BenchCouncil Transactions on Benchmarks, Standards and Evaluations 3 (2023) 100153

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.elsevier.com/locate/tbench

Full Length Article

Benchmarking ChatGPT for prototyping theories: Experimental studies using the technology acceptance model

Tiong-Thye Goh [a], Xin Dai [b], Yanwu Yang [b,*]

[a] School of Information Management, Victoria University of Wellington, Address: 23 Lambton Quay, Wellington 6011, New Zealand
[b] School of Management, Huazhong University of Science and Technology, Address: Luoyu Road, Wuhan 430074, China

ARTICLE INFO

Keywords:
ChatGPT
Large language model
Technology acceptance model
Prototyping Theory

ABSTRACT

This paper explores the paradigm of leveraging ChatGPT as a benchmark tool for theory prototyping in conceptual research. Specifically, we conducted two experimental studies using the classical technology acceptance model (TAM) to demonstrate and evaluate ChatGPT's capability of comprehending theoretical concepts, discriminating between constructs, and generating meaningful responses. Results of the two studies indicate that ChatGPT can generate responses aligned with the TAM theory and constructs. Key metrics including the factors loading, internal consistency reliability, and convergence reliability of the measurement model surpass the minimum threshold, thus confirming the validity of TAM constructs. Moreover, supported hypotheses provide an evidence for the nomological validity of TAM constructs. However, both of the two studies show a high Heterotrait–Monotrait ratio of correlations (HTMT) among TAM constructs, suggesting a concern about discriminant validity. Furthermore, high duplicated response rates were identified and potential biases regarding gender, usage experiences, perceived usefulness, and behavioural intention were revealed in ChatGPT-generated samples. Therefore, it calls for additional efforts in LLM to address performance metrics related to duplicated responses, the strength of discriminant validity, the impact of prompt design, and the generalizability of findings across contexts.

Introduction

ChatGPT (Generative Pretrained Transformer), powered by the generative large language model, possesses remarkable capabilities in generating human-like responses and engaging in naturalistic conversations across diverse topics. It is a state-of-the-art natural language processing model developed by OpenAI, trained on an extensive dataset. ChatGPT has been leveraged for a wide range of applications, ranging from aiding in creative writing and content generation to providing customer support and answering user queries [42,43].

Recent works have focused on profiling ChatGPT to explore its gender, personality, and political inclinations [34,39,49]. By prompting ChatGPT with specific instructions, researchers have investigated how these factors influence the model's responses, e.g., whether ChatGPT exhibits gender or political biases in its generated content and perception. Prompting ChatGPT with 630 political statements from voting advice applications and a political compass test, Hartmann et al. [20] uncovered ChatGPT's pro-environmental, left-libertarian ideology. In a study by Wong and Kim [49], 501 participants were recruited from Prolific to examine biases in perceiving ChatGPT's gender, where participants watched videos showcasing ChatGPT's capabilities and then provided gender ratings using an 8-point scale or a binary choice. Their results revealed a consistent tendency to perceive ChatGPT as more male than female, regardless of the response scale. These studies uncovered potential biases in AI systems, raised awareness about societal impacts of such biases, and developed methods to mitigate them [5,35]. Understanding the capabilities and limitations of ChatGPT in terms of profiling is crucial for responsibly deploying large-scale AI systems in real-world applications. By scrutinizing ChatGPT's responses, researchers can devote to developing more robust, inclusive, and unbiased AI technologies that can positively contribute to various domains and empower users with reliable and fair interactions [25].

Profiling ChatGPT assumes that ChatGPT represents a single user with stochastic attitudes and behaviours [15], thereby limiting its breadth of applications. In effect, it is possible to induce ChatGPT to simulate a population of different individual profiles, which significantly expands its capabilities. Jiang et al. [22] devised a method called chain prompting, which enables the language model to exhibit specific personalities and diverse behaviours in a controlled manner. This approach allows ChatGPT to cater to a broader range of user needs and preferences by considering different communication styles, cultural backgrounds, and domain-specific knowledge. Individual profiles enable personalized recommendations and guidance by capturing the unique characteristics of each simulated user. With ChatGPT capable of simulating a

* Corresponding author.
E-mail addresses: tiong.goh@vuw.ac.nz (T.-T. Goh), daixin@mail.hust.edu.cn (X. Dai), yangyanwu@hust.edu.cn (Y. Yang).

https://doi.org/10.1016/j.tbench.2024.100153
Received 30 October 2023; Received in revised form 25 January 2024; Accepted 31 January 2024
Available online 6 February 2024

population of individual profiles, it opens a potential avenue in conceptual understanding and theory prototyping which have not been thoroughly explored, to the best of our knowledge. By employing ChatGPT as a platform for prototyping a theory and evaluating its comprehension of related concepts, researchers can assess its ability to grasp and manipulate abstract ideas, as well as the interconnections between concepts.

The purpose of this study is twofold. First, it aims to explore the use of ChatGPT to respond to conceptual theories and assess its ability of comprehending constructs. Second, this study seeks to evaluate the validity of conceptual theories by examining relationships among different constructs with participants and survey responses generated by ChatGPT. The research questions are:

(1) How well does ChatGPT process various constructs within the context of provided conceptual theories?
(2) How valid are relationships between different constructs with survey responses generated by ChatGPT when being evaluated through a structural equation model?

This study has several potential contributions. First, it addresses a design shortcoming between engineering science prototyping and conceptual development in social science. While engineering commonly utilizes design tools for product prototyping, the realm of social science traditionally relied heavily on extensive human participation for conceptual theory development. The advent of Large Language Models (LLMs) like ChatGPT transforms this landscape by providing a design platform analogous to an engineering design tool. Software and system designers can integrate LLMs, such as ChatGPT, with other simulation tools for theoretical development. This would enable a novel and efficient approach to theory design in social science, offering a platform where human input and machine-generated insights collaborate seamlessly. The result is a more cohesive and comprehensive research methodology that amalgamates the strengths of human knowledge with the analytical capabilities of LLMs, fostering a dynamic and iterative process in social science research.

Second, researchers can follow the methodology used in this study to assess their theories' understanding, validate theoretical frameworks, and iteratively refine their theories through the analysis of responses generated by ChatGPT. This not only aids in enhancing the robustness of theoretical foundations but also provides a valuable tool for continuous improvement.

Third, the introduced research paradigm facilitates rapid exploration, fostering collaboration, and aiding in hypothesis generation. The interactive nature of ChatGPT allows for the swift identification of errors or inconsistencies, enabling researchers to promptly refine their theories. This accelerates the research prototyping process, saving valuable time and resources.

Lastly, the scalability of ChatGPT permits the testing of theories across a broad spectrum of contexts. This scalability, combined with interactive capabilities, promotes efficient theory development by allowing researchers to explore diverse scenarios and adapt their hypotheses accordingly.

In essence, the contributions of this study put forward the idea, drawing on two studies, that Large Language Models (LLMs) potentially play an important role in enhancing the efficiency, collaboration, and adaptability of research processes such as theory development in the fields of business, education, and social science using an interdisciplinary paradigm [50] which might not be possible in the past.

Related work

Large language model (LLM) - ChatGPT

ChatGPT has attracted interest in its potential for simulating human-like characteristics and generating perception responses. While past research solely focusing on using ChatGPT for prototyping theories is limited, we present studies that revealed connections that may support its relevance to theory prototyping.

Simulation of sample profile: An important aspect of prototyping theories is the ability to generate sample profiles based on a specific population. Madelyn [29] reported LLM such as GPT-3 can generate diverse data points and maintain relationships between columns, making it a useful platform for quickly generating data for testing proof of concepts. It can provide statistical relationships when explicitly requested. However, its limitations include the uncertainty of accurately modeling the complexities of real-world data.

Theory of Mind (ToM) Proficiency: Kosinski [26] demonstrated a high success rate of using ChatGPT in Theory of Mind (ToM) tasks. ChatGPT's ability to comprehend and respond to human intentions, beliefs, and emotions signifies its potential in prototyping theories related to social cognition and understanding others. Utilizing ChatGPT's ToM proficiency, researchers can explore and prototype theories on empathy, social interaction, and psychological processes. The study by Brunet-Gouet et al. [4] highlights the ability of ChatGPT to infer intentions, track beliefs, and respond to questions about mental states. These capabilities can be leveraged in the context of theory prototyping, where researchers seek to simulate and examine theoretical constructs related to human cognition, psychology, and social interaction.

ChatGPT Personalities and Psychologies: Machine personalities and psychologies associated with language models have been studied by G. Jiang et al. [22]. Their study suggests that ChatGPT's responses and interaction patterns are influenced by its machine personality characteristics. Personality priming ChatGPT into various pseudo personalities and behaviour tendencies using a psychological prompt, such as the human-like moral judgments [11], could transform ChatGPT as participants in a survey and inform the prototyping of theories related to personality psychology, human-computer interaction, and user experiences. Current research focuses on utilizing ChatGPT for various individualized assessments, leaving a research gap in assessing its capability of understanding concepts and constructs, and prototyping theories.

Technology acceptance model in education

The integration of new technologies into learning and teaching has become an area of great interest in the field of education. Digital technologies are a vital tool in achieving the objective of ensuring inclusive and equitable access for all [19]. As it is crucial to understand why users adopt or reject specific technologies in educational settings, research on technology acceptance in teaching and learning contexts has gained popularity. The Technology Acceptance Model (TAM) has gained prominence as a scientific paradigm for examining the acceptance of learning technology. TAM originated from the Theory of Reasoned Action (TRA) [2] and has evolved into a key model for understanding the predictors of human behaviours regarding technology acceptance. Davis [9] proposed the TAM framework that emphasizes factors such as perceived ease of use (PEOU), perceived usefulness (PU), attitude (AT) and behavioural intention (BI) towards using technology, which influences use motivation. It has demonstrated its applicability across a wide range of technologies and user groups.

In the realm of technology acceptance literature in education [16,40], TAM has been widely utilized by numerous studies expanding upon or applying the original model. Researchers have delved into user intentions toward e-learning technology using TAM as well as additional constructs such as subjective norms, perceived enjoyment, perceived compatibility, perceived trust, flow, and perceived social influence [14,23,24]. Moreover, the applicability of TAM has been explored in various learning technologies including mobile learning, personal learning environments (PLEs), learning management systems (LMSs), and emerging technologies like virtual reality (VR) and artificial intelligence (AI) [17]. Furthermore, the adoption of TAM in educational research highlights its significance in comprehending the factors that influence technology acceptance among students, teachers, and other

stakeholders. These studies contribute to a growing body of knowledge on the acceptance of learning technology, offering insights for effective implementation and utilization in educational contexts. Due to word limitations, interested readers can refer to Sukackė [45] and Granić and Marangunić [17] for a historical review of the technology acceptance model.

Measure development and metrics for construct validity

This study focuses on the notion of construct measurement in scale development [28]. A construct is a purposefully designed term within a scientific realm, serving to effectively organize knowledge and guide research endeavours in describing and explaining a particular aspect or phenomenon [31]. Each construct is measured with multiple items and could exist at a higher level of abstraction than concepts. In this context, a concept refers to an abstract idea operationalised through a construct that represents a particular attribute or dimension being measured. For example, in behavioural research, concepts like perceived usefulness, attitude or intention are often assessed using scales comprising multiple items or questions [9,10].

Measure development processes are essential for ensuring the validity and reliability of measures used in research and assessment. These processes involve various steps, including defining the construct's domain, generating items, specifying dimensions, and investigating dimensionality [32]. According to Peter and Churchill [32], adhering to careful measure development processes leads to higher construct validity, minimizing chance and method variance for more reliable and valid measures. ChatGPT can facilitate the measure development process by leveraging its language generation and understanding abilities to assist in generating and refining construct prototypes, thereby enhancing the overall reliability of measures.

Internal consistency reliability is a measure of the consistency or homogeneity of items within a construct [18]. It assesses the extent to which the items within the measure are measuring the same underlying construct. In behavioural research, internal consistency reliability is commonly used when a scale or questionnaire consists of multiple items or questions that are intended to measure a particular construct. The aim is to determine if these items are consistently measuring the same construct or if they are capturing different aspects. There are several commonly used statistical techniques for assessing internal consistency reliability, such as Cronbach's alpha [8] and Composite reliability [48]. These measures range from 0.0 to 1.0, with higher values indicating greater internal consistency reliability. A value closer to 1 indicates that the items in the scale are highly correlated and are consistently measuring the same construct. Generally, the accepted standard for both of these indices is 0.70 or above [18].

Convergent validity is the extent to which all indicators are related to the constructs they are meant to measure and are not related directly to constructs they are not intended to measure [7]. The metric used for evaluating a construct's convergent validity is the average variance extracted (AVE) for all indicators on each construct. AVE values above 0.5 or 0.6 are often considered indicative of good convergence, although the speific threshold can vary depending on the research field or context.

Discriminant validity, on the other hand, checks for the uniqueness of a measure and its independence from other variables. It is indicated by predictably low correlations between the measure of interest and other measures that are not supposed to measure the same variable or concept. Low correlations with unrelated measures indicate discriminant validity. Unlike reliability, discriminant validity is not enhanced by high reliability.

When assessing discriminant validity, the common Fornell-Larcker (FL) approach [12] and the relatively new Heterotrait-Monotrait ratio of correlations (HTMT) approach [21] can complement each other [13]. The FL approach relies on comparing the square roots of the average variance extracted (AVE) with the correlations between constructs. If the AVE square roots are greater than the corresponding inter-construct correlations, discriminant validity is established. On the other hand, the HTMT approach offers two ways to assess discriminant validity: as a criterion and as a statistical test [21]. Using HTMT as a criterion involves comparing its values to certain threshold values, such as 0.85 or 0.90. If the HTMT value exceeds these thresholds, it indicates potential issues with discriminant validity. The statistical test entails examining the null hypothesis (H0: HTMT $\geq$ 1) against the alternative hypothesis (H1: HTMT $<$ 1). If the confidence interval encompasses the value of one, it suggests the presence of discriminant validity concerns.

Both the FL approach and the HTMT approach provide researchers with valuable metrics to assess the distinctiveness of constructs in a measurement model. By utilizing multiple approaches, researchers can enhance the rigour and comprehensiveness of their evaluations of discriminant validity.

Nomological validity examines the degree to which a measure or construct behaves following established theoretical relationships and expectations. It involves assessing whether a measure demonstrates patterns of associations with other variables that are theoretically predicted or expected based on existing knowledge.

Measure development processes are crucial for establishing valid and reliable measures. Reliability, convergent validity, discriminant validity, and nomological validity are important metrics of measure evaluation: higher reliability generally leads to higher consistency, while convergent validity provides evidence of systematic variance; discriminant validity ensures the uniqueness of the measure, and nomological validity examines whether the measure behaves as expected. Appendix D provides a glossary of terms on theory and construct validation.

Research methodology

The methodology in this study involves a practical approach for leveraging ChatGPT in data generation, hypothesis evaluation, and metric assessment. We focus on designing prompts to guide ChatGPT in generating relevant data, supported by essential background information. The generated data undergoes a thorough evaluation of hypotheses to assess the model's fitness in addressing research inquiries. For quantifying and analysing performance metrics, we employ a combination of Partial Least Squares (PLS) and SPSS (Statistical Package for the Social Sciences). This framework ensures a robust evaluation of the generated responses, forming the basis for the study's findings and conclusions.

Prompt design

In ChatGPT, different prompt designs such as reframing can be utilized to achieve specific goals [30,51]. Instructional prompts offer explicit instructions, guiding ChatGPT's behaviours of generating responses aligned with a particular objective or style; contextual prompts provide background information, setting the conversation's context for better understanding; Socratic prompts use a questioning approach, encouraging critical thinking and exploring different perspectives [6]; seed prompts offer a starting point for ChatGPT to continue the conversation; evaluation prompts ask ChatGPT to assess given responses; creative prompts stimulate imaginative outputs like storytelling; and conditional prompts introduce specific constraints for controlled conversations. Choosing the appropriate prompt designs depends on the desired outcomes and the nature of the interaction with ChatGPT.

For the reported study, we used a mix of prompt designs to elicit responses from ChatGPT for generating participants' responses for theory prototyping, as shown in Fig. 1. The prompt design process, differs from other frameworks [27], as it necessitates GPT to act as human participants. Therefore, the prompt design framework needs to ensure that GPT comprehends the entire experiment requirements. The prompt aimed to create a structured and coherent conversation to elicit meaningful formatted responses from ChatGPT. An instructional prompt was used to provide the experiment scenario and clear instructions for answering survey questions. A combination of contextual prompts and in-

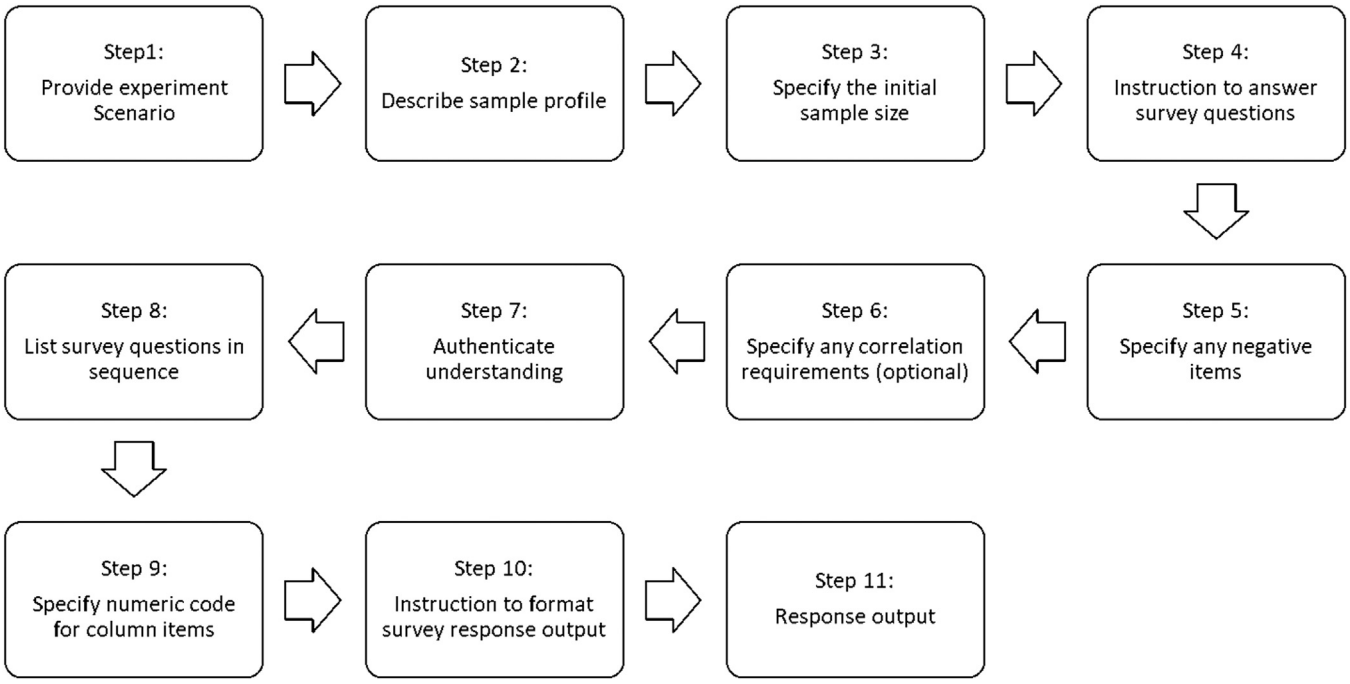

Fig. 1. The prompt design process.

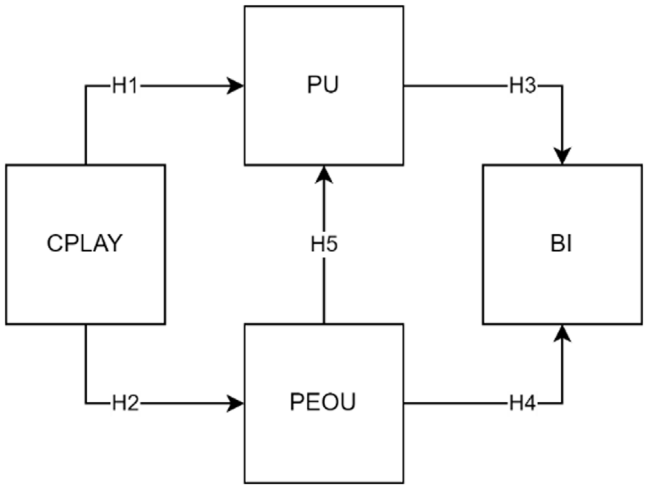

Fig. 2. Conceptual model in Study 1.

structional prompts was employed to describe the sample profile and authenticate understanding. The authentication prompt required ChatGPT to explain its reasoning before providing an answer. This process enhanced ChatGPT's reliability and overall performance by ensuring it understood the request [1]. The survey questions or conversation prompts were listed in sequence, and numeric codes were assigned to column items. Instructions were given for formatting the survey response output.

Background of study 1

Study 1 aims at creating a baseline model by leveraging the well-established and mature Technology Acceptance Model (TAM) [10] for subsequent comparison. The proposed conceptual model depicted in Fig. 2 includes the impact of computer playfulness (CPLAY) on students' perceived usefulness (PU) and perceived ease of use (PEOU) of ChatGPT, and how these factors collectively influence students' behavioural intentions (BI) towards adopting ChatGPT as a learning assistant. Computer playfulness (CPLAY) refers to the extent to which an individual's tendency to interact spontaneously, inventively, and imaginatively with computers [47]. It is an antecedent construct in the TAM3 model [46]. By incorporating the foundation Technology Acceptance Model (TAM), this study can replicate the original TAM and simultaneously validate a conceptual model over CPLAY, PU, PEOU, and BI and verify the hypotheses between these factors driving students' acceptance and adopting ChatGPT in a new learning context. The hypotheses in Study 1 were adapted from Davis [10] and Venkatesh and Bala [46] are as follows:

H1: Computer playfulness (CPLAY) positively influences students' perceived usefulness (PU) of ChatGPT as a learning assistant.
H2: Computer playfulness (CPLAY) positively influences students' perceived ease of use (PEOU) of ChatGPT as a learning assistant.
H3: Perceived usefulness (PU) positively influences students' behavioural intentions (BI) towards using ChatGPT as a learning assistant.
H4: Perceived ease of use (PEOU) positively influences students' behavioural intentions (BI) towards using ChatGPT as a learning assistant.
H5: Perceived ease of use (PEOU) positively influences students' perceived usefulness (PU) of ChatGPT as a learning assistant.

Data collection for study 1

In Study 1, the data collection involved utilizing ChatGPT to construct twenty student samples based on a student population profile. The profile included equal representation of gender with different majors, ages, years of study, and ChatGPT experience.

To gather responses, we prompted ChatGPT with the Technology Acceptance Model (TAM) questionnaires. By employing the TAM questionnaires, the intention is to gauge how ChatGPT's inner model comprehends the concept of perceived usefulness (PU), perceived ease of use (PEOU) and behavioural intention (BI) among the generated student samples. An additional construct of computer playfulness (CPLAY) [37] was included to assess its ability to discriminate against hedonic and utilitarian constructs. The scale used in this study ranged from 1

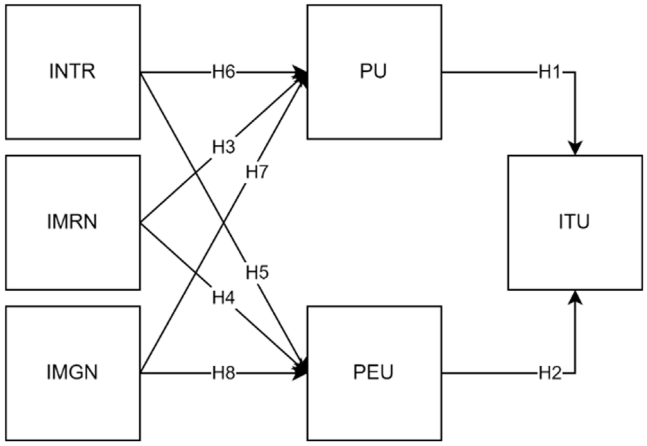

Fig. 3. Conceptual model in Study 1.

to 7, with 1 representing "Highly Unlikely," 2 representing "Unlikely," 3 representing "Somewhat Unlikely," 4 representing "Neutral," 5 representing "Somewhat Likely," 6 representing "Likely," and 7 representing "Highly Likely."

To ensure the robustness of the data collection, the process was repeated twenty times, generating a total of 400 samples. Each time ChatGPT generated twenty rows of responses based on a random set of student profiles. This methodology enabled ChatGPT to generate a sufficient sample size for conducting a structural equation analysis on how ChatGPT's inner model perceived the Technology Acceptance Model theory. The prompt is available in Appendix A. The wordings of the TAM constructs were modified from Davis [10] to align with the context of using ChatGPT as a learning assistant.

Background of study 2

The objective of Study 2 is to compare an existing TAM with different contexts and preferably to include a set of new constructs to assess ChatGPT's capabilities. To achieve this goal, we replicated the study by Barrett et al. [3]. Their study was based on user acceptance of a high-immersion virtual reality (VR) learning environment for English paragraph writing. A sample of 134 undergraduate students participated in their study, using a virtual reality system and a virtual reality learning program. A partial least squares structural equation modeling (PLS-SEM) analysis was employed to test the extended VR technology acceptance model, which was the same data analysis method used in our study. Their conceptual model as shown in Fig. 3 included exogenous variables such as Imagination (IMAG), Immersion (IMRN), and Interaction (INTR), with Perceived Ease of Use (PEU) and Perceived Usefulness (PU) mediating Intention to Use (ITU). The hypotheses in Study 2 follow exactly from Barrett et al. [3] are as follows:

H1: PU will have a strong, positive, and significant effect on learners' intention towards using the VR system.

H2: PEU will have a weak, positive, and nonsignificant effect on learners' intention towards using the high-immersion VR system.

H3: Immersion will be a strong, positive, and significant predictor for PU.

H4: Immersion will be a positive and significant predictor for PEU.

H5: Interaction will be a strong, positive, and significant determiner for PEU.

H6: Interaction will be a positive and non-significant predictor for PU.

H7: Imagination will exhibit a medium to large, positive, and significant effect on PU.

H8: Imagination will exhibit a medium, positive, and significant effect on PEU.

Table 1
Demographics summary of samples in Study 1.

Measure	Item	N (295)	%
Gender	Male	130	44.1 %
	Female	164	55.6 %
	No-binary	1	0.3 %
Age	18	7	2.4 %
	19	57	19.3 %
	20	76	25.8 %
	21	61	20.7 %
	22	63	21.4 %
	23	31	10.5 %
Year	1	74	25.1 %
	2	78	26.4 %
	3	76	25.8 %
	4	67	22.7 %
ChatGPT Experience	0	48	16.3 %
	1	74	25.1 %
	2	64	21.7 %
	3	65	22.0 %
	4	44	14.9 %

Data collection for study 2

In Study 2, the data collection method was identical to that in Study 1. It involved utilizing ChatGPT to create twenty student samples based on a defined student population profile. The profile included variables such as age, gender, majors, English ability, and AR experience.

We prompted ChatGPT with both AR and Technology Acceptance Model (TAM) questionnaires. The inclusion of the AR questionnaires aimed to assess how ChatGPT's inner model was able to comprehend the concepts of immersion (IMRM), imagination (IMGM), and interaction (INTR), which differed from the constructs of perceived usefulness (PU), perceived ease of use (PEU), and behavioural intention (BI) in the TAM questionnaire. By examining these responses, we can assess how ChatGPT's understanding of AR and its alignment with the theory of technology acceptance model.

Similarly, this study utilized a 7-point Likert scale, where a rating of 1 indicated 'strongly disagree', 2 represented 'disagree', 3 denoted 'somewhat disagree', 4 indicated 'neither agree nor disagree', 5 represented 'somewhat agree', 6 denoted 'agree', and a rating of 7 indicated 'strongly agree'.

Similar to study 1, the process was repeated twenty times, resulting in a total of 400 samples. Each iteration involved ChatGPT generating twenty rows of responses based on a randomly generated student profile. Appendix B depicts the specific prompt used in this study. The constructs and background information were adopted directly from Barrett et al. [3] to make meaningful comparisons.

Results

Study 1

Data analysis was conducted using SmartPLS4 [36] and SPSS 26 [44]. Table 1 illustrates the demographic distribution of samples in Study 1. During the sample generation process, the likelihood of ChatGPT's generating samples that deviate from the prescribed criteria was low. Note that an instance occurred in Study 1 where ChatGPT generated a sample with non-binary gender. After eliminating duplicate samples, Study 1 comprised 295 distinct data points, exhibiting an uneven gender distribution with 55.6 % females. Most of the generated participants were in the age range of 20 to 22, and the distribution of their years in university was fairly uniform. Participants' experiences with ChatGPT varied, ranging from a minimum rating of 0 (16.3 %) to a maximum rating of 4 (14.9 %), simulating a diverse range of users.

Table 2
Items loadings and constructs reliability in Study 1.

Construct	Loading	Cronbach's alpha	Composite reliability (rho_a)	Composite reliability (rho_c)	Average variance extracted (AVE)
BI1	0.982	0.960	0.965	0.98	0.961
BI2	0.979				
CPLAY1	0.947	0.959	0.96	0.97	0.891
CPLAY2	0.946				
CPLAY3	0.927				
CPLAY4	0.955				
PEOU1	0.906	0.968	0.971	0.974	0.862
PEOU2	0.941				
PEOU3	0.929				
PEOU4	0.933				
PEOU5	0.926				
PEOU6	0.937				
PU1	0.958	0.973	0.973	0.978	0.88
PU2	0.930				
PU3	0.933				
PU4	0.941				
PU5	0.918				
8PU6	0.947				

Table 3
The Heterotrait–Monotrait ratio of correlations (HTMT) in Study 1.

Construct	BI	CPLAY	PEOU
CPLAY	0.846		
PEOU	0.769	0.86	
PU	*0.93*	0.889	0.861

Table 2 illustrates the loading of items and construct reliability in Study 1. The loading and reliability values consistently exceeded the acceptable threshold of 0.7, indicating strong convergence validity [18]. Factor loading values in Study 1 were high, suggesting that the selected items effectively represented and measured the underlying constructs. Cronbach's alpha values for constructs in Study 1 showed strong internal consistency: Behavioral Intention (BI) at 0.960, Computer Playfulness (CPLAY) at 0.959, Perceived Ease of Use (PEOU) at 0.968, and Perceived Usefulness (PU) at 0.973. These scores demonstrate the reliability of the measurement scales for each construct.

In this study, we utilized the Heterotrait–Monotrait ratio of correlations (HTMT) to evaluate discriminant validity among the primary constructs. Table 3 displays the HTMT values for the relationships between Behavioral Intention (BI), Computer Playfulness (CPLAY), Perceived Ease of Use (PEOU), and Perceived Usefulness (PU). While most HTMT values hover around the threshold of 0.85, the correlation between Perceived Usefulness (PU) and Behavioral Intention (BI) produced an HTMT value of 0.93, slightly exceeding the recommended threshold. This finding prompts a potential need for further refinement of the measurement model to ensure robust discriminant validity.

Table 4 and Fig. 4 display the structural path coefficients and overall structural model from the SmartPLS analysis in Study 1. The coefficients are presented for the original sample (O), sample mean (M), and standard deviation (STDEV), along with T statistics (|O/STDEV|) and P values. The results indicated positive relationships between Computer Playfulness (CPLAY) to both Perceived Ease of Use (PEOU) and Perceived Usefulness (PU). Moreover, the path from Perceived Usefulness (PU) to Behavioral Intention (BI) was strong. However, the relationship between Perceived Ease of Use (PEOU) and Behavioral Intention (BI) was negative and lacked statistical significance. These findings offer valuable quantitative insights into the relationships among the studied constructs. When comparing the correlation values and path coefficients with the meta-analysis conducted by Yousafzai et al. [52], Study 1 exhibited correlation coefficients that fell towards the higher end of the range, while the path coefficients exceeded the upper limit.

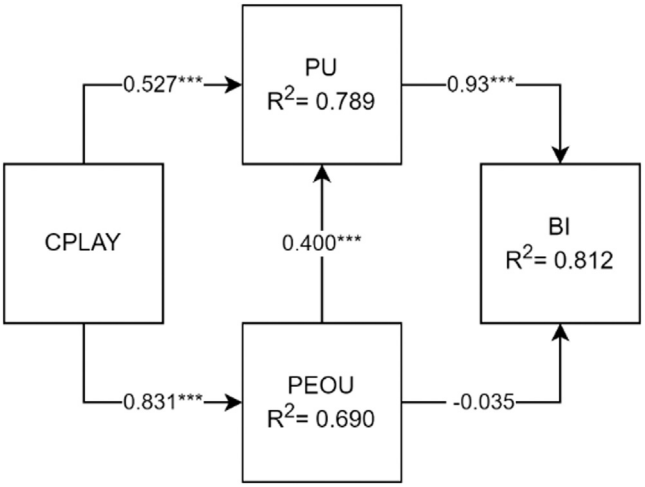

Fig. 4. Structure model in Study 1.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Study 2

In Study 2, the demographic profile of the generated sample, outlined in Table 5 after eliminating duplicates, encompasses 240 unique participants. The gender distribution is 37.5 % male and 62.5 % female, representing a ratio of 1 to 2. The age distribution displays a balanced spread, with 37.1 % at 18, 32.1 % at 19, and 30.8 % at 20. In terms of English proficiency, 6.3 % rated themselves at level 1, while 34.6 % assessed themselves at levels 2 and 3 each, and 24.6 % at level 4. Here, level 1 indicates low English ability, and level 4 signifies strong English ability. Concerning familiarity with virtual reality (VR), 13.8 % positioned themselves at level 1, 25.4 % at level 2, 33.8 % at level 3, and 27.1 % at level 4.

Table 6 presents the loading and construct reliability in Study 2, where the SmartPLS method was utilized to assess various constructs, including Imagination (IMGM), Immersion (IMRN), Interaction (INTR), Perceived Usefulness (PU), Perceived Ease of Use (PEU), and Intention to Use (ITU). The assessment involved factor loadings, Cronbach's alpha, composite reliability (rho_a), composite reliability (rho_c), and average variance extracted (AVE).

Table 4
Structure paths coefficients in Study 1.

Path	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T statistics (|O/STDEV|)	P values
CPLAY->PEOU	0.831	0.831	0.018	44.916	0.00
CPLAY -> PU	0.527	0.527	0.051	10.295	0.00
PEOU -> BI	−0.035	−0.036	0.049	0.71	0.477
PEOU -> PU	0.4	0.401	0.056	7.178	0.00
PU -> BI	0.93	0.931	0.042	21.902	0.00

Table 5
Demographics summary of samples in Study 2.

Measure	Item	N (240)	%
Gender	Male	90	37.5 %
	Female	150	62.5 %
Age	18	89	37.1 %
	19	77	32.1 %
	20	74	30.8 %
English Ability	1	15	6.3 %
	2	83	34.6 %
	3	83	34.6 %
	4	59	24.6 %
Familiar with VR	1	33	13.8 %
	2	61	25.4 %
	3	81	33.8 %
	4	65	27.1 %

In Table 6, factor loadings convey the strength of the relationship between items and their respective constructs, ranging from 0.759 to 0.962. Reliability measures, including Cronbach's alpha and composite reliability, provide insights into the internal consistency of constructs. For example, IMGM exhibits a Cronbach's alpha of 0.85, rho_a of 0.863, rho_c of 0.909, and AVE of 0.769. Similarly, ITU displays a Cronbach's alpha of 0.783, rho_a of 0.783, rho_c of 0.873, and AVE of 0.697, indicating high reliability for both constructs.

The acceptable criterion for factor loading is values above 0.7, signifying a strong relationship between items and constructs. Additionally, a Cronbach's alpha above 0.7 is deemed acceptable for reliability, ensuring internal consistency. Composite reliability values above 0.7 further indicate good reliability. All loading and reliability values were within the acceptable range, suggesting good convergence validity [18]. Loading values of TAM constructs in Study 2 appeared to be slightly lower than in Study 1.

Table 7 provides a detailed examination of the Heterotrait–Monotrait ratio of correlations (HTMT) in Study 2. The HTMT is a metric for assessing discriminant validity among constructs in the Technology Acceptance Model (TAM). However, an observation in Table 7 raised concerns, as two values among the TAM constructs exceeded 1, and one value exceeded 0.9. This suggests a potential issue with discriminant validity, indicating that certain constructs may not be sufficiently distinct from one another. A close examination revealed that the TAM

Table 6
Items loadings and constructs reliability in Study 2.

Construct	Loading	Cronbach's alpha	Composite reliability (rho_a)	Composite reliability (rho_c)	Average variance extracted (AVE)
IMGM1	0.848	0.85	0.863	0.909	0.769
	(0.831)	(0.837)			(0.754)
IMGM2	0.923				
	(0.896)				
IMGM3	0.857				
	(0.883)				
IMRN1	0.962	0.922	0.925	0.951	0.865
	(0.852)	(0.835)			(0.752)
IMRN2	0.898				
	(0.879)				
IMRN3	0.93				
	(0.871)				
INTR1	0.856	0.83	0.839	0.898	0.745
	(0.779)	(0.715)			(0.634)
INTR2	0.861				
	(0.755)				
INTR3	0.872				
	(0.851)				
ITU1	0.837	0.783	0.783	0.873	0.697
	(0.912)	(0.858)			(0.778)
ITU2	0.823				
	(0.874)				
ITU3	0.845				
	(0.860)				
PEU1	0.846	0.761	0.78	0.862	0.675
	(0.886)	(0.872)			(0.723)
PEU2	0.858				
	(0.856)				
PEU3	0.759				
	(0.847)				
PU1	0.874	0.719	0.722	0.877	0.78
	(0.911)	(0.801)			(0.716)
PU2	0.893				
	(0.873)				

Note: The bottom values were taken from Barrett et al. [3] for comparison.

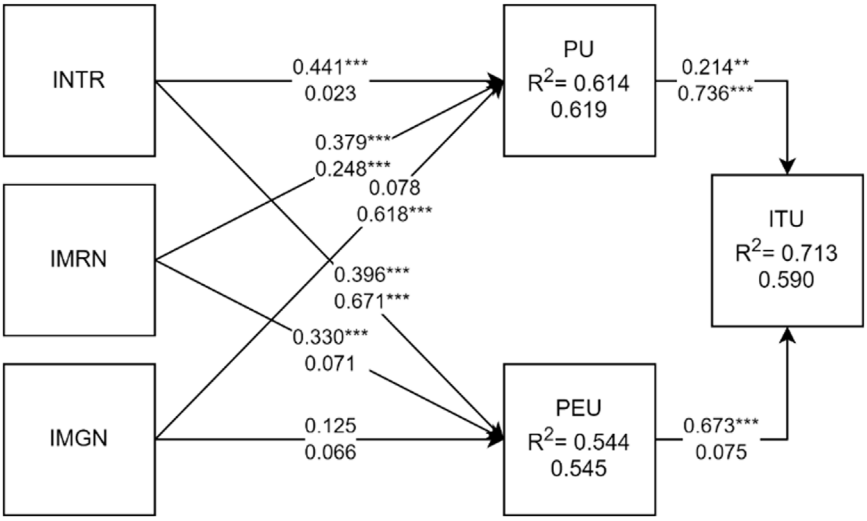

Fig. 5. Structural Model in Study 2.
Note: The bottom values were taken from Barrett et al. [3] for comparison. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7
The Heterotrait–Monotrait ratio of correlations (HTMT) in Study 2.

Construct	IMGM	IMRN	INTR	ITU	PEU
IMRN	0.841				
INTR	0.627	0.633			
ITU	0.822	0.823	0.749		
PEU	0.703	0.757	0.802	1.065	
PU	0.76	0.841	0.89	0.952	1.006

constructs' wordings in Study 1 maintained the standard phrasing from Davis [10], whereas, in Study 2, the wordings differed from the standard TAM, which could contribute to the observed variations in the HTMT ratios.

Table 8 and Fig. 5 depict the structural model and path coefficients in Study 2. Most of the path coefficients were significant except IMGM -> PEU and IMGM -> PU. R-squared values indicated a higher degree of variance explained in Study 2, suggesting a good overall fit.

For the TAM construct, Study 2's correlation coefficients were within the meta-analysis range, but the PEU->ITU path coefficient exceeded the upper range.

Discussion

This study centres on evaluating ChatGPT's processing of diverse constructs within the provided TAM theories and appraising the validity of relationships between different constructs using GPT-generated survey responses. This assessment is conducted through structural equation models in two studies. During the data gathering, we noted a tendency for ChatGPT to generate duplicated responses in both Study 1 and Study 2. The duplicate response rate was found to be 26.25 % in Study 1 and 40 % in Study 2. To ensure data quality and avoid redundancy, these duplicate responses were removed, resulting in a usable sample size of 295 for Study 1 and 240 for Study 2.

For the experiments, we selected a response length of 20 rows to overcome limitations encountered while manually working with ChatGPT. The slow response and frequent interruptions of the ChatGPT platform when reaching response limits necessitated the selection of a smaller response size. Despite these challenges, the selected response length enabled the collection of data for analysis and efficient data generation.

First, results of the two studies provided an evidence that ChatGPT can generate sample responses that align with the Technology Acceptance Model (TAM), as reflected in the loading and reliability analysis presented in Tables 2 and Table 6. Both Study 1 and Study 2 demonstrate valid models with high R-squared values of 82 % and 71 %, respectively. All the hypotheses in Study 1 were supported except H4. In Study 2, hypotheses that were not supported are H2, H6, H7 and H8.

It is interesting to note that in Study 2, the items PU3 and PEU4 contained negative wording, resulting in low loadings of 0.29 and 0.57, respectively, and they were subsequently removed from the measurement model. Further investigation is needed to understand how negative wording items impact ChatGPT's responses [38].

Second, despite the presence of high correlations among the constructs in Study 1 (see Table C3), the square roots of Average Variance Extracted (AVE) still exceeded the corresponding items in the correlation matrix, satisfying the Fornell-Larcker criterion [12]. This finding suggests that the indicators within each construct exhibit stronger internal consistency and stronger relationships with their respective constructs compared to other constructs. In Study 2, the intercorrelation coefficients among the items mostly fall within the moderate range, as shown in Table C4. Additionally, Study 2 also met the Fornell-Larcker criterion, indicating satisfactory discriminant validity. However, to fully assess the discriminant validity of the measurement model, the HTMT analysis should also be considered. The HTMT analysis revealed there was a tendency for high intercorrelation coefficients among the TAM constructs, which may impact ChatGPT's ability to discriminate between the constructs. In Study 1 the collinearity statistics (VIF) was between 1 and 4.7 while in Study 2 it was between 1.5 to 2.4. Although a VIF value of not larger than five suggests that low collinearity exists among the TAM constructs [18], additional investigation into the TAM constructs is still warranted.

Third, Study 1 exhibited higher reliability values compared to Study 2, as shown in Table 2 and Table 6. The wording style for the perceived usefulness (PU) statements in the two studies was not the same. In Study 1, the statements focused on the use of ChatGPT in a learning context, emphasizing benefits such as efficiency, performance improvement, productivity, and usefulness. These statements demonstrated a higher level of cohesion and a specific thematic focus, and the wording followed the standard TAM constructs. Conversely, the PU statements in Study 2 centred around the impact of a Virtual Reality (VR) system on understanding paragraph structure and its potential to enhance academic writing. While the statements in Study 2 were more specific, they also covered a narrower aspect compared to the broader scope in Study 1. The wordings were different from standard TAM constructs.

Fourth, when asking ChatGPT to assess the similarity between the items in the two sets of PU constructs, ChatGPT assigned a score of 8 in Study 1 and a score of 6 in Study 2, based on the construct items provided. This suggests that ChatGPT perceived greater similarity among the items in Study 1 which may contribute to the higher reliability val-

Table 8
Structure paths coefficients in Study 2.

Path	Original sample (O)	Sample mean (M)	Standard deviation (STDEV)	T statistics (|O/STDEV|)	P values
IMGM -> PEU	0.125	0.123	0.068	1.845	0.065
IMGM -> PU	0.078	0.072	0.091	0.858	0.391
IMRN -> PEU	0.33	0.335	0.073	4.543	0.000
IMRN -> PU	0.379	0.383	0.09	4.239	0.000
INTR -> PEU	0.396	0.397	0.058	6.819	0.000
INTR -> PU	0.441	0.443	0.05	8.798	0.000
PEU -> ITU	0.673	0.661	0.078	8.672	0.000
PU -> ITU	0.214	0.22	0.068	3.129	0.002

ues. In summary, the differences in wording style and thematic focus between the PU statements in Study 1 and Study 2, along with ChatGPT's scoring render further investigation.

Similarly, IMRN has a higher reliability value than INTR. We asked ChatGPT to determine the level of similarity between the items in the IMRN and INTR constructs. Here is ChatGPT's response: "Considering that all three statements in IMRN revolve around the subjective experience of immersion and engagement in the VR learning environment, I would assign a score of 8 out of 10 for the level of similarity. The statements share a common theme and address similar aspects of the VR experience.

INTR, on the other hand, consists of statements that focus on different aspects of interaction and ease of use within the VR system, such as identifying paragraph features, selecting colours, and browsing help menus. As a result, the level of similarity between the statements in INTR is lower. I would assign a score of 5 out of 10 for the level of similarity within INTR." In the context of prototyping theory, ChatGPT can assist in evaluating constructs that are not clearly defined or unambiguous [28].

Fifth, when we compared the structural models of Study 2 and the reference structural model as shown in Fig. 5, there were both similarities and differences between the structural models. The R-squared values indicate how well the models fit the data. In this case, the R-squared values were similar for both models, except for the "ITU" variable. ChatGPT had a higher R-squared value for "ITU" (0.71) compared to the reference model (0.59). This suggests that ChatGPT's model may better explain the variance in the "ITU" variable. R-squared values for PU and PEU were similar for the two models. However, path coefficients were not entirely similar. PEU→ITU was statistically significant for the simulated model but not significant in the reference model. In terms of the AR constructs, there were differences between the two models. Specifically, the relationships INTR→PU, IMRN→PEU, and IMGM→PU diverged between the models. These disparities may indicate that ChatGPT perceives the concepts related to the flow of augmented reality (AR) differently from the reference model. Another factor that may contribute to the difference was the presentation of the questionnaire. In the reference paper, all items were presented randomly, whereas, in this study, a sequential presentation of construct items was used.

Sixth, there were biases in the generated samples in terms of gender and ChatGPT experiences in Study 1. ChatGPT tended to generate more female students with higher ChatGPT experiences. As a result, the constructs' responses of PU and BI from these female students were significantly higher compared to the male samples. In Study 2, ChatGPT again generated more female students with higher English ability and AR experiences. Consequently, the constructs' responses from male students were lower compared to female students. Refer to Appendix C for the descriptive statistics in Study 1 and Study 2. These sample biases in both studies highlight the potential influence of gender and prior experiences on the constructs' responses generated by ChatGPT. It is important to consider and control for these biases when interpreting the conceptual capability of ChatGPT.

Overall, the experiments demonstrated that ChatGPT can generate relevant responses aligned with the TAM constructs which demonstrated the nomological validity of the constructs within the TAM framework based on the empirical evidence gathered from the study. The measurement models exhibited good validity, and while there were some challenges in discriminant validity due to high intercorrelations, the findings provided valuable insights into the abilities of ChatGPT to comprehend, discriminate and associate the relationship between theoretical constructs.

Practical implications

The research holds important implications for AI software developers interested in leveraging ChatGPT for theory prototyping.

First, there is an opportunity to integrate ChatGPT into software applications specifically designed for theory prototyping. By incorporating ChatGPT as a tool for generating responses aligned with theoretical constructs, developers can contribute to the development of interactive and intelligent systems for theory development.

Second, there is a need for dedicated platforms or software tools that harness the capabilities of ChatGPT for theory prototyping. Developers can design user-friendly interfaces and workflows that streamline the process, allowing researchers to create theoretical constructs, generate responses, and analyse results efficiently. Such platforms have the potential to accelerate the pace of knowledge advancement by providing researchers with a valuable tool for theory exploration and refinement.

Furthermore, the identification of biases in ChatGPT-generated samples highlights the responsibility of software developers to address and mitigate biases in their applications. Collaboration with researchers is crucial to implement strategies that enhance diversity and reduce biases in the responses generated by ChatGPT. By curating data, making algorithmic adjustments, and ensuring ongoing monitoring, developers can ensure that the outputs of the software are fair, unbiased, and reflective of diverse perspectives.

Lastly, prompt design considerations play a significant role in the outcomes of theory prototyping. Software developers, in collaboration with researchers, can establish guidelines and best practices for prompt design to improve the reliability and validity of generated responses. Clear instructions, contextual clarity, and minimization of ambiguity are essential factors that developers can focus on to enhance the quality of the data and support researchers in deriving meaningful insights.

Conclusion and future research

The study presented in this paper is not intended to be exhaustive but rather aims to open conversation for future research and underscore the potential of utilizing ChatGPT or LLMs as a benchmark tool for theory prototyping. The results indicate that ChatGPT can generate responses that align with the theoretical constructs of TAM, demonstrating its ability to process complex concepts. The high R-squared values obtained in the two experimental studies demonstrated the substantial explanatory power of the proposed models, which indicates that ChatGPT can capture and represent the underlying relationships among constructs. This paves the way for exploring and testing theories with ChatGPT as human participants in a simulated environment, which can save time and resources compared to traditional methods in behavioural research.

This research has identified several limitations that warrant further investigation. First, the presence of data duplications [41] poses a concern, potentially impacting the accurate evaluation of ChatGPT as a theory prototyping benchmark tool. The issues related to duplicated responses can be addressed through methodological refinements. Researchers can develop techniques to identify and filter out duplicated responses, ensuring data quality and integrity. This will enhance the reliability of findings and strengthen the validity of the conclusions drawn from the analysis. Second, variations in prompts such as correlation specification, negative wordings and priming used on ChatGPT may influence the generated survey responses, potentially introducing bias or inconsistencies [33] which needs further investigations. Third, it is important to note that the main theory examined in this research is the technology acceptance model specifically in the context of augmented reality (AR) and learning with ChatGPT. Further research is needed to validate the applicability and generalizability of the findings in other contexts beyond the scope of this research.

In addition to research issues triggered by these limitations, there are three interesting perspectives for future research. First, the validation of constructs, including factors loading, internal consistency reliability, and convergence reliability, provides a foundation for further investigation in this direction. In other words, researchers can build upon validated constructs to explore various theoretical frameworks and test hypotheses in a controlled and interactive manner. This has the potential to enhance the efficiency and effectiveness of theory development and refinement. Moreover, of most importance, the identification of potential biases in ChatGPT-generated samples presents an opportunity for further exploration. Understanding and mitigating these biases can contribute to the development of more robust and reliable models. Researchers can investigate feasible ways to enhance the diversity and representativeness of the generated responses, ensuring that the findings are applicable and generalizable across various demographic and contextual factors. Finally, future research can delve into more complex and nuanced theoretical models, expanding the scope of theory prototyping using LLMs.

Availability of data and material

Data is available on request from the author.

Funding

Not Applicable.

Declaration of competing interest

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

CRediT authorship contribution statement

Tiong-Thye Goh: Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Xin Dai:** Writing – original draft, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Yanwu Yang:** Writing – original draft, Project administration, Investigation, Formal analysis, Conceptualization.

Acknowledgements

The initial version of this manuscript was pre-printed on arXiv on June 4, 2023 (https://arxiv.org/abs/2307.05488).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tbench.2024.100153.

References

[1] Ajitesh, K. (2023). *ChatGPT prompts design tips & examples*. Retrieved 25-06 from https://vitalflux.com/chatgpt-prompts-design-tips-examples/.
[2] I. Ajzen, Understanding attitudes and predicting social behavior, Englewood cliffs (1980).
[3] A.J. Barrett, A. Pack, E.D. Quaid, Understanding learners' acceptance of high-immersion virtual reality systems: insights from confirmatory and exploratory PLS-SEM analyses, Comput. Educ. 169 (2021) 104214.
[4] Brunet-Gouet, E., Vidal, N., & Roux, P. (2023). *Do conversational agents have a theory of mind? a single case study of chatgpt with the hinting, false beliefs and false photographs, and strange stories paradigms*. https://hal.science/hal-03991530.
[5] A. Chan, GPT-3 and InstructGPT: technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry, AI. Ethics 3 (1) (2023) 53–64.
[6] E.Y. Chang, Prompting large language models with the socratic method, 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC), 2023.
[7] G.W. Cheung, H.D. Cooper-Thomas, R.S. Lau, L.C. Wang, Reporting reliability, convergent and discriminant validity with structural equation modeling: a review and best-practice recommendations, Asia Pacific J. Manage. (2023), doi:10.1007/s10490-023-09871-y.
[8] L.J. Cronbach, Coefficient alpha and the internal structure of tests, Psychometrika 16 (3) (1951) 297–334, doi:10.1007/BF02310555.
[9] F.D. Davis, A Technology Acceptance Model for Empirically Testing New End-User Information systems: Theory and Results, Massachusetts Institute of Technology, 1985].
[10] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS quarterly (1989) 319–340.
[11] D. Dillion, N. Tandon, Y. Gu, K. Gray, Can AI language models replace human participants? Trends Cogn. Sci. (Regul. Ed.) 27 (7) (2023) 597–600, doi:10.1016/j.tics.2023.04.008.
[12] C. Fornell, D.F. Larcker, Evaluating structural equation models with unobservable variables and measurement error, J. Market. Res. 18 (1) (1981) 39–50.
[13] G. Franke, M. Sarstedt, Heuristics versus statistics in discriminant validity testing: a comparison of four procedures, Internet Res. 29 (3) (2019) 430–447, doi:10.1108/IntR-12-2017-0515.
[14] T.T. Goh, B. Yang, The role of e-engagement and flow on the continuance with a learning management system in a blended learning environment, Int. J. Educ. Technol. High. Educ. 18 (1) (2021) 49, doi:10.1186/s41239-021-00285-8.
[15] Gozalo-Brizuela, R., & Garrido-Merchan, E.C. (2023). ChatGPT is not all you need. a state of the art review of large generative AI models. *arXiv preprint arXiv:*2301.04655.
[16] A. Granić, Educational technology adoption: a systematic review, Educ. Inf. Technol. (Dordr) 27 (7) (2022) 9725–9744.
[17] A. Granić, N. Marangunić, Technology acceptance model in educational context: a systematic literature review, Br. J. Educ. Technol. 50 (5) (2019) 2572–2593, doi:10.1111/bjet.12864.
[18] J.F. Hair Jr, G.T.M. Hult, C.M. Ringle, M. Sarstedt, N.P. Danks, S Ray, Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A workbook, Springer Nature, 2021.
[19] A. Haleem, M. Javaid, M.A. Qadri, R. Suman, Understanding the role of digital technologies in education: a review, Sustain. Oper. Comput. 3 (2022) 275–285, doi:10.1016/j.susoc.2022.05.004.
[20] Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: converging evidence on ChatGPT's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:*2301.01768.
[21] J. Henseler, C.M. Ringle, M. Sarstedt, A new criterion for assessing discriminant validity in variance-based structural equation modeling, J. Acad.Market. Sci. 43 (2015) 115–135.
[22] Jiang, G., Xu, M., Zhu, S.C., Han, W., Zhang, C., & Zhu, Y. (2022). MPI: evaluating and Inducing personality in pre-trained language models. *arXiv preprint arXiv:*2206.07550.
[23] X. Jiang, T.T. Goh, X. Chen, M. Liu, B. Yang, Investigating university students' online proctoring acceptance during COVID-19: an extension of the technology acceptance model, Aust. J. Educ. Technol. 39 (2023) 47–64, doi:10.14742/ajet.8121.
[24] X. Jiang, T.T. Goh, M. Liu, On students' willingness to use online learning: a privacy calculus theory approach [original research], Front. Psychol. 13 (2022), doi:10.3389/fpsyg.2022.880261.
[25] D. Kaur, S. Uslu, K.J. Rittichier, A. Durresi, Trustworthy artificial intelligence: a review, ACM Comput. Surv. 55 (2) (2022) Article 39, doi:10.1145/3491209.
[26] Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:*2302.02083.
[27] L.S. Lo, The CLEAR path: a framework for enhancing information literacy through prompt engineering, J. Acad. Librariansh. 49 (4) (2023) 102720, doi:10.1016/j.acalib.2023.102720.
[28] S.B. MacKenzie, P.M. Podsakoff, N.P. Podsakoff, Construct Measurement and Validation Procedures In MIS and behavioral research: integrating new and existing techniques, MIS Q. 35 (2) (2011) 293–334, doi:10.2307/23044045.
[29] Madelyn, G. (2022). *Can You generate realistic data with GPT-3? We explore fake dating with fake data*. Retrieved 20-06 from https://www.tonic.ai/blog/can-you-generate-realistic-data-with-gpt-3.
[30] Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2021). Reframing instructional prompts to GPTk's language. *arXiv preprint arXiv:*2109.07830.
[31] J.P. Peter, Construct Validity: a review of basic issues and marketing practices, J. Market. Res. 18 (2) (1981) 133–145, doi:10.1177/002224378101800201.

[32] J.P. Peter, G.A. Churchill, Relationships among research design choices and psychometric properties of rating scales: a meta-analysis, J. Market. Res. 23 (1) (1986) 1–10, doi:10.2307/3151771.
[33] Ramlochan, S. (2023). *Unlocking AI with priming: enhancing context and conversation in LLMs like ChatGPT*. Retrieved 05-06-2023 from https://www.promptengineering.org/unlocking-ai-with-priming-enhancing-context-and-conversation-in-llms-like-chatgpt/.
[34] Rao, H., Leung, C., & Miao, C. (2023). Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:*2303.01248.
[35] P.P. Ray, ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet Things Cyber-Phys. Syst. (2023).
[36] Ringle, C.M., Wende, S., & Becker, J.M. (2022). *SmartPLS 4. oststeinbek: smartPLS*. https://www.smartpls.com.
[37] F. Rondan-Cataluña, J. Arenas-Gaitán, P. Ramírez-Correa, A comparison of the different versions of popular technology acceptance models: a non-linear perspective, Kybernetes (2015) 44, doi:10.1108/K-09-2014-0184.
[38] M.J. Roszkowski, M. Soven, Shifting gears: consequences of including two negatively worded items in the middle of a positively worded questionnaire, Assess. Eval. High. Educ. 35 (1) (2010) 113–130.
[39] D. Rozado, The political biases of chatgpt, Soc. Sci. 12 (3) (2023) 148.
[40] R. Scherer, F. Siddiq, J. Tondeur, The technology acceptance model (TAM): a meta-analytic structural equation modeling approach to explaining teachers' adoption of digital technology in education, Comput. Educ. 128 (2019) 13–35, doi:10.1016/j.compedu.2018.09.009.
[41] Schwab, P.N. (2023). *ChatGPT: 1000 texts analyzed and up to 75,3% similarity*. Retrieved 05-06-2023 from https://www.intotheminds.com/blog/en/chatgpt-similarity-with-plan/.
[42] A. Shafeeg, I. Shazhaev, D. Mihaylov, A. Tularov, I. Shazhaev, Voice assistant integrated with chat GPT, Ind. J. Comput. Sci. 12 (1) (2023).
[43] M. Shidiq, The use of artificial intelligence-based chat-gpt and its challenges for the world of education; from the viewpoint of the development of creative writing skills, Proceeding of International Conference on Education, Society and Humanity, 2023.
[44] I. Statistics, IBM Corp. Released 2019. IBM SPSS Statistics for Windows, Version 26.0, IBM Corp. Google Search, Armonk, NY, 2019.
[45] V. Sukackė, Towards extending the original technology acceptance model (tam) for a better understanding of educational technology adoption. society. integration. education, in: Proceedings of the International Scientific Conference, 2019.
[46] V. Venkatesh, H. Bala, Technology acceptance model 3 and a research agenda on interventions, Decis. Sci. 39 (2) (2008) 273–315.
[47] J. Webster, J.J. Martocchio, Microcomputer playfulness: development of a measure with workplace implications, MIS Q. 16 (2) (1992) 201–226, doi:10.2307/249576.
[48] C.E. Werts, D.R. Rock, R.L. Linn, K.G. Jöreskog, A general method of estimating the reliability of a composite, Educ. Psychol. Meas. 38 (4) (1978) 933–938, doi:10.1177/001316447803800412.
[49] Wong, J., & Kim, J. (2023). ChatGPT is more likely to be perceived as male than female. *arXiv preprint arXiv:*2305.12564.
[50] Y. Yang, C. Zhang, K. Zhao, Q. Wang, The shifting role of information processing and management in interdiscipline development: from a collection of tools to a crutch? Inf. Process. Manage. 60 (4) (2023) 103388.
[51] Yaroslav, S. (2023). *The power of prompting: unleashing the full potential of ChatGPT*. Retrieved 25-06 from https://yarspirin.hashnode.dev/the-power-of-prompting-unleashing-the-full-potential-of-chatgpt.
[52] S.Y. Yousafzai, G.R. Foxall, J.G. Pallister, Technology acceptance: a meta-analysis of the TAM: part 2, J. Model. Manage. 2 (3) (2007) 281–304.