Unlocking the Black Box: Exploring the use of Generative AI (ChatGPT) in Information Systems Research

Rohani Rohan
School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
rohani.sabari@gmail.com

Lawal Ibrahim Dutsinma Faruk
School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
lawal.faruk@mail.kmutt.ac.th

Kittiphan Puapholthep
School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
kittiphan@sit.kmutt.ac.th

Debajyoti Pal
Innovative Cognitive Computing Research Center (IC2), School of Information Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
debajyoti.pal@mail.kmutt.ac.th

ABSTRACT

With the gaining popularity of generative AI tools like ChatGPT and their usage across several domains and disciplines, the question that naturally arises is how it can help the Information Systems (IS) researchers? Measuring hidden or latent constructs is one critical and primitive aspects of the IS domain that has always been challenging due to its abstractness. How good or bad these specially trained AI-based models are with respect to their conceptual understanding capabilities of specific IS constructs together with their usage for the purpose of testing IS theories is an unknown area. We set out to explore these unknown aspects in this work by conducting two separate experiments with ChatGPT using the already proven and robust Technology Acceptance Model (TAM) as the reference. Our results suggest that ChatGPT has good conceptual understanding of the presented latent constructs, although there might be certain validity issues in case of complex models. Therefore, it shows promise in the broader aspect of testing theories, but not without its limitations that we present in this research.

CCS CONCEPTS

• **Information systems**; • **Social and professional topics**; • **Applied computing**;

KEYWORDS

ChatGPT, information systems, latent constructs, scale, technology acceptance model

ACM Reference Format:
Rohani Rohan, Lawal Ibrahim Dutsinma Faruk, Kittiphan Puapholthep, and Debajyoti Pal. 2023. Unlocking the Black Box: Exploring the use of Generative AI (ChatGPT) in Information Systems Research. In *13th International Conference on Advances in Information Technology (IAIT 2023), December 06–09, 2023, Bangkok, Thailand.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3628454.3629998

IAIT 2023, December 06–09, 2023, Bangkok, Thailand

ACM ISBN 979-8-4007-0849-7/23/12…$15.00
https://doi.org/10.1145/3628454.3629998

1 INTRODUCTION

Based upon advanced AI language models, ChatGPT has gained very quick traction with regards to its capabilities towards understanding and responding in a natural language. It has been so popular, that till date it is considered to be the fastest growing innovation ever, and it has reached around 100 million active adoptions in just two months [13]. Presently, there is a lot of hype surrounding applications based on Large Language Models (LLM's) that include ChatGPT from OpenAI, Bard from Google, DALL -E 2 from OpenAI, Bing Chat from Microsoft, and several others. The ability of this technology to generate content that is almost human-like is its greatest strength and has the potential of bringing about revolutionary changes in the sociotechnical landscape. This potential of the LLMs has been made possible by the use of a special type of neural network architecture called the transformer, together with the huge datasets that have been used for training the AI models. The impact of ChatGPT may span across multiple domains ranging from healthcare, marketing, tourism, human resources, banking, retail industry, IT management to education [3].

However, there are several concerns related to ChatGPT usage too. As some of the leading researchers in the AI, IT, education, and healthcare fields have pointed out that the LLMs may produce responses that sound plausible but may not be right [13]. One of the main reasons for this is that LLMs are largely defined by the quality of the training data. If the training data has subjective bias, inaccurate, or does not have the information that is needed to answer a particular question, in that case it may hallucinate a response that might be still plausible, but not necessarily correct. Likewise, there is a high risk of the *propagation of misinformation*, which can occur when the LLMs are fed with false or misleading information during its training, leading to inaccurate or unreliable responses. Specifically, related to ChatGPT the training has been done based on vast amount of information available on the Internet (e.g., the GPT-3 model is trained from 175 billion parameters from web pages, books, research articles, and social media data), which includes both good and bad aspects of human behavior, and can result in the *propagation of inaccuracies* [11]. Therefore, there is a

need to assess how capable the LLMs might be in specific research domains and contexts.

In this work, we focus on the Information Systems (IS) research community, and specifically towards the technology adoption aspect. One major challenge that any IS researcher comes across is how to measure latent constructs, i.e., those constructs that are considered to be hidden in the mind of the users, and those which can be inferred only indirectly through mathematical models. In this regard, whenever any new concept or phenomenon needs to be developed, the aspect of scale development comes in. Previous research has shown that scale development is a complex procedure, time consuming by nature, has to be carried out in multiple stages, and needs several assessments related to various reliability and validity measures [15, 20]. With the emergence of Internet of Things (IoT), together with the rise in AI and other machine learning or natural language processing-based services, several new constructs have emerged very recently, e.g., machine personality [2], humanness [8], anthropomorphism [16], algorithmic bias [7], and several others. To make these and several other such constructs as a part of formal theory development there is a need to understand and evaluate these concepts and constructs. The bigger question is, can the LLMs like ChatGPT help in this regard? Since, these are used for a variety of NLP-based tasks, how is its potential in conceptual understanding of abstract ideas or the latent constructs that are often used for building theories is an unknown and under-explored area. However, we feel that this is an important aspect that needs further exploration to understand whether the generative AI platform in general can be used as a reliable research tool for developing scales and theories that will be of immense help for IS researchers.

Accordingly, in this work we decided to use one of the most popular acceptance models, i.e., the Technology Acceptance Model (TAM) as a reference. We feel that exploring the conceptual capabilities and understanding of ChatGPT towards the TAM constructs of perceived ease of use (PE), perceived usefulness (PU), behavioral intention (BI), and actual usage (AU) is justified, considering the popularity of TAM and its usage as one of the most widely tested models for technology acceptance. Since, its original formulation by Davis in 1986 [1], for more than the past three decades, TAM has established itself as a robust, parsimonious, and powerful model for predicting users' acceptance of technology [10, 21]. Hence, the choice of TAM seems to be justified. However, all the constructs of TAM measure the utilitarian aspects of an information system, and do not consider the hedonic aspects. Hence, we decided to use the concept of hedonic motivation (HM) that was first introduced by Venkatesh et al., in the extended version of the Unified Theory of Acceptance and Use of Technology (UTAUT), i.e., UTAUT2 [19]. The main motivation behind incorporating this hedonic aspect is to check if ChatGPT is able to differentiate between the proven theoretically distinct utilitarian and hedonic concepts. Moreover, we also evaluate the validity of the conceptual theory by exploring the relationships between the different constructs as comprehended by ChatGPT. Therefore, through this work we have attempted to answer the following two research objectives:

RQ_1: How capable ChatGPT is towards the conceptual understanding of the constructs that are part of well-established IS theories like the TAM?

RQ_2: How much relevant are the relationships between the different constructs produced by ChatGPT, i.e., its ability to test theories?

2 BACKGROUND: THE POTENTIAL OF LARGE LANGUAGE MODELS

The potential of the AI-based LLMs became more apparent to the general public after OpenAI released ChatGPT on 30 November 2022. Since, then ChatGPT has been the fastest growing consumer application ever and being used in a variety of domains. For example, the digital marketing field is witnessing its usage by creating novel advertisement formats, hyper-localized promotional offers, and personalized pages on social networking sites [13]. The generative AI-based chatbots and recommendation systems will help companies generate more revenue and understand their potential customers better. Likewise, in the healthcare domain generative AI can help patients to prepare their complete descriptions before visiting a doctor. Similarly, linguistically generative AI can suggest words, terms, or sentences for doctors and nurses to be recorded into medical systems after their interaction with patients. In the context of computer programming, authors in [17] have highlighted the potential of ChatGPT as a comprehensive debugging toolkit, and the benefits of combining its strengths with the strengths of other debugging tools to identify and fix bugs more effectively. Hence, the potential of applying ChatGPT into various types of applications is promising, however needs to be done with carefulness and caution [3].

Seeing the use of LLMs in various disciplines, it's an obvious question that how can it help the IS researchers? However, to the best of our knowledge, we could not find any research related to the use of any form of generative AI tool for solving specific IS research objectives, or one that answers how capable it might be in doing tasks related to the IS domain. However, research do exist that tries to capture the perceptions of various stakeholders towards the usage of generative AI or ChatGPT in different contexts. For example, authors in [9] have used an AI device use acceptance (AIDUA) model to explore factors on the acceptance of chatbots. Similarly, authors in [12] have explored the determinants of users' satisfaction and loyalty towards ChatGPT while also investigating the ethical concerns related to the usage of the AI-based chatbot. In an educational setting, authors in [18] have tried to identify the factors determining students' attitude towards using ChatGPT in various learning activities inside a classroom. In all these works, researchers have considered various theoretical models, e.g., TAM, Information Systems Success (ISS) model, coolness theory, affinity theory and others for explaining the respective perceptions of the users. However, these works delve into the adoption aspects of ChatGPT in various scenarios, and do not consider the potential of ChatGPT itself as to how good (bad) of a conceptual understanding it might have towards the well-established IS theories, and accordingly whether it can be used as a suitable human proxy during the data collection process. While we do agree that this might open up new ethical issues and concerns among our research community, however, that is not the focus of our present work and can be dealt as a separate issue.

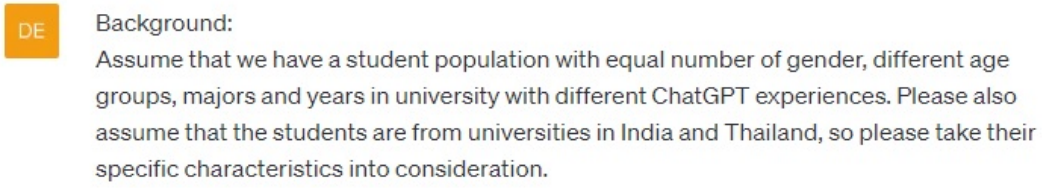

Figure 1: Background Prompt (Prompt 1)

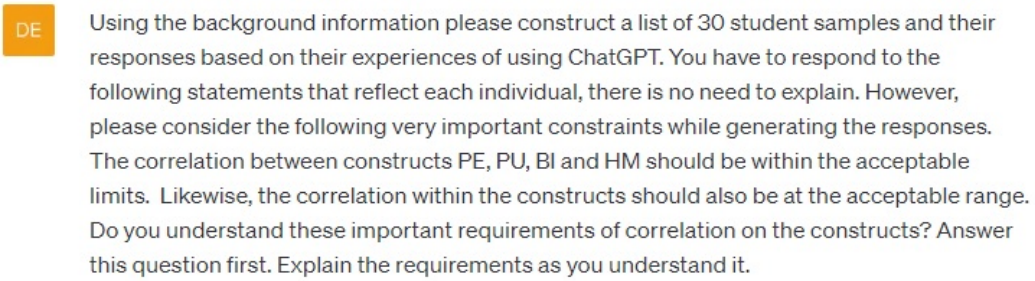

Figure 2: Sample Setup and Constraints (Prompt 2)

3 METHODOLOGY

For answering our research questions, we set-up two different interaction experiments with ChatGPT. The experiments are performed on ChatGPT version 3.5 that is freely available for public use. In experiment 1, we prompt ChatGPT to roleplay as 30 university students from India and Thailand based upon our supplied student profile. We specifically instruct ChatGPT to have equal number of genders, and belonging to different age groups, majors, years in university, and ChatGPT experiences. We supply this as a background information to ChatGPT (Figure 1). Before supplying ChatGPT with the items corresponding to PE, PU, BI, and HM constructs, we gave it with 2 constraints related to the acceptable correlation values related to "*between constructs*", and "*within constructs*". We specifically asked ChatGPT if it understands the constraints (to which it answered affirmative), and further asked for relevant explanations to examine if its understanding was correct (Figure 2).

Next, we prompted ChatGPT with the different items of PE (5 items), PU (5 items), BI (3 items), and HM (3 items) (Figure 3). These items were adapted from their original versions in [1, 19], so that their actual meaning remains coherent. In each iteration, we instructed ChatGPT to produce 30 responses based on random student profiles, and repeated it 15 times, thereby generating a total of 450 responses. Instead of instructing ChatGPT to provide with all the 450 responses in one shot, we chose to break it up into 15 rounds of iteration for avoiding lengthy specifications in our requests and asking it for less at once. This ensures that we ask ChatGPT within its limits, and do not overwhelm it by getting either *network error* or *connection timed out* error messages. Moreover, we instructed ChatGPT to produce the responses in a csv format that can be copied to Excel (Figure 3).

In experiment 2, we followed a similar approach that we did in the previous experiment for data collection, however, now focusing on testing a theory. Again, we used a total sample size of 450 students, divided into 15 rounds. In this experiment the student profile included variables like age, gender, study level, discipline, residential area, gadget used, experience of online learning, and frequency of online learning usage. We adapted this phase from our previous work in [14]. Accordingly, we prompted ChatGPT not only with the TAM related items, but also with items corresponding to technology characteristics (TC), individual characteristics (IC), and task-technology fit (TTF) constructs. Additionally, we also included some negatively worded items to check how ChatGPT comprehends those. By assessing the responses, we will be able to conclude that how ChatGPT understands the different constructs, their relationships, and whether it aligns with the theory of technology acceptance. As before we constrained the responses to satisfy the requirements of between and within constructs correlation. Additionally, to check how well ChatGPT understands our instructions, we specifically instructed it not to generate any responses if a particular sample had no experience of online learning. We did this purposely to check if ChatGPT understands filtering questions (criteria) that are often used in IS research.

4 RESULT

4.1 Measurement Model Analysis

We carried out the data analysis in SPSS version 17 and SmartPLS version 3. With regards to the demographic information generated by ChatGPT for both the experiments, Tables 1 and 2 present the relevant information for experiments 1 and 2 respectively. It should be noted that before carrying out the analysis we checked for the possibility of duplicate values.

To our surprise, in both the experiments more than 30% of the responses were duplicates. For experiment 1, we observed 201 duplicate values, whereas in experiment 2 we observed 157 duplicate ones. We removed all the duplicates from further data analysis,

On a 5-point scale, (1 - Highly unlikely, 2 - Unlikely, 3 - Neutral, 4 - Likely and 5 - Highly likely);
PE1: Learning to operate ChatGPT is easy for me
PE2: I find it easy to get ChatGPT to do what I want it to do
PE3: My interaction with ChatGPT is clear and understandable
PE4: It is easy for me to remember how to perform tasks using ChatGPT
PE5: Overall I find ChatGPT to be easy to use
PU1: Using ChatGPT will improve my study performance
PU2: Using ChatGPT in my study will increase my productivity
PU3: Using ChatGPT enhances the effectiveness of my study
PU4: Using ChatGPT makes me easier to do my job
PU5: Overall I find ChatGPT to be useful in my study
INT1: I plan to use ChatGPT in future
INT2: Assuming that I have ChatGPT, I plan to use it
INT3: I think that using ChatGPT for my study purpose will be a good idea
HM1: Using ChatGPT is fun for me
HM2: Using ChatGPT is enjoyable
HM3: Using ChatGPT is very entertaining
Please present the response in csv format. Each row should represent the response of a specific sample and the column should represent the item's number. Additionally, include columns on the left that indicate sequence, student age, gender (male:1, female:2), major, year in university (1 to 4), and ChatGPT experience (0 to 12 months). Please produce a total of 30 rows.

Figure 3: Response Generation (Prompt 3)

Table 1: Demographic Data (Experiment 1: N = 249)

Characteristics	Value	Frequency	Percentage
Gender	Male	120	48.19
	Female	129	51.81
Age (years)	18 – 20	83	33.33
	21 - 25	166	66.67
Major	Science	97	16.07
	Engineering	40	38.95
	Sociology	75	14.86
	Business	37	30.12
Chat experience (months)	0 – 3	74	29.72
	4 – 6	49	19.68
	7 – 9	55	22.09
	10 - 12	71	28.52

thereby finally giving us with 249 and 293 responses for experiments 1 and 2 respectively. Apart from the issue of very high number of duplicates, the responses generated by ChatGPT did not have much deviation from our prescribed criterion. With regards to the screening question of accepting responses from only those participants who had previous experience of online learning, ChatGPT was able to understand the requirements clearly. It only generated responses for samples having "*online learning experience = 1*", i.e., those having previous experiences, and did not generate any sample not having the relevant experience.

Next, we performed an Exploratory Factor Analysis (EFA) for examining the item' loadings onto their intended constructs, which would give some insight into the conceptual understanding capability of ChatGPT. We also checked the reliability measure by examining the Cronbach's alpha (α) values (should be greater than 0.70) [5].

Table 2: Demographic Data (Experiment 2: N = 293)

Characteristics	Value	Frequency	Percentage
Gender	Male	152	51.88
	Female	141	48.12
Age (years)	18 – 20	59	20.14
	21 – 25	111	37.88
	26 - 30	123	41.97
Study level	Undergraduate	165	56.34
	Postgraduate	128	43.66
Major	Engineering	110	37.57
	Science	103	35.18
	Business Management	80	27.26
Gadgets	Smartphone	148	50.51
	Laptop	145	49.49
Frequency of online learning	Daily	106	36.16
	Less than 3 days/week	96	32.82
	More than 3 days/week	91	31.02

Table 3: Reliability and Convergent Validity Measures (Experiment 1)

Construct	Items	Factor Loadings	Cronbach's α	CR	AVE
Perceived ease of use	PE1	0.902	0.960	0.963	0.834
	PE2	0.911			
	PE3	0.924			
	PE4	0.915			
	PE5	0.928			
Perceived usefulness	PU1	0.919	0.952	0.970	0.865
	PU2	0.927			
	PU3	0.936			
	PU4	0.915			
	PU5	0.952			
Behavioral intention	INT1	0.976	0.973	0.982	0.949
	INT2	0.968			
	INT3	0.979			
Hedonic motivation	HM1	0.947	0.945	0.964	0.900
	HM2	0.946			
	HM3	0.953			

Convergent validity is checked by examining the factor loadings (should be at least 0.5), Composite Reliability (CR) (should be at least 0.7), and the Average Variance Extracted (AVE) (should be at least 0.5) values [6]. For checking discriminant validity we performed the Heterotrait-monotrait (HTMT) analysis (values corresponding to the conceptually different constructs must be less than the threshold value of 0.85) [6]. Tables 3 and 4 present the reliability analysis and convergent validity measures for experiments 1 and 2 respectively. Tables 5 and 6 present the test for discriminant validity (HTMT analysis) for experiments 1 and 2 respectively. From the tables we observe that although there are no issues related to the reliability and convergent validity aspects, however, for experiment 2 the HTMT values of (PU and PE), and (PU and AU) are greater than 1, which is indicative of issues related to discriminant validity.

4.2 Structural Model Analysis

Before proceeding for analyzing the structural model, we checked the VIF values to look into the collinearity issues. For experiment 1, we obtained VIF values ranging between 2 to 3, whereas for experiment 2 the VIF values range from 3 to 5. Although, in both the cases the VIF values are within the recommended threshold of 5 [4, 6], yet for experiment 2 it is on the higher side. For evaluating the structural model, we used the Partial Least Squares (PLS) method. We used PLS since it is based on a prediction-oriented approach, and mainly used in exploratory studies. Although TAM is a well-established model and we do not want to re-establish its robustness or parsimoniousness, however, we do want to check the conceptual understanding of TAM by ChatGPT, and how it comprehends and explores the different TAM constructs, for which PLS is the best choice. We checked the structural relationships by using

Table 4: Reliability and Convergent Validity Measures (Experiment 2)

Construct	Items	Factor Loadings	Cronbach's α	CR	AVE
Perceived ease of use	PE1	0.841	0.804	0.906	0.706
	PE2	0.856			
	PE3	0.844			
	PE5	0.820			
Perceived usefulness	PU1	0.874	0.768	0.910	0.717
	PU2	0.893			
	PU3	0.802			
	PU4	0.815			
Technology characteristics	TC1	0.848	0.850	0.894	0.738
	TC2	0.872			
	TC3	0.857			
Individual characteristics	IC1	0.961	0.901	0.907	0.765
	IC2	0.834			
	IC3	0.822			
Task technology fit	TTF1	0.838	0.872	0.860	0.671
	TTF2	0.815			
	TTF3	0.804			
Actual usage	AU1	0.833	0.788	0.862	0.676
	AU2	0.823			
	AU3	0.810			

Table 5: Discriminant Validity – HTMT Analysis (Experiment 1)

Construct	PE	PU	HM	INT
PE	-			
PU	0.752	-		
HM	0.801	0.830	-	
INT	0.796	0.805	0.789	-

Table 6: Discriminant Validity – HTMT Analysis (Experiment 2)

Construct	PE	PU	TC	IC	TTF	AU
PE	-					
PU	1.045	-				
TC	0.784	0.758	-			
IC	0.773	0.771	0.763	-		
TTF	0.801	0.812	0.814	0.725	-	
AU	0.797	1.024	0.809	0.751	0.818	-

the bootstrapping procedure using 5000 re-samples. Figures 4 and 5 present the structural model for experiments 1 and 2 respectively.

For experiment 1, all the TAM relations are found to be valid and significant (Figure 4). The $adjusted - R^2$ value obtained is 0.788 for INT, which indicates that around 79% of the variances in behavioral intention can be explained by the model. For experiment 2 also, all the relationships between the constructs are significant (Figure 5). However, the $adjusted - R^2$ value obtained for the final dependent construct of AU is 0.647, which indicates that around 65% of the variances in actual usage can be explained by the model. The higher degree of variance explained in case of experiment 1 indicates that the overall model fit is better in case of the first experiment. Overall, comparing both the structural models, we observed that the standardized path weights, i.e., the β coefficients are generally higher for the first experiment for all the TAM related constructs. Additionally, we would like to point out that our intention in this

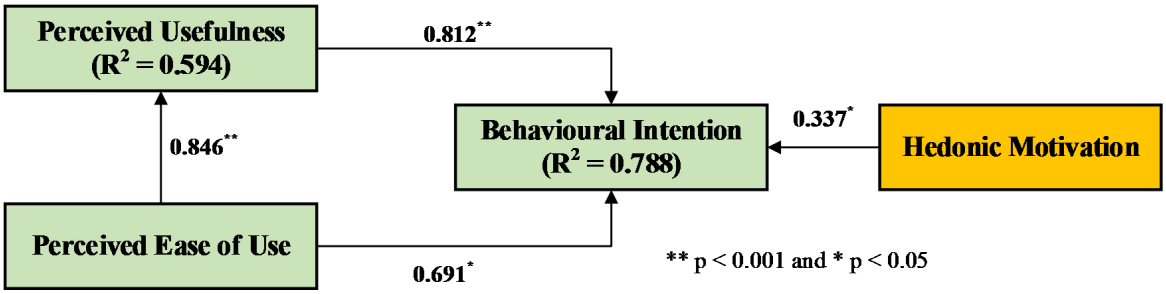

Figure 4: Structural Model (Experiment 1)

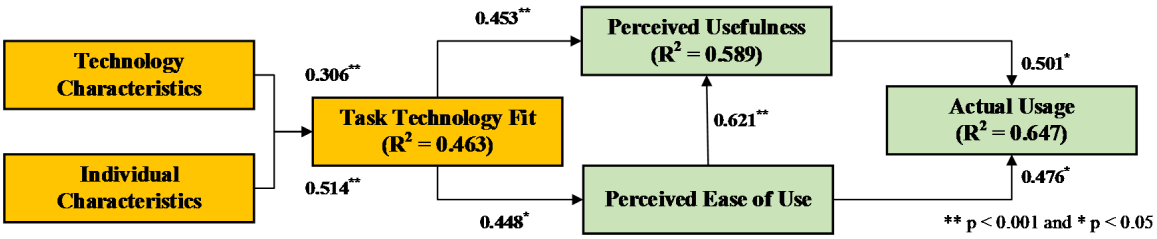

Figure 5: Structural Model (Experiment 2)

study was not to test any hypothesis, but to explore how ChatGPT comprehends the different abstract concepts, and how it oversees the relationships between the different constructs in terms of what we already know. The direct and indirect implications of the results together with the answers to both the research questions that we raised in the beginning are presented next.

5 DISCUSSION

5.1 RQ_1: How capable ChatGPT is towards the conceptual understanding of the constructs that are part of well-established IS theories like the TAM?

Based on our experimental results we can conclude that ChatGPT can generate responses that align well to the various TAM constructs. In both of our experiments ChatGPT generated responses that have a high internal consistency (reliability), as-well-as sufficient convergent validity (Tables 3 and 4). This indicates that ChatGPT is able to understand the meanings of the different constructs and it can accurately check that whether the intended meaning is getting captured or not. Additionally, in our experiment 2, intentionally we had included two negative items, one each corresponding to PE and PU respectively. We formulated PE_4 as "*I find it difficult to use the online learning system*", and PU_5 as "*Using the online learning system does not help me to fulfill my study needs*". In both the cases we obtained very low factor loadings of 0.302 and 0.295 respectively, due to which we eliminated them from data analysis. However, it becomes clear that ChatGPT can differentiate between positive and negatively worded items, but what effect such negatively worded items may have on the responses given by ChatGPT needs to be further explored.

Second, with respect to discriminant validity our observations are mixed. While for experiment 1 our HTMT correlations are within the acceptable limits, for experiment 2 some of the HTMT values lie outside the recommended range (PU and PE), and (PU and AU). Therefore, we decided to run an additional analysis to explore this disparity in greater depth. We re-ran the analysis using the Fornell Larcker criterion as it is also one standard and well accepted measure of checking discriminant validity [6]. Our observations indicate that the square-root of the AVE is greater than the correlational values between any other constructs for both the experiments, indicating that discriminant validity is satisfied for both the cases. To conclude, for experiment 2, HTMT analysis shows negative results, while Fornell Larcker criterion shows positive results. The key difference between the two methods is how discriminant validity is assessed. The Fornell Larcker focuses on comparing the AVE's and the squared correlations, whereas HTMT directly compares the HT correlations with the MT correlations, without calculating the AVE's. It means that Fornell Larcker examines the relationships between the latent constructs and their corresponding items, without considering any direct relationships between the constructs. However, HTMT considers both the within-construct and between-construct correlations. Therefore, if the measurement model is very simple, Fornell Larcker may provide a more lenient assessment of discriminant validity. For our present case, experiment 1 presents a very basic model, whereas the model complexity is higher for experiment 2. This explains as to why we have these different observations related to discriminant validity. Hence, ChatGPT is absolutely capable of generating responses that satisfy discriminant validity issues in case of simpler models, however, as model complexity increases the validity requirements may/may not be violated that needs further investigations.

Table 7: Descriptive Information of TAM Relationships (Based on [21, 22])

Path	Sample Size (N)	β Range	Correlation Status (in %)			β Values (This study)	
			Positive	Negative	Non-significant	Exp 1	Exp 2
$PU \rightarrow BI$	17,895	0.05 to 0.91	90	1	9	0.812	×
$PE \rightarrow BI$	16,518	-0.543 to 0.78	67	1	32	0.691	×
$PE \rightarrow PU$	24,110	-0.26 to 0.81	84	1	15	0.846	0.621
$PU \rightarrow AU$	14,387	-0.41 to 0.91	82	3	15	×	0.501
$PE \rightarrow AU$	11,456	-0.197 to 0.98	59	5	36	×	0.476

Third, we observed one peculiarity with our results. Across all the constructs, model 1 had better reliability and validity measures when compared with model 2 (Tables 3 to 6). Model complexity may be one of the reasons behind such an observation. Moreover, in experiment 1 the items focus on a specific context of ChatGPT usage. However, in experiment 2, the scope that the items cover is on the broader aspect of online learning systems. Hence, the measurement items are more coherent and cohesive in the former case than the later. Since, ChatGPT had generated all the responses, we felt that it should also be able to explain the differences in all these measures. Accordingly, we prompted ChatGPT to give an overall score (out of 100) for the two set of PU items and explain the reasons. ChatGPT gave a score of 80 to the PU items belonging to experiment 1, and 70 to the PU items belonging to experiment 2. The reason it gave was "*Version 1 is more specific to ChatGPT or similar tools, making it more suitable for such contexts and contributing to its higher score. Version 2, while still well-constructed, is more generic and adaptable to a broader range of online learning systems, making it suitable for a wider audience but receiving a slightly lower score due to its generality.*" Likewise, ChatGPT gave scores of 85 and 75 to the items belonging to PE for experiments 1 and 2 respectively. In addition to the above reasoning, in this case ChatGPT also mentioned that "*All items in Version 1 are positively worded, which is good for capturing the positive perception of ease of use, contributing positively to the score. In Version 2 while most items are positively worded, it includes a negatively worded item (PE4), which can help capture both positive and negative perceptions but is less favorable for response bias control, affecting its score.*"

Based on the above discussion we can conclude that ChatGPT is capable of understanding the conceptual concepts. When the model complexity is limited, it satisfies all the reliability and validity requirements. However, with an increase in model complexity certain issues may arise with regards to discriminant validity, especially with respect to HTMT analysis. These issues may be investigated in greater depth in further studies, especially in case of bigger and complex models.

5.2 RQ_2: How much relevant are the relationships between the different constructs produced by ChatGPT?

The main objective behind this second research question was to test the ability of ChatGPT for the purpose of testing theories. However, exploring the relationships between the different constructs and evaluating the structural models might vary based on the considered scenario. For example, the relationship between PE and BI maybe significant or may not be significant based on the considered context, and this is true for any relationships that a theoretical model might propose.

For the present case when we used the responses generated by ChatGPT to test the structural models (Figures 4 and 5), we found that all the hypotheses are significant and supported. Although, in terms of TAM theory this might be a positive observation, however, solely based on this we cannot conclude that ChatGPT is a suitable tool for testing theories. This points towards the need of having some baseline or reference data related to any theory, against which a particular observation may be compared to check how well it aligns towards that particular theory. Specifically, with reference to TAM we refer to the previous work by authors in [21] and [22]. Across 569 findings from 95 different TAM studies certain descriptive relationships about TAM relationships can be concluded that we present in Table 7. We just report those relationships that we have considered in this work, i.e., PU -> BI, PE -> BI, PE -> PU, PU -> AU, and PE -> AU. When we compare the results that we obtain in this work with that in Table 7, we observe that all of our path weights are positive, which is in agreement with majority of the cases. Moreover, all the values fall within the expected range, except the path from ease of use to usefulness (PE -> PU) that lies slightly outside the range (only for experiment 1). This confirms that ChatGPT is capable enough to be considered for the purpose of testing theories. Additionally, in both the experiments we obtained a decent predictive capability of 78.8% and 64.7%, which is indicative of a good model fit. However, the $adjusted - R^2$ is slightly greater in case of model 1, which might be attributed to its overall simplicity.

5.3 Additional Findings

In addition to answering the research objectives, we feel it is important to share some of our experiences and concerns with our research community. First, ChatGPT produced a lot of duplicate responses. The exact duplicate response rates were 44.7% and 34.9% for experiments 1 and 2 respectively. Over here we should emphasize that in order to overcome the limitations encountered while working with ChatGPT (in terms of slow response time, network errors or other frequent interruptions when reaching the response limit), for each of the prompts we limited the number of responses to 30. Moreover, for each subsequent prompt we instructed ChatGPT to provide with unique responses that it has not given previously.

Despite taking this precaution, we obtained such a substantial number of duplicate responses.

Second, we encountered certain scenarios where ChatGPT was not able to generate the responses in certain trials. Normally, this happened after the 7^{th} iteration, but in a total random manner. It gave with the following output when it was unable to generate the responses: "*I'm sorry for any confusion, but I can't generate additional responses for this request as it goes beyond the capacity for generating multiple sets of such data. If you have any other questions or need further assistance, please feel free to ask.*" However, upon rerunning the same prompt again the problem was solved. Another, aspect worth mentioning is since we had put forward the correlational constraints to ChatGPT when prompting it to generate responses, every time it presented us with the following message after generating the responses: "*Please note that these responses are randomly generated and may not precisely match your desired correlations. If you need more precise correlations, you may need to conduct a detailed statistical analysis or use specialized software*". Nevertheless, as we discussed previously ChatGPT has a decent capability to understand conceptual constructs and to be used for the purpose of testing theories.

6 CONCLUSION

The capabilities of generative AI tools like ChatGPT are well known and evident from their rapidly growing popularity. In this work we focused on how ChatGPT may help IS researchers by understanding the latent constructs together with its use as a tool for testing theories. Proper understanding of latent constructs is very important and often difficult since they are abstract and conceptual by nature. However, for the purpose of developing new scales for measuring novel items or exploring new theories, these latent constructs are indispensable. We find that ChatGPT has got its own set of advantages and disadvantages for this specific purpose. For simple models ChatGPT can generate accurate responses, however, as the model complexity increases there are concerns specifically related to discriminant validity issues.

This work is not without its limitations. First, we tested the conceptual capability of ChatGPT and its theory testing ability only with reference to TAM. We did this purposely, since TAM is a widely accepted robust and parsimonious model, however, in doing so we limit the research context only to this specific theory of technology acceptance. Although, this might be a starting point, future research should explore additional models and specifically those that are complex ones. Second, through this work we tried to fulfill the first step of answering whether ChatGPT is able to capture the abstract IS concepts in terms of the different latent constructs. However, we did not use ChatGPT for the purpose of generating items. Given a particular concept together with a proper working definition, it will be interesting to explore how ChatGPT generates items to measure such concepts. Future research can explore this aspect of item generation and compare how it performs against human generated items. A third area on which future work can focus on is to use ChatGPT for generating responses from diverse populations, and check whether demographics have any effect on the created responses and in terms of theory development.

REFERENCES

[1] Davis, F.D. 1986. A TECHNOLOGY ACCEPTANCE MODEL FOR EMPIRICALLY TESTING NEW END-USER INFORMATION SYSTEMS: THEORY AND RESULTS.

[2] Dutsinma, F.L.I. *et al.* 2022. Personality is to a Conversational Agent What Perfume is to a Flower. IEEE Consumer Electronics Magazine. (2022), 1–1. DOI:https://doi.org/10.1109/MCE.2022.3180183.

[3] Dwivedi, Y.K. *et al.* 2023. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management. 71, March (Aug. 2023), 102642. DOI:https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[4] Hair, J.F. *et al.* 2012. The Use of Partial Least Squares Structural Equation Modeling in Strategic Management Research: A Review of Past Practices and Recommendations for Future Applications. Long Range Planning. 45, 5 (2012), 320–340. DOI:https://doi.org/https://doi.org/10.1016/j.lrp.2012.09.008.

[5] Hair, J.F. *et al.* 2019. When to use and how to report the results of PLS-SEM. European Business Review. 31, 1 (Jan. 2019), 2–24. DOI:https://doi.org/10.1108/EBR-11-2018-0203.

[6] Joe, H. *et al.* 2017. An updated and expanded assessment of PLS-SEM in information systems research. Industrial Management & Data Systems. 117, 3 (Jan. 2017), 442–458. DOI:https://doi.org/10.1108/IMDS-04-2016-0130.

[7] Johnson, M.S. *et al.* 2022. Psychometric Methods to Evaluate Measurement and Algorithmic Bias in Automated Scoring. Journal of Educational Measurement. 59, 3 (Sep. 2022), 338–361. DOI:https://doi.org/10.1111/jedm.12335.

[8] Lu, L. *et al.* 2022. Measuring consumer-perceived humanness of online organizational agents. Computers in Human Behavior. 128, (Mar. 2022), 107092. DOI:https://doi.org/10.1016/j.chb.2021.107092.

[9] Ma, X. and Huo, Y. 2023. Are users willing to embrace ChatGPT? Exploring the factors on the acceptance of chatbots from the perspective of AIDUA framework. Technology in Society. 75, (Nov. 2023), 102362. DOI:https://doi.org/10.1016/j.techsoc.2023.102362.

[10] Marangunić, N. and Granić, A. 2015. Technology acceptance model: a literature review from 1986 to 2013. Universal Access in the Information Society. 14, 1 (Mar. 2015), 81–95. DOI:https://doi.org/10.1007/s10209-014-0348-1.

[11] Najafali, D. *et al.* 2023. Truth or Lies? The Pitfalls and Limitations of ChatGPT in Systematic Review Creation. Aesthetic Surgery Journal. 43, 8 (Jul. 2023), NP654–NP655. DOI:https://doi.org/10.1093/asj/sjad093.

[12] Niu, B. and Mvondo, G.F.N. 2024. I Am ChatGPT, the ultimate AI Chatbot! Investigating the determinants of users' loyalty and ethical usage concerns of ChatGPT. Journal of Retailing and Consumer Services. 76, (Jan. 2024), 103562. DOI:https://doi.org/10.1016/j.jretconser.2023.103562.

[13] Ooi, K. *et al.* 2023. The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. Journal of Computer Information Systems. 00, 00 (Oct. 2023), 1–32. DOI:https://doi.org/10.1080/08874417.2023.2261010.

[14] Pal, D. and Patra, S. 2021. University Students' Perception of Video-Based Learning in Times of COVID-19: A TAM/TTF Perspective. International Journal of Human–Computer Interaction. 37, 10 (Jun. 2021), 903–921. DOI:https://doi.org/10.1080/10447318.2020.1848164.

[15] Rohan, R. *et al.* 2023. A systematic literature review of cybersecurity scales assessing information security awareness. Heliyon. 9, 3 (Mar. 2023), e14234. DOI:https://doi.org/10.1016/j.heliyon.2023.e14234.

[16] Spatola, N. *et al.* 2021. Perception and Evaluation in Human–Robot Interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—A Multicomponent Approach of Anthropomorphism. International Journal of Social Robotics. 13, 7 (Nov. 2021), 1517–1539. DOI:https://doi.org/10.1007/s12369-020-00667-4.

[17] Surameery, N.M.S. and Shakor, M.Y. 2023. Use Chat GPT to Solve Programming Bugs. International Journal of Information technology and Computer Engineering. 3, 31 (Jan. 2023), 17–22. DOI:https://doi.org/10.55529/ijitc.31.17.22.

[18] Tiwari, C.K. *et al.* 2023. What drives students toward ChatGPT? An investigation of the factors influencing adoption and usage of ChatGPT. Interactive Technology and Smart Education. ahead-of-p, ahead-of-print (Aug. 2023). DOI:https://doi.org/10.1108/ITSE-04-2023-0061.

[19] Venkatesh *et al.* 2012. Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly. 36, 1 (Apr. 2012), 157. DOI:https://doi.org/10.2307/41410412.

[20] Worthington, R.L. and Whittaker, T.A. 2006. Scale Development Research. The Counseling Psychologist. 34, 6 (Nov. 2006), 806–838. DOI:https://doi.org/10.1177/0011000006288127.

[21] Yousafzai, S.Y. *et al.* 2007. Technology acceptance: a meta-analysis of the TAM: Part 1. Journal of Modelling in Management. 2, 3 (Nov. 2007), 251–280. DOI:https://doi.org/10.1108/17465660710834453.

[22] Yousafzai, S.Y. *et al.* 2007. Technology acceptance: a meta-analysis of the TAM: Part 2. Journal of Modelling in Management. 2, 3 (Nov. 2007), 281–304. DOI:https://doi.org/10.1108/17465660710834462.