., Shi, K., Rominger, A., & Afshar-Oromieh, A. (2023). Large language models (LLM) and ChatGPT: what will the impact on nuclear

medicine be? *European Journal of Nuclear Medicine and Molecular Imaging*, *50*(6), 1549–1552.

Ali Amer Jid Almahri, F., Bell, D., & Arzoky, M. (2019). Personas design for conversational systems in education. *Informatics*, *6*(4), 46.

ALQahtani, D. A., Rotgans, J. I., Mamede, S., ALAlwan, I., Magzoub, M. E. M., Altayeb, F. M., Mohamedani, M. A., & Schmidt, H. G. (2016). Does time pressure have a negative effect on diagnostic accuracy? *Academic Medicine*, *91*(5), 710–716.

Ankan, A., & Panda, A. (2015). pgmpy: Probabilistic graphical models using python. *Proceedings of the 14th Python in Science Conference (Scipy 2015), 10*.

Aydın, Ö., & Karaarslan, E. (2022). OpenAI ChatGPT generated literature review: Digital twin in healthcare. *Available at SSRN 4308687*.

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, *8*(4), 321–348.

Belda-Medina, J., & Calvo-Ferrer, J. R. (2022). Using chatbots as AI conversational partners in language learning. *Applied Sciences*, *12*(17), 8427.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., & others. (2021). On the opportunities and risks of foundation models. *arXiv Preprint arXiv:2108.07258*.

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: Changing user needs and motivations. *Interactions*, *25*(5), 38–43.

Broady, T., Chan, A., & Caputi, P. (2010). Comparison of older and younger adults' attitudes towards and abilities with computers: Implications for training and learning. *British Journal of Educational Technology*, *41*(3), 473–485.

Brooke, J. (2013). SUS: A Retrospective. *Journal of Usability Studies*, *8*, 29–40.

Cai, W., Jin, Y., & Chen, L. (2022). Impacts of personal characteristics on user trust in conversational recommender systems. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–14.

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 841–845.

Cho, J.-H., Cam, H., & Oltramari, A. (2016). Effect of personality traits on trust and risk to phishing vulnerability: Modeling and analysis. *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 7–13.

Choy, E. C., Patel, S. J., & Chaparro, A. (2022). Safety first: User needs analysis of advanced driver assistance systems (ADAS) to determine learning preferences. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1310–1314.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340.

DeGuzman, C. A., & Donmez, B. (2022). Drivers don't need to learn all ADAS limitations: A comparison of limitation-focused and responsibility-focused training approaches. *Accident Analysis & Prevention*, *178*, 106871.

Denecke, K., & May, R. (2022). Usability assessment of conversational agents in healthcare: A literature review. *Challenges of Trustable AI and Added-Value on Health*, 169–173.

Dergaa, I., Chamari, K., Zmijewski, P., & Saad, H. B. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biology of Sport*, *40*(2), 615–622.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*.

Dis, E., Bollen, J., Zuidema, W., Rooij, R., & Bockting, C. (2023). ChatGPT: five priorities for research. *Nature*, *614*, 224–226.

Dishaw, M., & Strong, D. (1998). Experience as a moderating variable in a task-technology fit model. *AMCIS 1998 Proceedings*, 242.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., & Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. *Advances in Neural Information Processing Systems*, *32*.

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters*, *53*, 103662.

Franke, T., Trantow, M., Günther, M., Krems, J. F., Zott, V., & Keinath, A. (2015). Advancing electric vehicle range displays for enhanced user experience: The relevance of trust and adaptability. *Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 249–256.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*(2), 131–163.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv*, 2022–12.

Gao, J., Lee, J. D., & Zhang, Y. (2006). A dynamic model of interaction between reliance on automation and cooperation in multi-operator multi-automation situations. *International Journal of Industrial Ergonomics*, *36*(5), 511–526.

Gero, K. I., Long, T., & Chilton, L. B. (2023). Social dynamics of AI support in creative writing. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–15.

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. *Cognition, Technology & Work*, *14*, 39–49.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660.

Gordijn, B., & Have, H. ten. (2023). ChatGPT: evolution or revolution? *Medicine, Health Care and Philosophy*, *26*(1), 1–2.

Gosling, S. D., Rentfrow, P. J., & Swann Jr, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528.

Grosse, K., Bieringer, L., Besold, T. R., Biggio, B., & Krombholz, K. (2023). Machine learning security in industry: A quantitative survey. *IEEE Transactions on Information Forensics and Security*, *18*, 1749–1762.

Guerino, G. C., Silva, W. A. F., Coleti, T. A., & Valentim, N. M. C. (2021). Assessing a Technology for Usability and User Experience Evaluation of Conversational Systems: An Exploratory Study. *ICEIS (2)*, 463–473.

Guerino, G. C., & Valentim, N. M. C. (2020). Usability and user experience evaluation of conversational systems: A systematic mapping study. *Proceedings of the XXXIV Brazilian Symposium on Software Engineering*, 427–436.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology* (Vol. 52, pp. 139–183). Elsevier.

He, D., & Donmez, B. (2019). Influence of driving experience on distraction engagement in automated vehicles. *Transportation Research Record*, *2673*(9), 142–151.

He, D., Kanaan, D., & Donmez, B. (2022). Distracted when using driving automation: A quantile regression analysis of driver glances considering the effects of road alignment and driving experience. *Frontiers in Future Transportation*, *3*, 772910.

Heckerman, D. (2008). A tutorial on learning with Bayesian networks. *Innovations in Bayesian Networks: Theory and Applications*, 33–82.

Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Human Factors*, *59*(3), 457–470.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5–53.

Hickerson, K., & Lee, Y.-C. (2022). A psychometric evaluation of a technology acceptance model for autonomous vehicle. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1289–1293.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434.

Holden, H., & Rada, R. (2011). Understanding the influence of perceived usability and technology self-efficacy on teachers' technology acceptance. *Journal of Research on Technology in Education*, *43*(4), 343–367.

Huang, C., He, D., Wen, X., & Yan, S. (2023). Beyond Adaptive Cruise Control and Lane Centering Control: Drivers' Mental Model of and Trust in emerging ADAS technologies. *Frontiers in Psychology*, *14*, 1236062.

Jannach, D., & Bauer, C. (2020). Escaping the mcnamara fallacy: Towards more impactful recommender systems research. *Ai Magazine*, *41*(4), 79–95.

Jensen, A. F., Cherchi, E., & Mabit, S. L. (2013). On the stability of preferences and attitudes before and after experiencing an electric vehicle. *Transportation Research Part D: Transport and Environment*, *25*, 24–32.

Jin, Y., Tintarev, N., & Verbert, K. (2018). Effects of individual traits on diversity-aware music recommender user interfaces. *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 291–299.

John, O. P., Srivastava, S., & others. (1999). *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*.

Jungherr, A. (2023). *Using ChatGPT and Other Large Language Model (LLM) Applications for Academic Paper Assignments*.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., & others. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Khakzad, N., Khan, F., & Amyotte, P. (2011). Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches. *Reliability Engineering & System Safety*, *96*(8), 925–932.

Kiraskowski, J., & Corbett, M. (1993). SUMI: the software usability measurement inventory, *British Journal of Educational Technology*, *24*(3), 210–212.

Knijnenburg, B. P., Reijmer, N. J. M., & Willemsen, M. C. (2011). Each to his own: How different users call for different interaction methods in recommender systems. *ACM Conference on Recommender Systems*. https://api.semanticscholar.org/CorpusID:7819904

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics*, *12*(3), 307–310.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & others. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, *2*(2), e0000198.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Liebrenz, M., Schleifer, R., Buadze, A., Bhugra, D., & Smith, A. (2023). Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *The Lancet Digital Health*, *5*(3), e105–e106.

Lim, S. L., Bentley, P. J., Peterson, R. S., Hu, X., & Prouty McLaren, J. (2023). Kill chaos with kindness: Agreeableness improves team performance under uncertainty. *Collective Intelligence*, *2*(1), 26339137231158584.

Longo, L. (2018). Experienced mental workload, perception of usability, their interaction and impact on task performance. *PloS One*, *13*(8), e0199661.

Luccioni, A., & Viviano, J. (2021). What's in the box? An analysis of undesirable content in the Common Crawl corpus. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 182–189.

Lyell, D., Magrabi, F., & Coiera, E. (2018). The effect of cognitive load and task complexity on automation bias in electronic prescribing. *Human Factors*, *60*(7), 1008–1021.

Manca, M., Palumbo, V., Paternò, F., & Santoro, C. (2023). The transparency of automatic web accessibility evaluation tools: Design criteria, state of the art, and user perception. *ACM Transactions on Accessible Computing*, *16*(1), 1–36.

Maslove, D. M., Podchiyska, T., & Lowe, H. J. (2013). Discretization of continuous features in clinical datasets. *Journal of the American Medical Informatics Association*, *20*(3), 544–553.

Matthew, H. (2022). Could AI help you to write your next paper? *Nature*, *611*.

Maule, A. J., & Edland, A. C. (2002). The effects of time pressure on human judgement and decision making. In *Decision making* (pp. 203–218). Routledge.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, *2*(2), 1–25.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520–534.

Miceli, S., de Palo, V., Monacis, L., Di Nuovo, S., & Sinatra, M. (2018). Do personality traits and self-regulatory processes affect decision-making tendencies? *Australian Journal of Psychology*, *70*(3), 284–293.

Mogavi, R. H., Deng, C., Kim, J. J., Zhou, P., Kwon, Y. D., Metwally, A. H. S., Tlili, A., Bassanelli, S., Bucchiarone, A., Gujar, S., & others. (2023). Exploring user perspectives on chatgpt: Applications, perceptions, and implications for ai-integrated education. *arXiv Preprint arXiv:2305.13114*.

Morris, M. R. (2023). Scientists' Perspectives on the Potential for Generative AI in their Fields. *arXiv Preprint arXiv:2304.01420*.

Omolayo, B. O., & Omole, O. C. (2013). Influence of mental workload on job performance. *International Journal of Humanities and Social Science*, *3*(15), 238–246.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, *7*(3), 101–134.

Pfeuffer, N., Benlian, A., Gimpel, H., & Hinz, O. (2019). Anthropomorphic information systems. *Business & Information Systems Engineering*, *61*, 523–533.

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. *Proceedings of the Fifth ACM Conference on Recommender Systems*, 157–164.

Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv Preprint arXiv:1704.04579*.

Rahman, M. M., Terano, H. J., Rahman, M. N., Salamzadeh, A., & Rahaman, M. S. (2023). ChatGPT and academic research: A review and recommendations based on practical examples. *Rahman, M., Terano, HJR, Rahman, N., Salamzadeh, A., Rahaman, S.(2023). ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples. Journal of Education, Management and Development Studies*, *3*(1), 1–12.

Rubin, V. L., Chen, Y., & Thorimbert, L. M. (2010). Artificially intelligent conversational agents in libraries. *Library Hi Tech*, *28*(4), 496–522.

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: Effects on trust, automation bias, complacency and performance. *Ergonomics*, *59*(6), 767–780.

Schmutz, P., Heinz, S., Métrailler, Y., Opwis, K., & others. (2009). Cognitive load in eCommerce applications—Measurement and effects on user satisfaction. *Advances in Human-Computer Interaction*, *2009*.

Scholtz, B. M., Mahmud, I., & Ramayah, T. (2016). Does usability matter? An analysis of the impact of usability on technology acceptance in ERP settings. *Interdisciplinary Journal of Information, Knowledge, and Management*, *11*, 309.

Schulte, O., Frigo, G., Greiner, R., Luo, W., & Khosravi, H. (2009). A new hybrid method for Bayesian network learning with dependency constraints. *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 53–60.

Scott-Parker, B. (2017). Emotions, behaviour, and the adolescent driver: A literature review. *Transportation Research Part F: Traffic Psychology and Behaviour*, *50*, 1–37.

Seber, G. A., & Lee, A. J. (2003). *Linear regression analysis* (Vol. 330). John Wiley & Sons.

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. *Recommender Systems Handbook*, 257–297.

Shaw, A., & Choi, J. (2023). Get creative to get ahead? How personality contributes to creative performance and perceptions by supervisors at work. *Acta Psychologica*, *233*, 103835.

Shawar, B. A., & Atwell, E. S. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, *10*(4), 489–516.

Siepmann, C., & Chatti, M. A. (2023). Trust and Transparency in Recommender Systems. *arXiv Preprint arXiv:2304.08094*.

Stevens, Horrock. (2019, October 7). *What Human Factors Isn't: 1. Common Sense*. https://humanisticsystems.com/2019/07/10/what-human-factors-isnt-1-common-sense/

Stokel-Walker, C. (2023). ChatGPT listed as author on research papers: Many scientists disapprove. *Nature*, *613*(7945), 620–621.

Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, *61*, 49–62.

Sun, L., Wang, L., Su, C., Cheng, F., Wang, X., Jia, Y., & Zhang, Z. (2022). Human reliability assessment of intelligent coal mine hoist system based on Bayesian network. *Scientific Reports*, *12*(1), 21880.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Turel, O., & Gefen, D. (2013). The dual role of trust in system use. *Journal of Computer Information Systems*, *54*(1), 2–10.

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2002). Does conscientiousness moderate the relationship between attitudes and beliefs regarding performance appraisal and rating behavior? *International Journal of Selection and Assessment*, *10*(3), 218–224.

Ullman, J. B., & Bentler, P. M. (2012). Structural equation modeling. *Handbook of Psychology, Second Edition*, *2*.

Van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: five priorities for research. *Nature*, *614*(7947), 224–226.

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. *arXiv Preprint Cmp-Lg/9704004*.

Wang, J., Huang, C., Tu, R., & He, D. (2023, January 21). *Influential Factors of Users' Trust in the Range Estimation Systems of Battery Electric Vehicles – A Survey Study in China*. 2022 Transportation Research Board. https://personal.hkust-gz.edu.cn/hedengbo/assets/publicationPDFs/Wang_TRB_2023a.pdf

Wang, J., Tu, R., Wang, A., & He, D. (2023). Trust in Range Estimation System in Battery Electric Vehicles – A Mixed Approach. *In Processing*.

Witt, L., Burke, L. A., Barrick, M. R., & Mount, M. K. (2002). The interactive effects of conscientiousness and agreeableness on job performance. *Journal of Applied Psychology*, *87*(1), 164.

Wu, C., Lin, Y., & Zhang, W.-J. (2005). Human attention modeling in a human-machine interface based on the incorporation of contextual features in a bayesian network. *2005 IEEE International Conference on Systems, Man and Cybernetics*, *1*, 760–766.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*.

Yuviler-Gavish, N., & Gopher, D. (2011). Effect of descriptive information and experience on automation reliance. *Human Factors*, *53*(3), 230–244.

Zhang, G., Chong, L., Kotovsky, K., & Cagan, J. (2023). Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Computers in Human Behavior*, *139*, 107536.

Zhu, Y., Han, D., Chen, S., Zeng, F., & Wang, C. (2023). *How can chatgpt benefit pharmacy: A case report on review writing*.

Jiyao Wang received B.Eng. degree from the Sichuan University in 2021, and M.Sc. degree from the Hong Kong University of Science and Technology in 2022. He is now a Ph.D. student in the Thrust of Robotics and Autonomous Systems at the Hong Kong University of Science and Technology (Guangzhou).

Chunxi Huang received his bachelor's degree in industrial engineering from Zhejiang University, Hangzhou, China in 2018 and his master's degree in industrial and systems engineering from Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2020. He is currently a Ph.D. candidate at the Hong Kong University of Science and Technology. His research interests include human factors, driver behavior, and traffic safety.

Song Yan received his bachelor's degree in automotive engineering from Zhejiang University, China, and master's degree in mechanical engineering from the University of Tokyo, Japan. He is currently a PhD student in Intelligent Transportation Thrust at the Hong Kong University of Science and Technology (Guangzhou). His research interests include automated driving systems, human factors, and driver behavior.

Weiyin Xie received B.Sc. degree and M.Sc. degree from the University of Duisburg-Essen. He is now a Ph.D. student in the Thrust of Robotics and Autonomous Systems at the Hong Kong University of Science and Technology (Guangzhou).

Dengbo He received his bachelor's degree from Hunan University in 2012, M.S. degree from the Shanghai Jiao Tong University in 2016 and Ph.D. degree from the University of Toronto in 2020. He is currently an assistant professor from the Thrust of Intelligent Transpiration and Thrust of Robotics and Autonomous Systems, the Hong Kong University of Science and Technology (Guangzhou). He is also affiliate with the Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology.