Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars

Jiyao Wang*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
jwanggo@connect.ust.hk

Haolong Hu*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
hhu574@connect.hkust-gz.edu.cn

Zuyuan Wang
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
zwang534@connect.hkust-gz.edu.cn

Yan Song
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
zwang534@connect.hkust-gz.edu.cn

Youyu Sheng
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
ysheng330@connect.hkust-gz.edu.cn

Dengbo He†
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
dengbohe@hkust-gz.edu.cn

ABSTRACT

The rapid advancement of large language models (LLMs) such as ChatGPT makes LLM-based academic tools possible. However, little research has empirically evaluated how scholars perform different types of academic tasks with LLMs. Through an empirical study followed by a semi-structured interview, we assessed 48 early-stage scholars' performance in conducting core academic activities (i.e., paper reading and literature reviews) under different levels of time pressure. Before conducting the tasks, participants received different training programs regarding the limitations and capabilities of the LLMs. After completing the tasks, participants completed an interview. Quantitative data regarding the influence of time pressure, task type, and training program on participants' performance in academic tasks was analyzed. Semi-structured interviews provided additional information on the influential factors of task performance, participants' perceptions of LLMs, and concerns about integrating LLMs into academic workflows. The findings can guide more appropriate usage and design of LLM-based tools in assisting academic work.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

large language model, academic tasks, user perception, human-AI collaboration

ACM Reference Format:
Jiyao Wang, Haolong Hu, Zuyuan Wang, Yan Song, Youyu Sheng, and Dengbo He. 2024. Evaluating Large Language Models on Academic Literature Understanding and Review: An Empirical Study among Early-stage Scholars. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (CHI '24)*. ACM, New York, NY, USA, 18 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The rapid advancement of artificial intelligence (AI) and natural language processing (NLP) has led to the development of sophisticated large language models (LLMs), such as ChatGPT [1], GPT4 [2], and Claude [3]. These models have demonstrated impressive capabilities in generating human-like text, understanding context, and solving complex language tasks. The application scope of LLM technology is extensive, and relevant scholars have been actively analyzing the impact of such technology on fields such as healthcare [1, 39], education [35, 36], and creative writing [26] (e.g., helping journalists extract ideas from document [56], enabling scholars to communicate findings mutually [25], and assisting writers in exploring more ways of writing story [12]).

In recent years, LLMs have been employed for various academic tasks [33], such as literature reviews, paper reading, writing polishing, etc. Being different from other areas where LLMs are applicable, the academic work requires extensive training in acquiring, judging, and synthesizing relevant information [46]. Moreover, the academic community also demands high standards for logical coherence, information accuracy, and idea novelty [6] and thus requires more responsible AI tools compared to other domains. However, the application of LLMs in academic contexts was under-investigated.

Further, as emerging conversational systems, the promotion of LLM results in more diverse user behaviors as well as new social

*Both authors contributed equally to this research.
†Corresponding author.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

[1]https://openai.com/chatgpt
[2]https://openai.com/gpt-4
[3]https://www.claude.co.id/

norms and user expectations [7]. Hence, it is imperative to evaluate the capacity boundary of LLMs in academic settings through user studies, so that the LLMs can be better designed to be integrated into academic workflows, and subsequently, contribute to academic research. A few studies have evaluated the effectiveness of LLMs in assisting selected academic tasks. For example, Gordijn and Have [27] argued that the capacity of ChatGPT to develop a whole scientific paper is restricted. LLMs have also been found to be able to alleviate some of the time pressures by automating certain processes during their academic tasks [14]. However, academic tasks are diverse, and different tasks may require completely different cognitive resources. For instance, compared to extracting key information from a paper (in which the source of the information is known, but the information is unknown), literature reviews require locating and summarizing information from a wider range of studies (in which the targeted information is partially known, but the source of the information is unknown). Actually, some research [2, 16, 52] has also pointed out that LLMs may introduce inaccuracies and biases in academic tasks, especially in understanding and summarizing the content of the literature as a priority concern for humanity [16]. Furthermore, task complexity can also be moderated by time pressure [44], which is prevalent in academia [11]. Thus, a more comprehensive investigation is needed to better understand the role of LLMs in different academic tasks with different task complexity, which can be moderated by task type and time pressure.

On the other hand, given that LLMs can be regarded as a special type of automation that can help gather, analyze, and summarize information, users' perceptions of the LLM may also influence task performance. Although a few empirical studies discussed the implication and limitations of LLMs when they are used for specific academic tasks (e.g., literature review [3], idea generation [22, 48]), no research has discussed different strategies young scholars may take when using the LLMs for different tasks, nor compared how task difficulties (e.g., as moderated by time pressure) and training may influence users' performance, although these factors have been widely acknowledged as influential factors of users' reliance on automation [32].

Thus, using a mixed-methods approach combining an experimental study with semi-structured interviews, this study aims to investigate:

- When using LLMs, whether and why there are discrepancies in young academic users' performance in conducting different academic tasks, as defined by time pressure and required cognitive resources.
- How can young academic users' perceptions of LLMs limitations affect their performance or strategies when using LLMs for academic tasks?
- What young academic users' expectations of the LLMs and LLM training are when the LLMs are being used for different academic tasks?

Given that the younger generation has a higher acceptance of emerging technologies [9], and may lack experience in conducting academic tasks, this study was planned to target young scholars, specifically, graduate students who have just started their academic careers as researchers. This decision was based on the fact that LLMs are new to most scholars, and based on the research in other domains, new users may have highly uncertain and potentially inappropriate strategies when they first start to use the LLMs [30]. Thus, understanding the strategies new users adopt can help support young scholars to better use the LLM tools or at least shorten their familiarization process, by providing new users with appropriate training materials. In the study, an onsite experiment was conducted, followed by a semi-structured interview regarding the usage of LLM in the experiment or their daily life. Together, the empirical and interview data offered a nuanced perspective on opportunities and challenges of using LLMs for academic tasks.

2 RELATED WORKS

2.1 Natural Language Technologies

In recent years, a remarkable evolution has been happening in the field of Computational Linguistics, also known as Natural Language Processing (NLP), primarily driven by the development of neuron-based network models trained on vast datasets [45, 68]. Compared to traditional rule-based systems, recent data-driven models have shown remarkable results across various NLP tasks [19, 53]. Deep learning techniques have become mainstream in developing these NLP models [40]. Current popular architectures include Long Short-Term Memory (LSTM) [67] and transformer models [63].

A significant paradigm shift in natural language technologies occurred over the past half-decade, primarily attributed to the advent of large language models (LLMs) [17]. These techniques involve an initial training phase on a comprehensive dataset, followed by fine-tuning for specific tasks. Pre-trained models like BERT [15], BART [42], XLNet [66], and LLaMa [62] have demonstrated substantial performance improvements across a variety of NLP tasks.

However, the challenges of smaller models persist in LLMs. For instance, the LLMs still lack an explicit factual model, which makes them prone to producing inaccurate information [34]. Even innocuous prompts can lead to the generation of toxic content from these models [24, 55]. Their performance varies, excelling in some areas while faltering in others [25]. Guiding these models to deliver specific outputs remains a challenge, leading to the emergence of prompt engineering as a sub-field [4, 45]. Ethical concerns surrounding these models are wide-ranging, from environmental to socio-political considerations [5].

Our research acknowledges the limitations of current LLMs and assumes that they cannot fully replace human in creative writing tasks. However, they can significantly aid academic writing across various contexts to a certain extent. This perspective motivates our exploration of users' concerns when using LLMs as a peer-level writing tool.

2.2 Large Language Model in Academic Tasks

Large Language Models (LLMs), exemplified by ChatGPT, harness broad internet-based datasets to mimic human language patterns and create realistic text [57]. This capability has attracted interest across academia. For instance, the broader implications of AI in academic research have been scrutinized by Grimaldi and Ehrler [28] and Hutson [33]. These tasks include the compilation of essential components of the manuscript such as the abstract, introduction,

literature review, methodology, results, discussions, and conclusions.

Scholars, researchers, and students in the academic community have utilized LLMs like ChatGPT for a variety of academic and non-academic tasks. Dowling and Lucey [18] explored the application of ChatGPT and found it to be particularly effective for initial idea generation, literature synthesis, and creating testing frameworks. Yet, according to Gordijn et al., [27], ChatGPT still fails to produce a complete scientific article on par with a skilled researcher. However, it is expected that the capabilities and uses of these tools will continue to grow, and will be capable of conducting more academic tasks, including experiment design, manuscript writing, peer reviews, and editorial decision support [16]. Additionally, the ability of ChatGPT to generate and understand texts in multiple languages is believed to improve the efficiency of publishing and accessing literature for non-native English speakers [43]. In general, scientists in many fields are positive about the potential of using LLM in academic tasks [52].

However, the performance of the LLM in academic tasks is still less than ideal. For example, Aydın and Karaarslan [3] pointed out that when using ChatGPT for a literature review in healthcare, the content generated by LLM still lacks synthesis, and may suffer the problem of plagiarism. In another study, Gao et al. [23] reported that the abstract generated by the LLMs can still be identified as AI-generated using an AI output detector. Particularly, through multiple experimental trials, Dis et al. [16] reminded researchers to pay extra attention and remain vigilant when applying LLM to literature comprehension and summarization tasks. However, to the best of our knowledge, no empirical research has been conducted to understand how scholars use the LLMs and how the LLMs can influence scholars' performance in academic tasks. Given that the LLM can still only work as a collaborator, it is necessary to consider the characteristics of the user-LLM combined system instead of the LLM alone. Furthermore, most of the existing research focused on the attitudes and opinions of senior researchers on the use of LLM [2, 52]. Few research focused on younger scholars, who may have a higher propensity to accept new technologies and lack the necessary expert knowledge to supervise the application of LLM in academic tasks [16].

3 METHODOLOGY

We adopted a mixed approach consisting of an empirical experiment and a semi-structured interview. Quantitative performance data was gathered to evaluate the performance and the strategies participants took for different academic tasks. Post-experiment interviews focused on researchers' evaluations of current LLM limitations in academia, their subjective understanding of the factors influencing their performance across tasks, and their concerns about integrating LLM into their workflows.

3.1 Participants

In the scope of this study, a diverse (in terms of academic background) pool of 48 young participants (age <= 30, 30 males and 18 females) was selected. All from research institutions or universities in China where English is the principal language of instruction. Recruited through online posters in the social network and on-campus posters, all participants were native Mandarin speakers. Each participant was given a unique experiment ID number from P1 to P48. Table 1 provides a comprehensive overview of the academic profiles of the participants. All participants were actively involved in academia, including 22 Ph.D. students, 17 MPhil students, and 9 research assistants (with a minimum of a bachelor's degree). An examination of their academic publication history reveals that 35 participants had 1 to 3 publications (including journal articles, conference proceedings, and edited books); 3 participants had 4-6 publications; while 10 participants were still striving for their first publication. Moreover, according to their self-reported current research topics, we classified them into three types, i.e., AI-related, Other STEM (Science, technology, engineering, and mathematics), and Social Science & Business. Basically, we tried to balance the participant distribution of background within each experiment condition.

Significantly, the study focused on participants with limited exposure to LLM in their academic career, specifically those who "sometimes used LLMs for academic purposes" or less. This criterion was adopted given that frequent users of LLM tools may have developed their own strategies for using the LLM, which can hardly be controlled in the experiment. More importantly, as illustrated in previous human-automation interaction domains [31, 32], new users may encounter performance and trust degradation when using unfamiliar LLMs. Thus, focusing on this group of users can help optimize the design of LLMs to better support the users.

Table 1: Background Statistics of Each Group of Participants.

Background	Type	Group 1	Group 2
Academic Position	Ph.D. student	11	11
	Mphil student	8	9
	Research assistant	5	4
Publication Number	0	3	7
	1-3	19	16
	4-6	2	1
Gender	Male	15	15
	Female	9	9
Usage Experience	Never used	3	3
	Rarely used	9	9
	Sometimes used	12	12
Research Interest	AI-related	3	3
	Other STEM	13	12
	Social Science & Business	8	9

Notes: In this table and the following tables, Group 1 refers to the group of participants who received additional training on LLM limitations; while the participants in Group 2 only receive basic training on how to use the LLM. In our recruitment questionnaire, options of LLM tool usage experience include: Never used; Rarely used; Occasionally used, but not frequent; Sometimes used - about half the time.

3.2 Tasks

Two types of academic tasks (i.e., literature review (LR), and paper understanding (PU)) were used in the study, given that they require different levels of skills, and are the type of tasks that LLM users need to pay the most attention to [16]. The literature review

task requires the users to search and identify relevant information when the sources of the information are unknown but the targeted information is partially known; while in the paper understanding task, the source of the information is known, but the information is unknown. In addition, given that the task complexity can moderate the relationships between users' trust and reliance on the system [54] and that time pressure is common in academia [11], we set two levels of time constraints for the tasks (i.e., 10 minutes, and 20 minutes) to construct comparable pair under the same task type. These two-time limits were set based on users' feedback in pilot tests. Note that, given that we aim to understand how young scholars use LLM for academic tasks, it would be unfair comparisons if the selected topics are within the research domain the participants are familiar with. Thus, we chose to provide experimental materials (i.e., scientific papers and review topics) from a field that no participants were from nor familiar with so that all participants were at a similar level of familiarity with the materials. We ended up choosing the topics from the human factors in transportation, because this field is minor in our target universities, and the experimenters are all familiar with this field. This decision was also based on the belief that the human factors domain has long been regarded as 'common sense' [61]. While this assumption may not be entirely accurate, it suggests that the research in this area should be relatively comprehensible for non-experts. The task type and time constraints led to four task conditions as follows:

- **Paper Understanding-More Time** (PU-MT) Given a published scientific paper, answer five questions related to the paper we provided within **20** minutes.
- **Paper Understanding-Limited Time** (PU-LT) Given another published scientific paper, answer the other five questions related to the paper we provided within **10** minutes.
- **Literature Review-More Time** (LR-MT) Given a topic, complete a literature review of approximately 500 words on the topic within **20** minutes.
- **Literature Review-Limited Time** (LR-LT) Given another topic, complete a literature review of approximately 500 words on the topic within **10** minutes.

To control the level of difficulties within each experimental condition, for the PU task, we used five similar questions regarding the two target papers, and the two papers were of the same length (5 pages) and were from the same academic conference in the same year; for the paper understanding task, we chose the topics from the same fields and a preliminary search in Google Scholar showed that the two tasks yield a similar number of publications in recent years.

ChatGPT (GPT3.5 version [4]), a popular LLM tool that leverages advanced language technology, was selected for the experiment. To maintain fairness, we restricted the use of other LLM tools, allowing only the official ChatGPT interface. While it is challenging to determine the popularity of such tools, ChatGPT appeared to be the most widely recognized at the time of the study. Participants were free to use ChatGPT for the tasks when they felt necessary. We established a virtual machine (VM) on Microsoft Azure for users to get access to ChatGPT. The VM also featured pre-installed Google Chrome [5] and Microsoft Office Packages. Mimicking real-world situations, participants were permitted to use Chrome and ChatGPT during tasks voluntarily. All experiments were conducted in the same meeting room with minimal external interference.

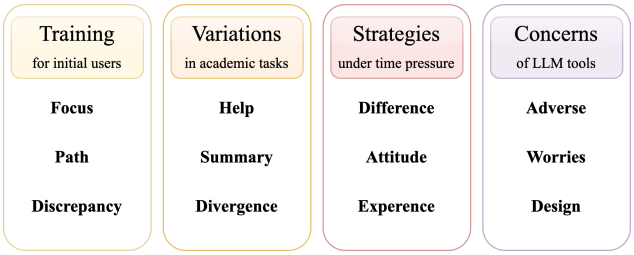

Figure 1: Coding framework and themes.

3.3 Experiment Design and Procedures

In addition to the four within-subject experimental conditions, we also provided participants with or without training materials regarding the limitations and potential errors that may be made by the LLMs on top of the basic training for using LLM tools (e.g., how to turn on the interface and what LLM can do in general) as the between-subjects factor. Thus, participants were divided into two between-subjects study groups, one of which (Group 1) was informed about potential errors and limitations of the selected ChatGPT in a pre-experiment video (provided in the supplement materials), while the other group (Group 2) was not.

Upon arrival, participants' informed consent was obtained. Subsequently, participants received basic training (around 10 minutes) regarding how to use the LLM. Half of the participants received an additional pre-experiment training video regarding the LLM limitations. Participants were then asked to complete four academic tasks (i.e., PU-MT, PU-LT, LR-MT, LR-LT) on the same laptop. To eliminate the learning and fatigue effects from task execution order, we counterbalanced the four experimental conditions. Throughout the experiment, the experimenter strictly adhered to a non-intrusive approach, refraining from interrupting the participants unless they sought assistance unrelated to the ongoing tasks. Participants were allowed a maximum of 5 minutes break between two tasks. Following the experiment, we conducted semi-structured interviews. The questions used in the semi-structured interview can be found in Appendix C.

The entire experiment lasted approximately two hours and participants received 120 Chinese Yuan as compensation. The Hong Kong University of Science and Technology's Human and Artefacts Research Ethics Committee approved this study (protocol number: HREP-2023-0159).

3.4 Analysis and Coding Methods

3.4.1 Quantitative Analysis. Two experts (senior Ph.D. students who had authored at least one peer-reviewed publication in the field) from the human factors field were invited to evaluate the answers participants generated. The two raters followed the same scoring standard that was decided before they started the evaluation and

[4] https://openai.com/chatgpt

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

Table 2: Individual and Group Performance Statistics of Paper Understanding.

PU-MT				PU-LT			
Group 1		Group 2		Group 1		Group 2	
LLM Usage	Grade:Time	LLM Usage	Grade:Time	LLM Usage	Grade:Time	LLM Usage	Grade:Time
y	90:100	n	70:100	y	70:100	y	15:100
y	75:80	y	65:100	y	60:100	y	5:100
n	70:100	y	55:100	n	75:100	y	95:100
y	55:100	y	75:80	y	65:100	y	65:100
y	85:85	y	75:100	y	65:100	y	85:90
y	65:90	n	60:100	y	80:100	y	60:100
y	80:100	y	75:100	y	90:100	y	45:100
y	80:100	y	85:95	y	90:100	y	80:50
y	80:85	n	75:100	y	70:100	n	50:100
y	90:95	y	90:80	y	70:100	y	85:100
y	80:80	y	65:100	y	65:90	y	45:100
y	75:100	y	55:90	y	55:100	y	60:100
n	35:100	y	20:100	y	55:100	y	90:100
y	75:75	y	85:90	y	80:100	y	85:100
y	50:100	y	70:70	y	40:100	y	50:100
y	55:100	y	65:100	n	60:100	y	50:100
y	30:100	y	45:100	y	15:100	n	65:100
y	75:100	n	50:80	y	70:100	y	25:100
y	85:80	y	65:90	y	30:100	n	75:80
n	35:100	y	85:100	n	35:100	y	80:100
y	50:100	n	75:100	y	40:100	n	55:100
y	90:100	n	55:100	y	70:100	n	75:100
y	65:100	y	20:100	y	65:100	y	25:100
y	50:100	n	45:100	n	25:100	y	20:100
n/y=3/21	average=67.5:94.6	n/y=7/17	average=63.5:94.8	n/y=4/20	average=60.0:99.6	n/y=5/19	average=57.7:94.8

Notes: The column 'LLM Usage' indicates whether participants used the LLM tool to assist in completing the task. The symbol 'y' means 'used' and 'n' is 'not used' in each task. The unit of time is minute.

conducted the evaluation independently. An Intraclass Correlation Coefficient (ICC) analysis was conducted and the two raters reached an ICC of 0.94 (95% CI = [0.91, 0.97], $p < .0001$), which indicates high consistency and inter-rater reliability of the grades (i.e., from 0 to 100). The guiding principles employed for scoring, as well as detailed experimental materials, can be found in Appendix B. Other metrics of the task performance in the empirical experiment part of the study include task completion Time (%) (i.e., the actual completion time / the time allowed for the current task.) and LLM tool adoption rate (i.e., the number of participants in each group who used LLM during that task / the total number of participants in each group). The criterion of LLM tool usage is whether they had fully accepted responses by LLM in their answers while fulfilling each task.

For the quantitative analysis method, in order to quantify the combined effects of participants' background and controlled experiment conditions, regression analyses were performed by "SAS OnDemand for Academics". Mixed linear regression models (using Proc MIXED) were built for two continuous dependent variables (Time and Grade), which included all demographic factors, three experimental conditions, and their two-way interactions as independent variables. Repeated measures were accounted for through a generalized estimating equation, which can be used to model multiple responses from a single subject. Backward stepwise selection procedures were employed based on model fitting criteria and 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 [38] were conducted. Variables demonstrating a significance level of $p < .05$ were considered statistically significant in the analyses.

3.4.2 Qualitative Analysis. As for the answers in the semi-structured interview, Figure 1 illustrates the coding framework and its corresponding themes. We transcribed interviews from 48 participants using automated transcription software[6], followed by content calibration to ensure the alignment between the original audio and transcribed text. Our approach blends the strengths of qualitative and quantitative analysis to investigate textual content. This dual approach not only facilitates more robust inferences but also opens avenues for additional reflection, hypothesis refinement, and further investigation [47].

To gain a deeper understanding of the interview content, two researchers (co-authors of this paper) identified several topics of interest based on the research questions and interview outline, including training, academic task types, pressure, concerns, and individual differences. They independently read all the interview texts, and extracted segments related to these topics. At the same time, they performed open coding (i.e., taking apart the information collected, assigning concepts, and then reassembling it in new ways) and

[6]https://www.feishu.cn/product/minutes

Table 3: Individual and Group Performance Statistics of Literature Review.

LR-MT				LR-LT			
Group 1		Group 2		Group 1		Group 2	
LLM Usage	Grade:Time	LLM Usage	Grade:Time	LLM Usage	Grade:Time	LLM Usage	Grade:Time
y	50:85	y	46:90	y	65:70	y	26:90
y	58:85	y	46:100	y	33:100	y	40:100
y	41:100	y	49:60	y	26:100	y	60:100
y	46:100	y	59:55	y	17:100	y	34:100
y	58:75	y	50:75	y	50:100	y	18:100
y	66:100	y	55:90	y	46:100	y	41:80
y	52:95	y	42:70	y	31:100	y	27:100
y	41:100	y	38:100	y	36:100	y	50:60
y	31:100	y	32:55	y	33:100	y	47:40
y	26:100	y	50:100	y	11:100	y	32:100
y	41:65	y	48:100	y	41:70	y	34:100
y	53:85	y	42:75	y	49:50	y	45:100
y	12:100	y	42:75	y	23:100	y	16:100
y	43:100	y	50:100	y	29:100	y	47:100
y	15:85	y	55:80	y	12:100	y	38:100
y	30:100	y	26:100	y	33:100	y	30:100
y	12:100	y	39:100	y	3:100	y	16:100
y	83:100	y	28:100	y	80:100	y	44:100
y	41:75	y	17:100	y	51:100	y	16:100
y	17:100	y	40:100	y	19:100	y	3:100
y	64:100	y	37:100	y	63:100	y	45:100
y	39:100	y	37:100	y	49:100	y	45:100
y	69:60	y	10:100	y	67:100	y	23:100
y	13:100	y	33:80	y	14:100	y	33:100
n/y=0/24	average=41.7:92.1	n/y=0/24	average=40.5:87.7	n/y=0/24	average=36.7:95.4	n/y=0/24	average=33.8:94.6

assigned descriptive labels to key paragraphs or viewpoints in the text. Subsequently, the two researchers jointly integrated the high-frequency repetitive labels and established overall themes based on their discussions. To achieve a comprehensive understanding and delve deeper into the participants' perspectives, attitudes, and emotions, the two researchers jointly developed broader categories and labels. By comparing and integrating different themes, they established the final coding framework, capturing the core content of the discussion through keywords, phrases, or topic sentences.

After completing all coding work, a third researcher (another co-author of this paper) joined the discussion to check and further analyze the constructed coding framework. In addition, we conducted a statistical analysis of the qualitative data to explore the frequency, distribution, and correlations of the coding. Finally, the three researchers discussed the results of the qualitative and quantitative analyses, cross-referenced, and synthesized each thematic category to analyze the participants' strategies, attitudes, and perceived changes when using LLM tools. It is important to note that the interview outline took open-ended and semi-open-ended questions, and during the course of the interviews, we were flexible in adjusting the questions based on the responses of the interviewees. Thus, not all 48 interviewees were asked and responded to the same questions, even though the outline of the interviews was fixed. Therefore, only the interviewees who responded to a particular question were coded and discussed rather than the entire group of 48 interviewees.

4 RESULTS

In this section, we present both quantitative and qualitative results of the study.

4.1 Quantitative results from the empirical experiment

As mentioned previously, the quantitative metrics are extracted from the empirical experiment and are summarized in Table 2 and 3. The results of the regression analysis are shown in Table 4, 5.

To verify the difficulty of experimental material before formal quantitative analysis, through paired samples t-tests, we first compared the scores of two cohorts who conducted the same type of academic task with different materials but under the same experimental conditions (same time pressure, and same training level). We did not find a significant discrepancy neither between the scores of the two cohorts who read different papers nor between those who performed literature review for different topics (p>.05). Therefore, we argue that the difference in the difficulty levels of the experimental materials we prepared for the same type of academic tasks was minor.

Next, as shown in the last rows of Table 2 and Table 3, under low time pressure, 10 out of 48 participants chose to finish the paper understanding task without the LLM tool; while under high time pressure, 9 participants chose to finish the tasks without LLM. Among these 9 participants, un-use of LLM occurred only in paper understanding tasks, and all participants chose to use LLM tools

Table 4: Summary of Statistical Results.

Dependent Variable (DV)	Independent Variable (IV)	F-value	*p*
Time	Usage Experience	F(2, 42) = 2.55	.09*
	Training	F(1, 42) = 7.50	.009**
	Training * Usage Experience	F(2, 42) = 4.84	.01**
	Task Type	F(1, 47) = 4.64	.04**
	Time Pressure	F(1, 47) = 5.51	.02**
Grade	Usage Experience	F(2, 45) = 2.21	.1
	Task Type	F(1, 45) = 105.19	<.0001**
	Usage Experience * Task Type	F(2, 45) = 8.98	.0005**
	Time Pressure	F(1, 47) = 9.59	.003**

Notes: In this table and the following tables, * marks marginal significant results (p<.1), ** marks significant results (p<.05).

Table 5: Significant Post-hoc Results for Discrete Independent Variables.

DV	IV	IV Level	IV Level compared to	Estimation (95% CI)	t value	*p*
Time	Task Type	PU	LR	3.49 [0.23, 6.79]	t(47)=2.15	.04**
	Time Pressure	MT	LT	-3.80 [-7.06, -0.54]	t(47)=-2.35	.02**
	Training	Without training	With training	-5.72 [-9.93, -1.50]	t(42)=-2.74	.009**
	Training*Usage Experience	Without training*Never used	Without training*Sometimes used	-13.65 [-25.40, -1.90]	t(42)=-3.47	.015**
		Without training*Never used	With training*Never used	-15.00 [-29.86, -0.14]	t(42)=-3.01	.047**
Grade	Task Type	PU	LR	24.60 [19.77, 29.43]	t(45)=10.26	<.0001**
	Time Pressure	MT	LT	6.26 [2.19, 10.33]	t(47)=3.10	.003**
	Usage Experience*Task Type	Never used*PU	Never used*LR	30.00 [12.99, 47.02]	t(45)=5.25	<.0001**
		Rarely used*PU	Rarely used*LR	12.97 [3.15, 22.80]	t(45)=3.93	.004**
		Rarely used*PU	Sometimes used*PU	-16.25 [-30.11, -2.39]	t(45)=-3.49	.01**
		Sometimes used*PU	Sometimes used*LR	30.83 [22.34, 39.34]	t(45)=10.78	<.0001**

Notes: estimate is the difference between IV level and IV level compared to.

when they were conducting the literature review task. This indicates that scholars presented varied preferences for the use of LLM tools on different tasks. Refer to [13], such attitude may be determined by the perceived ease of use and usability of the tool, which we will further discuss in qualitative analysis.

Second, to better model the influence of users' background and three experimental conditions, as well as their interaction effects, we built two models for Time (%) and Grade of participants. Refer to Table 4, we found that the type of training, task type and time pressure were significant predictors of time spent on task; task type and time pressure were influential factors of grades. Specifically, as shown in Table 5, one would spend more time and gain higher scores when conducting a paper understanding task compared to when conducting a literature review task. At the same time, people under higher time pressure spent a higher percentage of time on tasks but obtained lower scores. We also found that the training made a difference - participants who received limitation-related training used more time on task compared to those who received only basic training, while no significant effects of training were observed on grades. Finally, two significant interaction effects related to LLM usage experience were identified. We found that, within the group without limitation, participants who had more experience in utilizing LLMs in academic tasks spent more time in tasks compared to those who used LLMs less frequently. At the same time, when conducting paper understanding tasks, more experienced LLM users were always more likely to obtain higher grades compared to less experienced users of LLM.

4.2 Qualitative results from semi-structure interview

We extracted four categories of topics from the interview: training for initial users, variations in two academic tasks, strategies under time pressure, and concerns about LLM tools. Each category was further divided into three subtopics, which encompass the common themes emerging from participants' responses. Figure 2 illustrates the detailed statistics. Through coding and discussing diverse topics, we aim to delve into participants' attitudes, strategies, and reflections on various aspects of LLM tools.

4.2.1 Training for initial users. The majority (47/48) of participants would like to obtain some kind of guidance or training before using the LLM tools, but one subject explicitly stated that she did not need to know any information or knowledge to use the LLM tool for the first time and she could use it in a straightforward away. A total of 16 types of information that participants wished to know before using the LLM were identified, and these were categorized into 6 themes through thematic analysis. These categories, listed in descending order of frequency in Figure 2, are pre-use techniques, features and limitations of LLM, basic methods and operations to use LLM, ethics and compliance, historical or current tool development, and others. In addition to the most frequently mentioned and

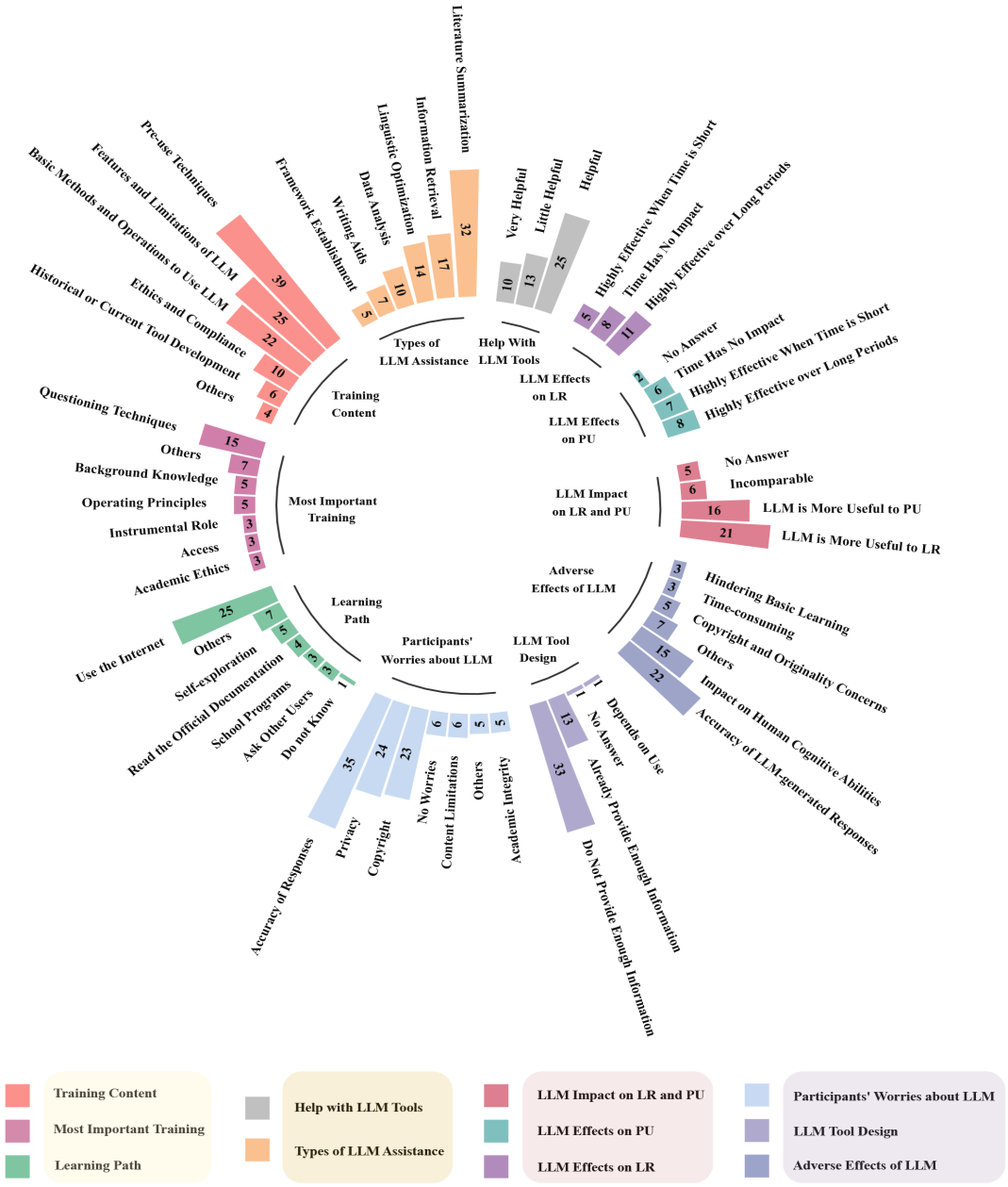

Figure 2: Interview quantitative statistical data. The categorization of coding and theme group corresponds to Figure 1.

emphasized "questioning techniques", many participants emphasized the importance of crafting effective "prompts." For instance, P43 said, "*If you don't use prompt engineering and instead express yourself naturally, there's a good chance you won't get what you're looking for; if you don't get the results you intended, LLM is not very useful in academic assignments.*" It is noteworthy that although 16 participants felt that understanding the limitations or flaws of the LLM tool was necessary, only 2 of them considered it the most crucial skill when using LLM for academic tasks.

When discussing the learning resources for the LLM tool, we find that the official guides or documentation provided by the LLM tools were not the primary learning resources. Only 4 participants indicated that they would read or watch the official learning materials. Most individuals tended to rely on third-party educational resources when acquiring new skills. P23 mentioned that he would learn how to use LLM tools through some user-generated content platforms; P38 said that he would check out posts shared on the Internet to learn; P30 emphasized the role of watching reviews

of LLM tools through short video platforms; and P40 said that he would check online forums or use a search engine to find relevant information.

To compare the effect of different training, we provided general training for all participants and conducted limitation-related training for half of the participants, in which we emphasized the shortcomings, limitations, and academic integrity issues related to the LLM tool. By comparing these two training methods, we found that the individuals who did not receive LLM-limitation-related training expressed greater satisfaction with the actual effectiveness of the tools and a higher percentage of them (83.3% versus 62.5%) believed that the LLM tool provided important assistance in completing the tasks. Further, participants who did NOT receive limitation-related training mentioned more content outside of our 'limitation-related training', e.g., they mentioned limitations of content generation more frequently in the semi-structured interview compared to those who received limitation-related training. For example, p8 said *"The current training data of LLM is also based on a more general data site, so my current experience is that there is still a lack of specialized knowledge. The generated answers are still limited and not professional enough."*

4.2.2 Variations in the role of LLMs in different academic tasks. Only 27% of individuals stated that the LLM tool was merely useful in assisting the two academic tasks in the experiment, while the rest of the respondents indicated that the LLM tool was useful to some extent. For example, P37 said, *"I find the ChatGPT very helpful, especially for summarizing existing literature and quickly locating answers. I think it's incredibly useful."*

Regarding the types of assistance gained from the LLM tools, we have categorized them into six primary themes using thematic analysis, ordered by frequency from high to low: 1. literature summarization, which aims to help users understand and summarise the content of the literature; 2. information retrieval, which helps find relevant information for a specific problem or to give advice on how to solve the problem; 3. linguistic optimization, which involves polishing the texts, and correcting grammar, spelling, and expression; 4. data analysis, which helps users process and analyze data; 5. writing aids, which support users with writing inspiration, content continuation, and so on; 6. framework establishment, which helps users create a framework or structure to present their ideas or research results. Figure 2 presents detailed data on these six themes.

Information search is a noteworthy feature of LLM tools, which is believed to have the potential to replace search engines and encyclopedias. As P15 said, *"I study chemistry, and when I come across some unfamiliar compounds, I will ask the LLM tool directly, which is more accurate and direct than the results obtained from a search engine."* P8 also mentioned that "asking questions to the LLM tool is like asking a Wikipedia." It is worth mentioning that, a few participants (2/48) mentioned assistance of LLMs in personalized tasks (e.g., Language translation, coding). For example, P37 mentioned that *"the LLM tool can judge my solutions, then identify some shortcomings, and help me to correct them"*

It is also interesting to find that participants exhibit significant divergence regarding the role of the LLM tool in different tasks. As illustrated in Figure 2, 21 respondents believed that the LLM was more helpful in assisting literature review tasks compared to in assisting paper understanding tasks. P19 said, *"I think it (LLM) was more useful in the Literature review, it not only helps us target some key information but at the same time relieves us of writing burdens."* In contrast, 16 participants held an opposite view, and the rest 6 participants expressed uncertainty about the comparison of the role of the LLM in two tasks used in the experiment. For example, P21 mentioned that *"The LLM is more useful in supporting paper understanding. The LLM tool can give me a general outline. It can explain terms I don't understand, and it also can summarise the paper a little bit." "When it comes to the literature review, I think it's better to refer to relevant published literature reviews that are more capable or conduct this myself, instead of referring to a bunch of literature summarized by the current LLM tool."*

In addition, some thought-provoking ideas were identified. For example, P41 emphasized that: *"Reading a paper is a process of comprehension, and the use of the LLM tool removes this purpose.* Implicitly, there is a concern that LLM tools may negatively impact one's capability in reading and comprehension of paper. While, P13 indicated that LLM can help the comprehension process as he mentioned that using LLM for paper reading is like *"going on a treasure hunt with a treasure map"*, highlighting the function of LLM tools as an aid. Nevertheless, although the LLM tool can guide and speed up the paper reading task, a deeper comprehension of the paper still requires the involvement of one's personal reflection.

4.2.3 Strategies under time pressure. Under different levels of time pressure, there are significant divergences in the impact of LLM tools on paper understanding and literature review tasks. 21 out of 48 participants felt that the LLM tool was more useful under less time pressure compared to that under high time pressure. For instance, P40 said, *"During the literature review, the tool is more useful when 20 minutes were allowed for the task."*. At the same time, 10 out of 48 participants felt that the LLM tool worked better under higher time pressure compared to that under low time pressure. The rest 17 participants thought that the time pressure did not make a difference.

At the same time, referring to Figure 2, 23 out of 48 participants specifically compared the role of the LLM tool in completing the paper understanding task under different time pressures. Among these 23 participants, 6 participants believed that time pressure would not affect the completion of the paper understanding task. In comparison, 2 participants stated that they did not use the LLM tool at all in the paper understanding task regardless of time pressure. Further, 24 out of 48 participants specifically compared the role of the LLM tool in completing the literature review task under different time pressures. Nearly half (11/24) of the participants indicated that the LLM tool would be more effective under low time pressure, while 5 participants held the opposite opinion.

These discrepancies and divergences also led to variations in participants' attitudes toward tasks under different levels of time pressure. We employed creative coding to differentiate these attitudes and found that under lower time pressure, participants tended to exhibit more positive attitudes toward LLM. When the time pressure was low, only 3 respondents regarded the help from LLM tools as ignorable, while the rest held positive attitudes towards LLM in accomplishing academic tasks. It is likely that as the time pressure reduced, participants could engage more in introspective

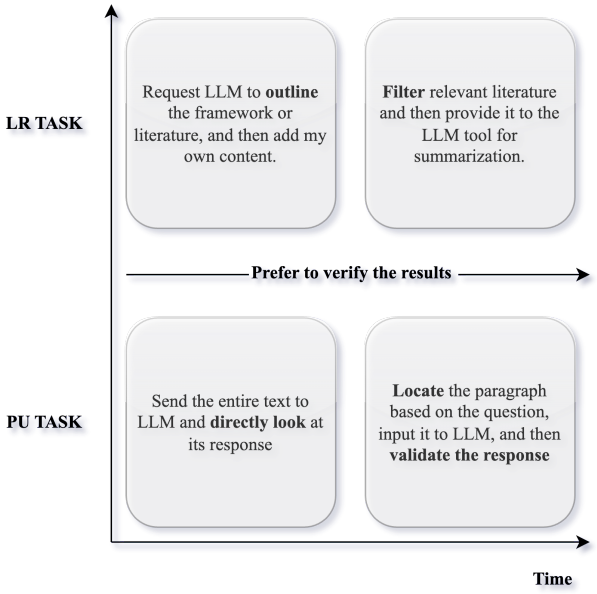

Figure 3: The 2x2 LLM tool usage strategies matrix. The x-axis indicates the time dimension, where more time means less time pressure. The two rows in the y-axis represent the Literature Review and Paper Understanding tasks, respectively.

thinking (e.g., contemplating ways to ask questions, strategies of using LLM, and double-checking the accuracy of the responses generated by LLM), which has been mentioned in the interview. Conversely, when time pressure got higher, almost all participants leaned towards negative attitudes toward the LLM, mainly due to concerns about the lack of time to check the replies generated by LLM. Interestingly, a few participants chose to prioritize completion of the task over concerns about the LLM, e.g., P8 said *"I will use the LLM to generate an approximate answer to satisfy the basic requirement of completing the task first, and then check if the answer is what I want when I have the time later"*. These distinct attitudes also affected participants' strategies in using LLMs. Figure 3 summarises the most widely adopted strategies when using the LLM tool in different situations. As time pressure decreased, participants showed a stronger willingness to examine the content generated by the LLM tool.

The lack of familiarity with LLM was one of the primary obstacles preventing the timely completion of tasks in the study. Throughout the interviews, the reasons for not being able to finish tasks within the designated time frame were mentioned 15 times. Some highly-mentioned reasons include unfamiliarity with the LLM tool (mentioned 3 times), slow responses to LLM (mentioned 7 times), and inaccurate time management (mentioned 5 times). Notably, unfamiliarity with using LLM tools stood out as one of the prominent barriers, as mentioned by P25, *"I might not be proficient with that software, so when I use it, I feel a bit flustered."* Hence, appropriate training for tool usage should be necessary. Throughout the interviews, there were a total of 39 instances of tasks being completed ahead of schedule. Most cases happened in literature review tasks with low time pressure, where 18 participants completed tasks ahead of schedule. The primary facilitating factor for early completion was experience in usage, as indicated by P45: *"After completing the task once, I gained experience or a sense of how to finish this task more quickly."* Participants were likely to become more familiar with the experimental process, leading to better comprehension of response and optimized strategies. This familiarity can also play a role in real-world academic tasks, manifesting as increasing efficiency when conducting similar tasks or using LLM tools repeatedly. Another intriguing discovery was that some participants had to lower their interactions with the LLM tool due to time pressure. They stopped scrutinizing the generated content before adopting it, which paradoxically led to early task completion. For example, P44 said, *"Because it might just be time pressure, I didn't expect as much from the LLM tool. So I didn't bother to make any further adjustments to the answer, and finished the task ahead of schedule."*. This raises concerns about over-reliance on AI tools in high-pressure situations [8, 54].

4.2.4 Concerns about LLM tools. Most participants expressed concerns about the impacts of the LLM tool. The top five most frequently mentioned negative effects include 1) concerns about the accuracy of LLM-generated responses, where the answers provided by the tool may be erroneous or imprecise; 2) impact on human cognitive abilities, where overuse of the tool may weaken the user's ability to think independently and dependency on LLM tools may develop; 3) copyright and originality concerns, wherein the tool may infringe upon others' intellectual property rights while generating content and users may be questioned about the originality of their work when utilizing AI-generated materials; 4) time-consuming, where users might spend excessive amounts of time seeking accurate answers or rectifying incorrect content; 5) hindering basic learning, where users may stop developing basic skills due to over-reliance on the tool. The statistics of these five negative impacts can be found in Figure 2. For instance, P38 said, *"Relying too much on the LLM for assistance in academic tasks might lead to academic misconduct or errors within the academic process."* He also mentioned, *"If you overly depend on this tool and turn to it for solutions whenever you encounter problems, it might hinder critical thinking and innovation by impeding our natural thought processes."*

More specifically, among the 48 participants, concerns regarding the current LLM tool are primarily about the accuracy of responses rather than other issues such as privacy and copyright. Indeed, although many participants did not explicitly mention concerns over the accuracy of their responses, they always mentioned this concern implicitly in their words. For example, though P14 did not mention the accuracy issue directly, he still expressed concerns about the correctness of the LLM-generated content when describing his strategy in using LLM: *"I may double-check the LLM responses, and beyond the logic, I will also pay attention to some of the parts that may not match my perception and may do further validation "*. Surprisingly, 5 participants indicated that they did not have any concerns about the LLM tool. For example, P42 mentions that *"LLM has not been used in a particularly bad way, so I do not have any obvious concerns,"* while P19 also indicates that *"I have no concern about LLM. I think that as long as the responses are scrutinized and*

full, while reasonable inputs are provided, a high degree of accuracy can be achieved.

When the academic background of participants is considered, we were also surprised to find that none of the participants without publication mentioned copyright concerns proactively. Even after reminding, only 2 (out of 10) of them said they would consider the copyright as an issue. Further, only one of them mentioned academic integrity issues when using LLM. While for participants who had at least one publication, a larger portion of them (23 out of 38) regarded the copyright or academic integrity as a potential issue of using LLM. For example, P46 said "There may be academic misconduct ..... I'm also afraid that my intellectual property will be compromised. I prefer not to send the paper I'm working on directly to LLM. Instead, I'll probably send small segments and have the GPT do some writing polishing." It seems that researchers who received more extensive academic training were more aware of violating academic rules when an AI-based tool was used.

Regarding the issues in the design of the LLM, the majority of participants believed that the current design of LLM tools does not provide users with sufficient information. Among the 48 participants, 33 of them expressed that the current design of LLM tools does not provide official guidance on how to use prompts efficiently. For example, P44 said, *"The design only offers an interface for input and output, but it doesn't provide specific guidance on how to better utilize and master the tool. Most of the learning comes from seeking information through other channels."* Similarly, P31 said, *"The interface is very simple, and the content is quite brief. It doesn't provide me with proper guidance."* However, some interviewees held different opinions. They believed that the simplicity of the interface makes the tool easy to operate, as P39 mentioned, *"The LLM tool itself is quite simple. After having several conversations with it, you naturally become familiar with the pattern. It doesn't require excessive design."*

5 DISCUSSION

5.1 Strategies in using LLMs for different academic tasks

Combining results from quantitative and qualitative analysis, our research indicates that young scholars performed better when using LLM tools for paper understanding (PU) tasks compared to literature review (LR) tasks. However, young scholars spent more time and had lower intentions to use LLMs for PU tasks versus LR tasks. In the field of human-computer interaction, it is widely recognized that user reliance on automation can be moderated by task complexity [54]. Similarly, in our experiments, compared to the PU task, where the information source was known, participants needed to search a wider range of unknown sources in LR tasks. Further, most participants perceived LLM tools as being good at handling complex tasks such as developing process frameworks. Both may explain why participants relied on LLM more in LR tasks, especially when the time pressure was high, as many participants felt that copying and typing text from PDF files into LLM in PU tasks was more complicated and time-consuming compared to the procedures in LR tasks. However, it should be noted that, given the limitations of the LLM we provided in the experiment (i.e., source of bias [2]), current LLM tools cannot provide the most up-to-date results in LR tasks. Thus, the LLMs can provide limited assistance in LR tasks, which may explain why participants obtained lower scores in LR tasks (which were judged based on scoring standards like the number of literature, source accuracy, and citation quality. Details please refer to supplement materials) compared to PU tasks.

We also found that participants can adaptively change their strategies when conducting different types of academic tasks. In order to further reveal the participants' strategies, a flowchart was obtained by summarising the interview data and the experimenter's report of observation notes[21]. We combined various factors such as interview transcripts, task materials, task completion time, scores, and interaction styles to create a basic flowchart showing the process of completing the task for most of the participants (Figure 4). In the figure, we aggregated and abstracted the steps the majority of participants took.

In general, at the beginning of a task, participants would judge whether they need help from LLM tools, taking the task type, the difficulty of the task (as moderated by the time pressure), and their capabilities into consideration. Then, during their interactions with LLM tools in tasks, participants may repeatedly modify their strategies (e.g., adjusting the context in their prompts) to optimize the LLM-generated results. Different strategies were adopted for different types of tasks. Specifically, participants were highly uniform in their strategies when using LLM tools for the PU tasks. Most of them would divide the articles into small segments and ask questions based on the segments. When conducting LR tasks, participants chose a more diverse strategy. For example, some participants asked the LLM to generate a complete review; others only let the LLM generate the outline. Some participants even chose to provide a framework for the LLM to refer to. Figure 4 depicts two strategies that the participants used the most in LR tasks. Finally, another difference in strategies for the LR task and PU task was that participants usually used LLM throughout the whole task procedures for the PU task; whereas participants preferred to conduct self-modification and refinement for the responses generated by LLM tools in LR tasks. It is likely that the participants had different levels of trust in the LLM tool when completing different tasks, which led to different levels of reliance on the LLM tool in the tasks.

In addition, when conducting PU tasks, participants were more inclined to complete the task on their own compared to when conducting the LR (see Tables 2 and 3). Further statistical tests show that those who did not use the LLM tool obtained lower scores and took a significantly longer time to complete the task (paired t-test $p<.0001$). Based on the interviews, we found two potential reasons explaining the low usage rate of LLM in PU tasks. First, participants were confident in their ability to comprehend the scientific literature; second, they were trying to avoid the deterioration of their learning ability as a result of over-reliance on LLM. We speculate that the abandonment of LLM in tasks may also be related to one's personality traits and the early-stage scholars' wish to develop their skills for future academic success. However, in this experiment, we cannot validate these assumptions given that only early-stage scholars participated in the experiment, and future experiments are needed.

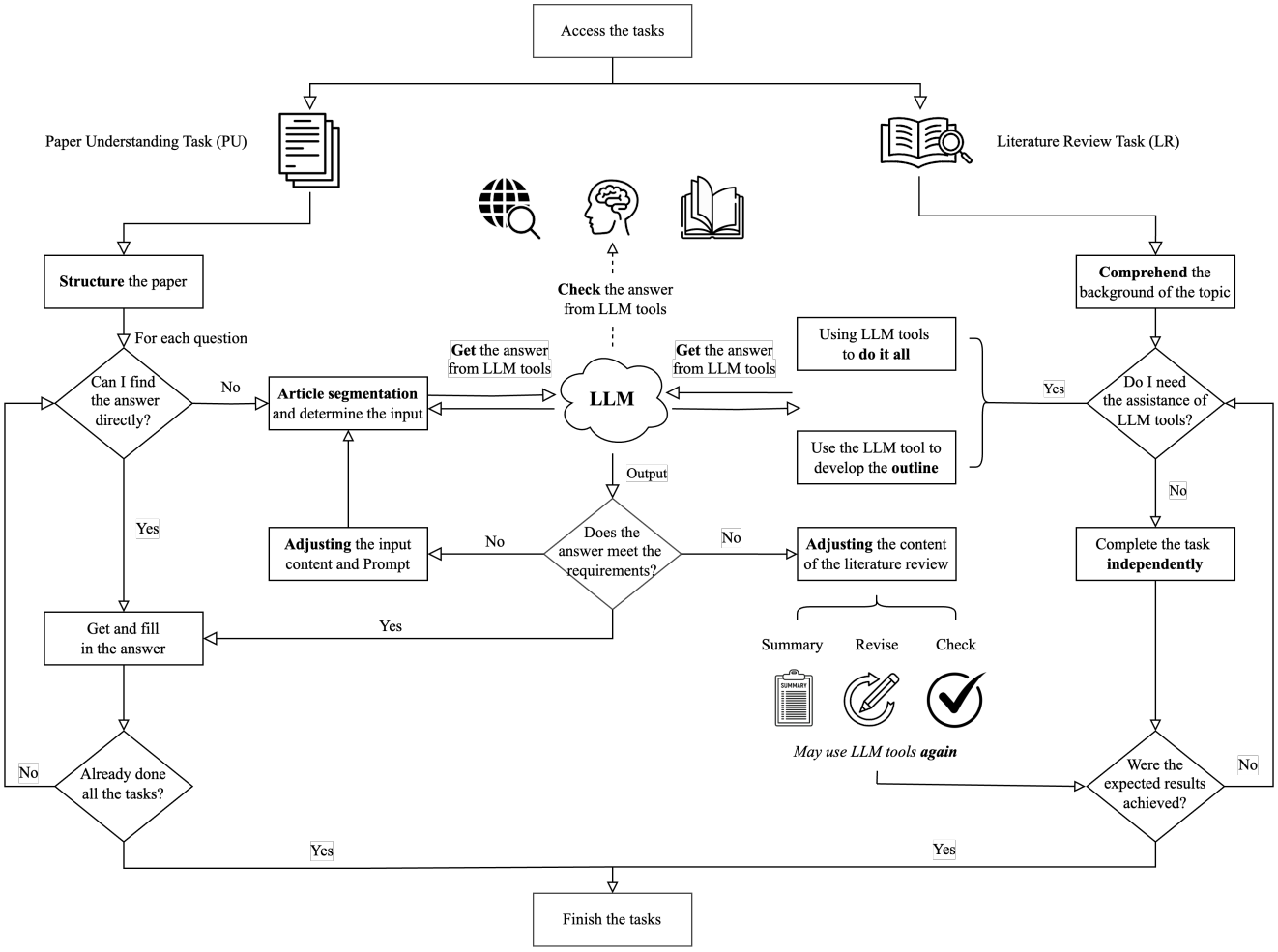

Figure 4: Flowchart of conducting two tasks with LLMs. Notes: This flowchart only summarizes the basic processes for the two tasks in the experiment. The actual behaviors of participants were more diverse when under different levels of time pressure, which are presented in Figure 3.

5.2 Strategic choices under time pressure

Time pressure can influence researchers' strategies when using the LLMs and their attitudes toward LLM tools. Although previous results have shown that time pressure can affect the strategies that users adopt to learn new knowledge or skills [50], it is still unknown how time pressure may influence one's strategies when an AI-based assistant, the LLM tool, was available for academic tasks. The qualitative analyses in our study indicate that with relatively low time pressure, researchers exhibited a more positive attitude toward the LLM and were more confident in fulfilling tasks using LLM tools. In contrast, under high time pressure, most researchers showed a more hesitant and negative attitude toward using LLM for academic tasks. It is possible that researchers still have concerns over the capability of LLM. Thus, under high time pressure, participants tended to adopt more conservative methods rather than use new tools [64], so that they do not need to double-check the content generated by LLM.

However, users' attitudes toward the LLM tools may not directly reflect their choice of strategies. The observational data in our study shows that, in PU tasks, with low time pressure, participants were more likely to abandon LLM tools; whereas under high time pressure, participants exhibited higher LLM tool usage rates. At the same time, under high time pressure, some participants chose to skip verifying the responses generated by LLM tools. This indicates that when faced with more urgent deadlines, researchers may prioritize efficiency over skepticism and potentially sacrifice the quality of their work. This result is in line with previous findings in the human-automation interaction domain, which suggested that external stressors of the tasks may influence their adoption of new technologies [41]. Specifically, when the users are under a high workload or in a stressful situation, they tend to rely more on the technologies, even if they do not fully trust them. Thus, LLM tool designers should carefully balance efficiency and effectiveness to better support users. For instance, the trade-offs between response speed and the accuracy of the responses can be customizable to

better cater to users' needs when they are under different levels of time pressure.

5.3 Users' attitudes and concerns of LLM

In general, researchers hold a positive and forward-looking attitude toward LLM tools. They mentioned more about the functionality of LLM tools and how to effectively utilize them, rather than the limitations of the tools. On the one hand, this is an encouraging discovery, as it suggests that young scholars focused more on harnessing the benefits of LLM tools rather than dwelling on the shortcomings and thus they may be more willing to use them. On the other hand, this may lead to misuse of the LLM tools. For example, during the interviews, although most participants mentioned their concerns about the limitations of the LLM tool (similar to the findings from [2, 29, 52]), very few participants could comprehensively and systematically acknowledge the constraints and boundaries of LLM tools and even fewer participants were awareness of the potential privacy and copyright issues of the LLMs. Especially, those with little academic experience (i.e., had no academic publications) were inclined to overlook the potential personal privacy, academic copyright, and ethics issues caused by LLM tools. This finding provides a different perspective on the opinions of adopting the LLMs for academic tasks compared to the previous study, which focused more on senior scholars [52]. Additionally, young scholars may intentionally choose to ignore the limitations of the LLMs, similar to how human beings rely on heuristics to make decisions in urgent situations [58]. In our study, under time pressure, some participants indicated that they intentionally ignored the deficiencies of LLM tools, even when they were aware of these issues. For instance, when striving to complete PU tasks under high time pressure, some participants indicated that they lowered their expectations of the performance of the LLM tool and may cease to verify the content generated by these tools.

The associations among academic experience, attitudes toward LLMs, and strategies when using LLM indicate that, in the academic community, users' willingness to use the LLM tools is a dynamic process and there is a chance that young scholars prefer to use the LLM tools, especially under high time pressure. Thus, LLM tool designers should try to make the users aware of the limitations and boundaries of LLM tools so that the users can use the LLMs more effectively and responsibly. For instance, appropriate system transparency [37, 49] can be an effective way to address the concerns regarding the accuracy of tool outputs. Specifically, designers can incorporate features such as confidence scores or explanatory annotations [60] in the responses generated by the LLMs, which would help users better understand the reliability of the generated content so that the users can make more informed decisions when using the tool, even under high time pressure. On the other hand, before adopting the LLMs, young scholars should also receive training regarding the limitations of LLM tools so that they can make more informed decisions on when and how to use the LLM tools.

At the same time, from the perspective of Human Machine Interface (HMI) design for LLM, understanding the differences in users' performance and level of reliance can aid in designing LLM tools to fit the needs of different academic tasks and improve users' satisfaction with the LLM tools. Sakirin et al. showed that users preferred to use the dialogue interface supported by the LLM tool [59], but our study found that the interface of the LLM tool did not provide users with enough information, and the oversimplified design may not give users enough hints and feedback. It is recommended that the functionality and interface of the LLM should be improved so that it is more suitable for different academic tasks. For example, the option of "Prompt" for different scenarios can be proactively provided, to reduce the learning cost of the researcher.

What is more alerting is that, some participants overstated the abilities of LLM tools (i.e., overlooked potential limitations or risks in the LLM usage to academic tasks), which coincides with several voices supporting the use of LLM tools in academic tasks [51]. However, overestimating the capabilities of the LLMs may lead to over-reliance on LLM tools. From an academic performance perspective of view, this may result in erroneous or inaccurate conclusions in academic tasks. From an educational perspective of view, this may negatively impact young scholars' critical thinking and academic skills, potentially affecting their overall academic development. This finding points to another important topic in LLM usage, the training of the users.

5.4 The role and future improvement of training

Training is pivotal for the appropriate use of the LLM tool. Previous research has pointed out that it is important to train users to refine their mental models, and subsequently facilitate user-LLM collaboration performance [65]. Our study reveals that individuals who received limitation-related training expressed lower satisfaction with the effectiveness of the LLM tool and discussed more of the accuracy of the LLM-generated responses in the post-experiment interview. This implies that the trained individuals were more skeptical of the content generated by the LLM, which may explain why, among the ones who never used LLM, those who received limitation-based training spent more time on academic tasks compared to those who did not receive limitation-related training.

It is also interesting to notice that, in addition to the experience passed in the limitation-related training, users can also gain experience during interactions with LLMs, before the experiment. For example, we found that compared to those who had little to no prior experience with the LLMs, the participants who had relatively more experience (i.e., those who self-reported sometimes using LLMs, of which 12 of them received limitation-based training and 12 did not) with the LLM tended to be more aware of the strengths and weaknesses of the LLM tool and tried to find the best strategies when using LLM tools. Specifically, they adjusted their interactions with LLM tools more constantly compared to those who had less experience and they also provided more insightful comments on LLMs. For example, 7 out of the 12 LLM users who did not receive limitation-based training mentioned that they double-checked and double-examined the answers generated by LLM, which cost additional time in the task. In contrast, users who lacked LLM experience and did not receive limitation-related training tended to show low confidence in the LLM tools. In particular, in the study, among the 3 participants who had no LLM experience and did not receive limitation-related training at the same time, 2 of them abandoned LLM tools during the PU task, and they did not

adopt the strategies most experienced users would take (as shown in Figure 4) in LR tasks. The above findings indicate that limitation-related training can not only shorten the period that users may take to develop appropriate strategies to use new technologies, but also help promote the adoption of LLMs among new users.

Unfortunately, the training or materials provided by the official providers of the LLM tools may not be enough. Many participants reported receiving their training from third-party platforms on the Internet rather than from official sources. This could be attributed to uncertainties regarding the comprehensibility of official documentation or the usability issues of the official documents. Such a phenomenon has also been observed in other domains. For example, researchers found that a very low portion of users read the manual of their vehicles regarding driving automation [20]. Therefore, it is recommended that LLM developers or maintainers explore better ways to present necessary information to new users, or actively engage with relevant online forums and social media groups to assist users in addressing their usage-related queries. However, it should be noted that the training methods adopted in our study are still preliminary, and future research should continue to optimize training methods and content, and better incorporate the training methods in the LLM tool design to improve users' performance in academic tasks with the LLM tools.

Especially, for the PU tasks, the results showed that most early-stage scholars would prefer to read and understand the literature on their own, as they did not want to "*rely too much on the LLM to constrain their learning ability*". Hence, future LLMs can provide more translation or search functions for key information in PU scenarios. For LR tasks, designers should try to reduce the chances of noisy responses appearing or provide confidence scores [60] for the LLM-generated responses. Personalized training and support services may also be necessary to help participants make better use of LLM tools. By tailoring support based on researchers' experience, proficiency with the tools, and the type of tasks they are undertaking, individualized assistance can be provided. This could include training on specific usage techniques and strategies for a particular type of task, thus enabling scholars to perform better in their academic endeavors.

6 LIMITATIONS

We recognize that although our study has followed standards in the field of human-computer interaction to some extent [10], there are still some limitations. First, as we intentionally limited our targeted user group to young scholars, the findings may not be well-generalizable to the senior academic community. Users with different levels of familiarity with academic tasks may hold different attitudes toward AI-based tools and may adopt different strategies for using them. In future research, senior scholars with different backgrounds should be recruited. Secondly, limited by the sample size, we had to focus on two types of common but typical academic tasks in a single academic domain in this experiment, which may not cover all scenarios when LLM tools are used in academic tasks. We also only considered the task difficulty controlled by the time pressure. In daily academic tasks, the task difficulties may be moderated by many factors. Future research may consider introducing more types of academic tasks (e.g., academic writing, data analysis, and experimental design) from more academic domains, and modeling the influence of other task-difficulty-related factors to assess the impact of LLM tools on academic tasks more comprehensively. Further, as an empirical study, though we tried to replicate realistic scenarios in daily life, the scenarios the users encountered were still artificial to some level and users may have biased behaviors in the experiment. Future research may consider observational studies to better reveal the strategies users may adopt when LLMs are used for academic tasks. Lastly, considering the rapid development of LLMs, more advanced models or interfaces are being introduced (e.g., GPT-4V [7], Semantic Reader [8]). In this work, we were not able to adopt these up-to-date tools, as they were not publicly accessible when our experiment began. Thus, the readers should be aware that some findings in our study may not apply to some emerging LLM tools and future assessments of how users' behaviors change adaptively with the evolvement of LLM tools are needed.

7 FINDINGS AND CONCLUSIONS

In this study, we conducted an empirical study involving 48 early-stage scholars, to understand how LLM tools can be utilized for academic tasks and affect early-stage scholars' workflow. Specifically, we discussed the influences of user perspectives on LLMs, evaluated users' performance when using LLM in two typical but different academic tasks, and analyzed the influence of time pressure in these tasks. Besides, the qualitative analysis based on a post-experiment interview revealed the strategies users adopted when using LLM tools. In general, several key findings are summarized as:

- We found that young scholars can adaptively change their strategies when using the LLM for different tasks. Specifically, we observed more diverse questioning styles and less reliance on the LLM tools when using LLM for LR task; while a more monotonic strategy was observed when the LLM was used for PU tasks. Future LLM design may consider customizing the tool to better satisfy users' needs in different scenarios.
- Time pressure can influence users' attitudes toward the LLM tools and the strategies they take to cooperate with the LLM. However, the strategies they took may not necessarily match the attitudes they hold. High time pressure led to declined attitudes toward the LLM, but increased the adoption rate of the LLM. It is likely that the users were not satisfied with the performance of LLM tools, but they had to use them to reduce the time pressure. Future LLM tools may need to allow users to customize the LLM tools to reach a balance between accuracy and efficiency.
- Young scholars had an overall positive attitude towards the LLMs in academic tasks, but due to their lack of academic experience, they were also inclined to ignore the academic ethical and privacy risks introduced by LLM tools, and tended to voluntarily give up their concentration on the risks from LLMs when the complexity of the task increases. Thus, training might be necessary to support better use of LLM tools among young scholars.

[7] https://openai.com/research/gpt-4v-system-card
[8] https://openreader.semanticscholar.org/

- We investigated a specific training, the limitation-based training, on users' attitudes towards strategies and performance in academic tasks when using the LLM tools. The results show that training can increase users' awareness of the limitations of the LLMs and lead to more appropriate strategies when using LLM tools, similar to what experience with LLM can do. Given that users criticized the current LLM developers for not providing adequate training materials, more authoritative and comprehensive online training materials for LLM are expected in the future.

REFERENCES

[1] Ian L Alberts, Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi, Axel Rominger, and Ali Afshar-Oromieh. 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 (2023), 1549–1552.

[2] Muath Alser and Ethan Waisberg. 2023. Concerns with the usage of ChatGPT in Academia and Medicine: A viewpoint. American Journal of Medicine Open 100036 (2023).

[3] Ömer Aydın and Enis Karaarslan. 2022. OpenAI ChatGPT generated literature review: Digital twin in healthcare. Available at SSRN 4308687 (2022).

[4] Stephen H Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, et al. 2022. Promptsource: An integrated development environment and repository for natural language prompts. arXiv preprint arXiv:2202.01279 (2022).

[5] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. 610–623.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021).

[7] Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: changing user needs and motivations. Interactions 25, 5 (2018), 38–43.

[8] Dawn Branley-Bell, Rebecca Whitworth, and Lynne Coventry. 2020. User trust and understanding of explainable ai: Exploring algorithm visualisations and user biases. In International Conference on Human-Computer Interaction. Springer, 382–399.

[9] Tim Broady, Amy Chan, and Peter Caputi. 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 (2010), 473–485.

[10] Kelly Caine. 2016. Local standards for sample size at CHI. In Proceedings of the 2016 CHI conference on human factors in computing systems. 981–992.

[11] Lydia Carson, Christoph Bartneck, and Kevin Voges. 2013. Over-competitiveness in academia: A literature review. Disruptive Science and Technology 1, 4 (2013), 183–190.

[12] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: visual sketching of story generation with pretrained language models. In CHI Conference on Human Factors in Computing Systems Extended Abstracts. 1–4.

[13] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly (1989), 319–340.

[14] Ismail Dergaa, Karim Chamari, Piotr Zmijewski, and Helmi Ben Saad. 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 (2023), 615–622.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

[16] Eva Dis, Johan Bollen, Willem Zuidema, Robert Rooij, and Claudi Bockting. 2023. ChatGPT: five priorities for research. Nature 614 (02 2023), 224–226. https://doi.org/10.1038/d41586-023-00288-7

[17] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. Advances in Neural Information Processing Systems 32 (2019).

[18] Michael Dowling and Brian Lucey. 2023. ChatGPT for (finance) research: The Bananarama conjecture. Finance Research Letters 53 (2023), 103662.

[19] Armin Esmaeilzadeh and Kazem Taghva. 2022. Text classification using neural network language model (nnlm) and bert: An empirical comparison. In Intelligent Systems and Applications: Proceedings of the 2021 Intelligent Systems Conference (IntelliSys) Volume 3. Springer, 175–189.

[20] Yannick Forster, Sebastian Hergeth, Frederik Naujoks, Josef Krems, and Andreas Keinath. 2019. User education in automated driving: Owner's manual and interactive tutorial support mental model formation and human-automation interaction. Information 10, 4 (2019), 143.

[21] Patricia H Fowler, Janet Craig, Lawrence D Fredendall, and Uzay Damali. 2008. Perioperative workflow: barriers to efficiency, risks, and satisfaction. AORN Journal 87, 1 (2008), 187–208.

[22] Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. BioRxiv (2022), 2022–12.

[23] Catherine A. Gao, Frederick M. Howard, Nikolay S. Markov, Emma C. Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T. Pearson. 2022. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. BioRxiv (2022). https://doi.org/10.1101/2022.12.23.521610 arXiv:https://www.biorxiv.org/content/early/2022/12/27/2022.12.23.521610.full.pdf

[24] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462 (2020).

[25] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In Designing Interactive Systems Conference. 1002–1019.

[26] Katy Ilonka Gero, Tao Long, and Lydia B Chilton. 2023. Social dynamics of AI support in creative writing. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–15.

[27] Bert Gordijn and Henk ten Have. 2023. ChatGPT: evolution or revolution? Medicine, Health Care and Philosophy 26, 1 (2023), 1–2.

[28] Gianluca Grimaldi and Bruno Ehrler. 2023. AI et al.: machines are about to change scientific publishing forever. ACS Energy Letters 8, 1 (2023), 878–880.

[29] Mohanad Halaweh. 2023. ChatGPT in education: Strategies for responsible implementation. (2023).

[30] Dengbo He, Dina Kanaan, and Birsen Donmez. 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 (2022), 772910.

[31] Sebastian Hergeth, Lutz Lorenz, and Josef F Krems. 2017. Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. Human factors 59, 3 (2017), 457–470.

[32] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors 57, 3 (2015), 407–434.

[33] Matthew Hutson. 2022. Could AI help you to write your next paper? Nature 611, 7934 (2022), 192–193.

[34] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? Transactions of the Association for Computational Linguistics 8 (2020), 423–438.

[35] Andreas Jungherr. 2023. Using ChatGPT and Other Large Language Model (LLM) Applications for Academic Paper Assignments. (2023).

[36] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences 103 (2023), 102274.

[37] René F Kizilcec. 2016. How much information? Effects of transparency on trust in an algorithmic interface. In Proceedings of the 2016 CHI conference on human factors in computing systems. 2390–2395.

[38] CY KRAMERß. 1956. Extension of multiple range tests to group means with unequal numbers of replication. Biometrics 12 (1956), 307–310.

[39] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLoS Digital Health 2, 2 (2023), e0000198.

[40] Ivano Lauriola, Alberto Lavelli, and Fabio Aiolli. 2022. An introduction to deep learning in natural language processing: Models, techniques, and tools. Neurocomputing 470 (2022), 443–456.

[41] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. Human Factors 46, 1 (2004), 50–80.

[42] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 (2019).

[43] Michael Liebrenz, Roman Schleifer, Anna Buadze, Dinesh Bhugra, and Alexander Smith. 2023. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. The Lancet Digital Health 5, 3 (2023), e105–e106.

[44] Peng Liu and Zhizhong Li. 2012. Task complexity: A review and conceptualization framework. International Journal of Industrial Ergonomics 42, 6 (2012), 553–568.
[45] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. Comput. Surveys 55, 9 (2023), 1–35.
[46] Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In 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.
[47] Thorleif Lund. 2012. Combining qualitative and quantitative approaches: Some arguments for mixed methods research. Scandinavian Journal of Educational Research 56, 2 (2012), 155–165.
[48] Muneer M Alshater. 2022. Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT. Available at SSRN (2022).
[49] Jan Maarten Schraagen, Sabin Kerwien Lopez, Carolin Schneider, Vivien Schneider, Stephanie Tönjes, and Emma Wiechmann. 2021. The role of transparency and explainability in automated systems. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 65. SAGE Publications Sage CA: Los Angeles, CA, 27–31.
[50] Giuliana Mazzoni and Cesare Cornoldi. 1993. Strategies in study time allocation: Why is study time sometimes not effective? Journal of Experimental Psychology: General 122, 1 (1993), 47.
[51] Jesse G Meyer, Ryan J Urbanowicz, Patrick CN Martin, Karen O'Connor, Ruowang Li, Pei-Chen Peng, Tiffani J Bright, Nicholas Tatonetti, Kyoung Jae Won, Graciela Gonzalez-Hernandez, et al. 2023. ChatGPT and large language models in academia: opportunities and challenges. BioData Mining 16, 1 (2023), 20.
[52] Meredith Ringel Morris. 2023. Scientists' Perspectives on the Potential for Generative AI in their Fields. arXiv preprint arXiv:2304.01420 (2023).
[53] Michael Muller, Lydia B Chilton, Anna Kantosalo, Charles Patrick Martin, and Greg Walsh. 2022. GenAICHI: generative AI and HCI. In CHI conference on human factors in computing systems extended abstracts. 1–7.
[54] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. Human Factors 39, 2 (1997), 230–253.
[55] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. arXiv preprint arXiv:2202.03286 (2022).
[56] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. Anglekindling: Supporting journalistic angle ideation with large language models. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–16.
[57] Md Mizanur Rahman, Harold Jan Terano, Md Nafizur Rahman, Aidin Salamzadeh, and Md Saidur Rahaman. 2023. ChatGPT and academic research: a review and recommendations based on practical examples. Journal of Education, Management and Development Studies 3, 1 (2023), 1–12.
[58] Torsten Reimer and Jörg Rieskamp. 2007. Fast and frugal heuristics. Encyclopedia of Social Psychology (2007), 346–348.
[59] Tam Sakirin and Rachid Ben Said. 2023. User preferences for ChatGPT-powered conversational interfaces versus traditional methods. Mesopotamian Journal of Computer Science 2023 (2023), 24–31.
[60] Anuschka Schmitt, Thiemo Wambsganss, and Andreas Janson. 2022. Designing for conversational system trustworthiness: the impact of model transparency on trust and task performance. (2022).
[61] Horrock Stevens. 2019. What Human Factors Isn't: 1. Common Sense. https://humanisticsystems.com/2019/07/10/what-human-factors-isnt-1-common-sense/
[62] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in Neural Information Processing Systems 30 (2017).
[64] Christopher D Wickens, William S Helton, Justin G Hollands, and Simon Banbury. 2021. Engineering psychology and human performance. Routledge.
[65] Bart D Wilkison, Arthur D Fisk, and Wendy A Rogers. 2007. Effects of mental model quality on collaborative system performance. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Vol. 51. SAGE Publications Sage CA: Los Angeles, CA, 1506–1510.
[66] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems 32 (2019).
[67] Lirong Yao and Yazhuo Guan. 2018. An improved LSTM structure for natural language processing. In 2018 IEEE International Conference of Safety Produce Informatization (IICSPI). IEEE, 565–569.
[68] Nan Zhong, Zhenxing Qian, and Xinpeng Zhang. 2021. Deep neural network retrieval. In Proceedings of the 29th ACM International Conference on Multimedia. 3455–3463.

A INTERVIEW QUANTITATIVE STATISTIC DATA

It is important to note that a large amount of data was obtained during the coding and analysis phases. Only the data involved in Figure 2 is shown in Table 6 as an additional illustration for ease of reference.

B EXPERIMENT MATERIALS AND QUESTIONS

B.1 Paper Understanding 1 (P1)

Paper link:
https://journals.sagepub.com/doi/pdf/10.1177/1071181322661442

(1) What is ADAS? Please list some typical functions it concludes.
(2) What is the purpose of this research, and how this study can benefit future studies?
(3) Please briefly describe the procedures of how the survey data was collected, the participants' criteria, and how valid data was selected after the data collection.
(4) Please briefly conclude the findings in this paper, and how these findings can benefit future studies.
(5) Please indicate the limitations of this paper.

B.2 Paper Understanding 2 (P2)

Paper link:
https://journals.sagepub.com/doi/pdf/10.1177/1071181322661400

(1) Please give a definition for AV and TAM respectively and indicate how TAM is relevant to AV.
(2) What is the purpose of this research, and how this study can benefit future studies?
(3) Please summarize the types of information collected in the survey and how the valid data was selected after the data collection.
(4) Please briefly conclude the findings in this paper, and how these findings can benefit future studies.
(5) Please indicate the limitations of this paper.

B.3 Literature Review Topic 1 (T1)

Topic No.1: Novice driver training

B.4 Literature Review Topic 2 (T2)

Topic No.2: Hazard perception in driving

Table 6: Interview quantitative statistical data.

Type	Theme	Label	Number of Mentions
Training	Training Content	Pre-use Techniques	39
		Features and Limitations of LLM	25
		Basic Methods and Operations to Use LLM	22
		Ethics and Compliance	10
		Historical or Current Tool Development	6
		Others	4
	Most Important Training	Questioning Techniques	15
		Others	7
		Background Knowledge	5
		Operating Principles	5
		Instrumental Role	3
		Access	3
		Academic Ethics	3
	Learning Path	Use the Internet	25
		Others	7
		Self-exploration	7
		Read the Official Documentation	4
		School Programs	4
		Ask Other Users	3
		Do not Know	1
Variations	Help with LLM tools	Helpful	25
		Little Helpful	13
		Very Helpful	10
	Types of LLM Assistance	Literature Summarization	32
		Information Retrieval	17
		Linguistic Optimization	14
		Data Analysis	10
		Writing Aids	7
		Framework Establishment	5
Strategies	LLM Impact on LR and PU	LLM is More Useful to LR	21
		LLM is More Useful to PU	16
		Incomparable	6
		No Answer	5
	LLM Effects on PU	Highly Effective over Long Periods	8
		Highly Effective When Time is Short	7
		Time Has No Impact	6
		No Answer	2
	LLM Effects on LR	Highly Effective over Long Periods	11
		Time Has No Impact	8
		Highly Effective When Time is Short	5
Concerns	Participants' Worries about LLM	Accuracy of Responses	35
		Privacy	24
		Copyright	23
		No Worries	6
		Content Limitations	6
		Others	5
		Academic Integrity	5
	LLM Tools Design	Do Not Provide Enough Information	33
		Already Provide Enough Information	13
		No Answer	1
		Depends on Use	1
	Adverse Effects of LLM	Accuracy of LLM-generated Responses	22
		Impact on Human Cognitive Abilities	15

	Adverse Effects of LLM	Others	7
		Copyright and Originality Concerns	5
		Time-consuming	3
		Hindering Basic Learning	3

C OUTLINE OF INTERVIEW

Table 7: Interview Outline

Type	No.	Question	Time/min
Training	1	What information/knowledge do you think a user should have before using LLM tools for academic tasks? * Classify and rank them by importance. * Why do you think XXX is necessary? * Do you think the designers have provided adequate information to the users? If no, how can the drivers acquire the necessary information about LLM tools?	10
Academic task	2	Would you find it helpful to have LLM tools to finish academic tasks? And why? * If not, under what circumstances do you think the use of LLM tools would have a negative impact? In our experiment, did your evaluation of the LLM tool change when you were doing different academic tasks?	5
Pressure	3	When you were doing our experiments with sufficient time, do you think LLM tool helped you to finish the task well? Why? When you were asked to finish the task in a shorter time, do you think the LLM tool helped you to finish the task well? Why?	5
Individual specific	4	On <task x> you did not finish it in the required time: * Do you think what are the main obstacles? * Did the LLM tool help you? Why?	5
	5	On <task x> you finished it before the required time: * Which factors do you think contributed to that? * Did the LLM tool help you? Why?	5
Ending	6	Do you have any concerns about LLM tools used for academic tasks? From the perspective of * Privacy * Copyright * Reliability of responses	5
	7	Have we missed anything?	2