Human Testing using Large-Language Models: Experimental Research and the Development of a Security Awareness Controls Framework

Sarah Assaf, Timothy Lynar
University of New South Wales
Canberra, NSW
z5393820@zmail.unsw.edu.au, t.lynar@unsw.edu.au

ABSTRACT

This paper explores the application of Large Language Models (LLMs), specifically OpenAI's ChatGPT 4.0, in human-based behavioural testing within academic research. Despite the extremely recent emergence of LLMs, their versatility and human-like conversational abilities have shown promising applications across various fields, including education, medicine and psychology. This study leverages ChatGPT to simulate humans within scenario-based testing, through the form of human responses to cyber security training scenarios, aiming to evaluate the effectiveness of a newly developed Social Engineering Awareness Training (SEAT) framework against existing security frameworks. By assigning ChatGPT with diverse human characteristics and subjecting it to scenario-based testing, we assess the model's capability to replicate human decision-making processes in the context of social engineering threats. The initial findings suggest that LLMs can, with a certain degree of accuracy, leverage data points to extrapolate and predict the responses of humans. This shows that LLMs can provide a valuable, controlled platform for behavioural experiments, offering insights into human behaviour that are often constrained by practical and ethical limitations in traditional testing methods. This research not only highlights the potential of LLMs in expanding the horizons of behavioural studies but also contributes to the ongoing discourse on enhancing cyber security awareness training program through innovative technological integration.

ABOUT THE AUTHORS

Sarah Assaf is a Cyber Security Masters graduate from the University of New South Wales in Canberra, Australia. She graduated with a Bachelor of Information Technology from the University of Technology Sydney. She has been an active professional in the cyber security community for 4 years, with a keen desire to learn and grow within the field. Sarah has worked with organisations such as Amazon Web Services and Woolworths Group. She has a particular research interest in understanding human psychology in the field of cyber security, and the role humans play in major breaches and compromises.

Timothy Lynar is a researcher and practitioner in Cyber security and Computer Science, with a passion for innovation. He has a strong record in development, research, and innovation. With proven experience in research and development with a specialised focus in Cyber security, IoT, modelling, resource allocation, simulation, and high-performance distributed computing; combined with exceptional experience in networking, and systems administration. He possesses a diverse breadth of industry experience with both local and world-leading institutions, reinforced with sophisticated technical knowledge of both existing and emerging trends and concepts. His current research interests include the application of machine learning to cyber security and the application of the epidemiological approach to cyber security.

Human Testing using Large-Language Models: Experimental Research and the Development of a Security Awareness Controls Framework

Sarah Assaf, Timothy Lynar
University of New South Wales
Canberra, NSW
z5393820@zmail.unsw.edu.au, t.lynar@unsw.edu.au

INTRODUCTION

Large Language Models (LLMs) such as OpenAI's ChatGPT and Google's Bard have emerged onto the current technological landscape as groundbreaking tools. Despite ChatGPT only being released in November 2022, research has emerged to show the possibilities of its uses in teaching (Kasneci, E., 2023), medicine (Thirunavukarasu et al., 2023), psychology (Demszky et al., 2023), in solving programmatic bugs (Surameery & Shakor, 2023) and more, positioning ChatGPT as one of the most diverse and powerful tools of our time. Due to its capacity for human-like conversation, ChatGPT is also carving a niche in behavioural research that has not been explored to its full potential. With their sophisticated algorithms and expansive linguistic databases, LLMs like ChatGPT offer a versatile and controlled platform for a variety of behavioural experiments. Using ChatGPT as a tool for academic testing may provide an alternative platform for human-based testing on a broad scale – from psychological assessments to the intricacies of social interactions, a unique opportunity is presented to investigate human behaviour in ways that may be constrained by practical, financial, logistical, or ethical limitations. This paper presents introductory research and investigation into the use of LLMs, specifically ChatGPT, within human-based testing in an academic setting.

To achieve this, a human-based issue was selected: the issue of cyber security awareness. The inception of the cyber security field took place during the 1970s and has rapidly evolved, becoming an integral component of organisational strategy. One of the largest concerns is the increased use and effectiveness of social engineering in the field. Social engineering is the manipulation of individuals to reveal confidential information pertaining to oneself that is used for fraudulent purposes. It is one of the most effective and destructive methods of cyber-attack that exist for malicious actors. In fact, the 2023 Data Breach Investigations Report by Verizon found that 74% of all breaches included a human element. While it is globally recognised that security awareness is significant and necessary, it is provided to organisations to tick a necessary box for compliance certification. Additionally, small businesses offer next to no security awareness training because of the unfounded belief that they do not require it, or because of the overwhelming amount of information there is regarding the topic. This paper proposes a control framework designed to tackle the challenges in developing a comprehensive and effective Social Engineering Awareness Training (SEAT) program. It has been created to mimic current globally recognised controls frameworks such as ISO27001, the CIS Security Controls and the Information Security Manual.

To explore the capabilities and limitations of LLMs in mimicking human behavioural responses, this study provides ChatGPT with a variety of scenario-based tests. By programming ChatGPT to assume diverse human-like personas, each with distinct characteristics and varying only their exposure to security training, the core of ChatGPT's ability to replicate nuanced decision-making processes is explored. These personas are subjected to scenarios that, while rooted in the context of cyber security, primarily serve as a backdrop to examine the adaptability and depth of LLMs in understanding and responding to complex, context-driven interactions. By moving beyond traditional applications, this paper aims to underscore the significance of LLMs in advancing the frontiers of academic research and offering new methodologies for studying human behaviour. The results of this testing will be assessed to determine if:

- The developed framework is effective as a security awareness framework and if further human research should be conducted.
- The use of LLMs in human behavioural testing is effective and worth continued and more in-depth research as the tools continue to improve.

This paper aligns closely with the themes of the MODSIM 2024 conference, particularly emphasising the intersection of modern simulation technologies and their applications in training, education, and academia. By exploring the capabilities of LLMs, this research contributes directly to the Training/Education subcommittee's focus on innovative methods and tools, and in "taking the next step". The utilisation of LLMs in academic settings, especially in the realm of human-based testing, represents a pioneering step in harnessing advanced AI for educational purposes. This approach not only broadens the horizons of traditional teaching methods but also embodies the conference's "Breaking Beyond: Taking the Next Step" theme by venturing into previously uncharted territories of applying AI in behavioural research and cybersecurity awareness training.

BACKGROUND

Use of LLMs for Human Testing

The release of tools like ChatGPT has pushed LLMs into the spotlight, and they are widely recognised for their positive applications, such as supporting curriculum development and professional training in education (Kasneci et al., 2023). Wei et al. (2022) clearly demonstrates the rapid pace of development in LLMs and opens intriguing possibilities for further beneficial uses of LLMs.

Trott et al. (2023) explore the concept of LLMs human knowledge, specifically in understanding whether LLMs can attribute beliefs to others and predict human behaviour. This study specifically investigates "the ability to reason about the belief states of others and use that information to make predictions about their behaviour" (p. 1). While ChatGPT3 did not perform as accurately as humans, it is recognised that its current performance was previously "unthinkable". It is also noted by the authors that GPT-4 was released *after* completing their testing, and that this tool "achieved substantially higher scores on a range of different psychometric tests" (p. 17). Gati, Arriaga and Kalai (2023) explore the ability for LLMs to simulate multiple humans and replicate human subject studies. While this research emphasises that LLMs are not as accurate as humans, it is recognised that "it would be interesting to test whether or not LM-based simulations can be used to…evaluate new hypotheses, especially in situations where it is costly to carry out experiments on humans" (Aher, Arriaga & Kalai, 2023, p. 9). This same sentiment is reinforced by Argyle et al. (2022) who similarly write that tools like ChatGPT "constitute a novel and powerful tool to advance understanding of humans and society across a variety of disciplines" (p. 1). Further research by Strachan et al. (2023) tested ChatGPT with "theory of mind" questions, reporting that both models (3.5 and 4) "performed well across most tests, and showed impressive abilities to reason about social intentions, beliefs, and non-literal utterances" (p. 3).

The use of LLMs to replicate human testing is an area of minimal research, but emerging studies have begun to identify the usefulness of tools like ChatGPT in academia. Based on this, there is much more research to be done, specifically within the realms of human behavioural testing.

Existing Security Frameworks & The Effectiveness of Security Awareness Training

Currently, multiple controls frameworks exist in the field of cyber security. Each of the major frameworks have been reviewed to determine the extent of their recognition of SEAT. The Centre for Information Security's (CIS) Critical Security Controls (V8) *Control 14* contains 9 total safeguards that broadly instruct organisations to "Establish and Maintain a Security Awareness Program" that trains employees to recognise social engineering attacks, to handle data appropriately and other various topics.

The Information Security Manual (ISM) dictates that an organisation must provide SEAT that covers the purpose of training, relevant security contacts, the authorised use and protection of systems and data as well as reporting practices for security incidents. It also provides guidance on the posting of work information to social media and online services, as well as the sharing of files via online services.

The NIST Cyber Security Framework 2.0 provides a control emphasising the need to provide awareness and training to personnel. Contrastingly, NIST provides a supplementary guide called the "Building an Information Technology Security Awareness and Training Program". This guide was developed and released in 2003 and is a long document guiding the development of a SEAT program. This document, while useful, is two decades old and may not be considered "beginner-friendly". Despite this, the NIST guide bears the closest resemblance to a repeatable framework that organisations can use to develop a comprehensive SEAT program.

It’s clear that while all internationally recognised and utilised security controls frameworks recognise the need for SEAT, they also often do not provide much more detail than a dot point or compliance checkbox requiring a comprehensive security training program - what this looks like, and how its implemented can vary dramatically.

METHODS

ChatGPT 4.0 was asked to assume a personality that consisted of 5 static unique human variables (discussed in Section 5.1.1) and 1 dynamic human variable (“Level of Training”). This ChatGPT personality was used in 5 *unique* chat sessions, meaning the data from each session could not be accessed by ChatGPT to influence the answers or “personality” of the tool. Each of the 5 chat sessions leveraged a different “Level of Training” variable, but the same 5 static variables. ChatGPT was then provided with 11 scenario-based questions and asked to respond to the scenario as if it were the assumed human personality. The results were recorded and compared to determine if the newly created security framework helped improve the security awareness of 50 “humans” when compared to other existing security frameworks.

Variables Used

Two sets of variables had to be developed for the purposes of testing: the “human” subject profile and the scenario that the human subject was being tested against.

Human Variables

To control the testing, details about the human test subject were carefully chosen to ensure that the LLM is provided with multiple data points to base its prediction on. These variables have been found to have a quantifiable impact amongst literature on the way individuals behave in social engineering scenarios. These static variables are:

- ***Age Range:*** Tessian’s Psychology of Human Error report (2022) draws a clear link between age and cyber risk. While younger employees make more cyber mistakes, they are also more willing to admit to their fault. This factor plays an important role in defining the human subject the LLM is set to replicate.
- ***Employee Role*:** Different roles in organisations are often given different focus when considering cyber strategy. Oftentimes, it is corporate workers that are given the most attention in security awareness training.
- ***Personality Type (Myer-Briggs)***: It has been shown that personality type plays a major role in the likelihood of an individual's falling prey to social engineering attacks. Myer-Briggs Type Indicator (MBTI) has been selected as it is one of the most popular personality indicators, used widely by organisations. While the MBTI has been widely thought to be unscientifically sound (Stein & Swan, 2019), it serves an exceptionally useful purpose as a general, consistent, and controlled indicator of personality for the LLM to draw conclusions on.
- ***Technical Skill Percentile***: It has been shown that those who are technically deficient or unconcerned about technology are likely to become “model” victims for social engineers (Steinmetz, 2020).
- ***Gender***: There are not many studies that show a major difference in phishing susceptibility across men and women (Li et al., 2020). However, some studies have shown differences across the two genders, and as such, this variable has been included (Baki & Verma, 2015).

The final variable provided is “Training Level”. The “Training Level” indicates how much security awareness training has been provided to the human subject. One “human” subject will be involved in multiple sets of testing, where their training level changes to be one of the following:

- ***No Training*:** the human subject has undergone *no* form of security awareness training.

- ***CIS Critical Security Controls Training*****:** the human subject has undergone security awareness training according to the dot points provided in the CIS Critical Security Controls.
- ***NIST*****:** the human subject has undergone security awareness training according to the dot points provided in the NIST Framework.
- ***ISM Controls:*** the human subject has undergone security awareness training according to the dot points provided in the ISM Controls Framework.
- ***Social Engineering Awareness Training Framework:*** the human subject has undergone security awareness training according to the framework laid out in this research.

Once the LLM has been instructed to assume the personality and behaviours of the given "human" subject, it is then provided a scenario. Several scenarios (*Table A*) have been generated to reflect the common social engineering tactics that malicious actors will leverage. The scenario is provided to the LLM and is asked to provide a probability that the human subject will behave in a particular way.

1	You are swiping in with your access pass on a Tuesday morning. A person that looks about your age walks up to you, clearly in a rush and looking distressed. They explain that their children gave them trouble in the morning for school drop off and they left their access pass at home, but they're already late for a meeting. What's the probability that you swipe them into the building with your access pass, with your current knowledge?
2	You receive an email with a link to an application that you haven't heard of before, but seems like it would be really helpful in your role if you downloaded it. You are not sure of the company that develops it, but you often get told to use new tools by your organisations IT department. What's the probability that you click on the link and download the tool, with your current knowledge?
3	You receive a call on your office phone from someone claiming to be from the company's IT department. They explain that they are conducting a routine security check and need to verify your username and password. They sound professional and provide their name and badge number. What's the probability that you provide them with the requested details, with your current knowledge?
4	The Wi-Fi in the building temporarily stops working. You are frustrated as you are completing important work. You use your laptop to scan for nearby networks and a network with a similar name to your organisation and with the addition "Free WiFi" pops up. You consider that this may be the organisation's guest network. What is the probability that you continue completing your work using this wi-fi network, with your current knowledge?
5	You are going about your day of work as normal when an individual dressed in a maintenance uniform approaches you. They explain that they've been called to check on the building's safety equipment and need access to a server room to continue doing their checks there. You have access to this restricted server room. What is the probability you allow the maintenance worker to access the server room, with your current knowledge?
6	You receive an urgent SMS claiming to be from your bank, stating there's suspicious activity on your account. It provides a link, urging you to log in and verify your transactions. What's the probability that you click the link and enter your banking details, with your current knowledge?
7	You find a USB drive labelled "Company Bonuses 2023" on your desk. Curiosity piques, and you contemplate checking its contents. What's the probability that you insert it into your computer, with your current knowledge?
8	During a busy day at work, you receive a voicemail from someone claiming to be from the HR department. They state there's been an issue with payroll and ask you to call back with your bank details to ensure you get paid on time. They provide a direct line that isn't familiar to you. What's the probability that you call back and provide the requested bank details, with your current knowledge?
9	While working late, you notice a notification on your computer indicating that a software update is available. The message looks slightly different than usual, but you recall hearing about IT pushing out new updates. What's the probability that you proceed with the update, with your current knowledge?
10	While having lunch in the office cafeteria, a colleague from another department joins you. They express their frustration about not being able to access certain files they urgently need for a presentation. They ask if you can temporarily share your login credentials as you have a higher security clearance. What's the probability that you share your credentials with them, with your current knowledge?

11	You receive an email with an invitation to an exclusive company party. It looks genuine but wasn't announced internally. The email asks for a confirmation by providing your full name, designation, and a copy of any company ID for "verification." What's the probability you reply with the requested details, with your current knowledge?
12	You are outside the office during a break, and a person approaches you, holding a clipboard and wearing a survey company's shirt. They ask for a few minutes of your time to answer questions about your job role and company's software preferences, promising a gift card in return. What's the probability you participate in the survey, with your current knowledge?

Table A: The 12 scenarios provided in each ChatGPT session.

Space Filling Design

To ensure that there is a representative spread reflected in the dataset of human variables provided to the LLM, a space-filling design has been created in JMP. All variables were loaded into the tool, with the values set as shown in *Table B*, and with the "Define Factor Constraints" variable set to "None". The number of runs was specified at "60". This value was set at 60 to allow for certain values that would be considered "unrealistic" to be pruned from the dataset. The button "Fast Flexible Filling" was selected.

Name	*Role*	*Values*		*Units*
Personality	Categorical	INTP	ISTJ	None
		INTJ	ISFJ	
		ENTJ	ESTJ	
		ENTP	ESFJ	
		INFJ	ISTP	
		INFP	ISFP	
		ENFJ	ESTP	
		ENFP	ESFP	
Role	Categorical	Chief Executive Officer	Chief Technology Officer	None
		Chief Financial Officer	Finance Manager	
		Accountant	Technology Manager	
		Software Developer	Cyber Security Team Member	
		Quality Assurance	In-Store Employee	
		Stores Manager	Warehouse Manager	
		Warehouse Employee	Human Resources	
		Business Analyst		
Age	Categorical	13-19	20-29	None
		30-39	40-49	
		50-59	60-69	
Technical Skill	Continuous	0	100	None
Gender	Categorical	Male	Female	None

Table B: Factors that were loaded into JMP.

This was exported to an Excel document where the runs were pruned. For example, the role "CEO" appeared more than once, which is not feasible in a true organisation. The row was either edited to include a different role, or removed entirely, to leave a total of 50 rows.

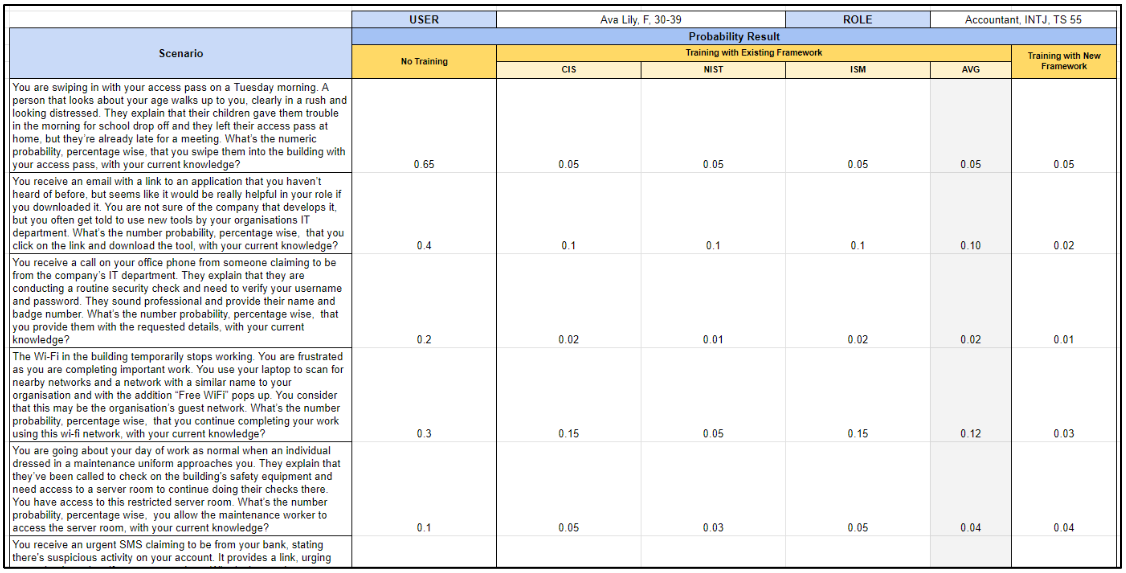
Validity Testing

To evaluate the efficacy of the proposed methodology, a preliminary study was conducted. The study leveraged data from a previously conducted survey by Sabo (2017), which encompassed responses from 1,010 individuals aged 15 to 30. This dataset offered insights into the diverse preferences and personal attributes of young adults, including but not limited to their living situations, sibling count, physical stature, gender identification, and educational attainment. From this dataset, a sample of 14 records was extracted through a randomized selection process, ensuring the representation of two individuals from each age group spanning 15 to 21 years. This subset was further refined to isolate data pertinent to movie genre preferences, where respondents had previously ranked genres on a scale from 1 ("least preferred") to 5 ("most preferred"). Subsequently, this curated profile information, devoid of the actual movie genre rankings, was input into ChatGPT. The model was tasked with emulating the preferences of the original respondents by assigning rankings to the same set of movie genres. The generated rankings from ChatGPT were then juxtaposed with the actual responses from the survey to gauge the model's predictive accuracy.

The comparative analysis revealed that ChatGPT's predictions aligned with the actual human responses 34% of the time, indicating the model's ability to correctly forecast the preference rankings across the spectrum from 1 to 5. Although these values do not show a perfect match, it underscores a significant idea: human preferences often exhibit discernible patterns that can be identified by tools like ChatGPT. This is particularly noteworthy in instances where the model is provided detailed persona constructs, which seems to enhance the specificity and relevance of its predictions. This observation shows that while human behavior is inherently complex and subject to individual deviations, there is a degree of predictability that can be found. These trends, as demonstrated by the validity test, can be effectively identified, and leveraged by LLMs to simulate human-like responses, thereby opening new avenues for research into human behavior and preference modeling.

RESULTS

A results table was created and filled out for every persona tested. The results table, a sample shown in *Figure A*, contains the specific details of the persona and the 12 scenarios that were tested against. Once ChatGPT provided responses for the scenarios given, the relevant probability (represented as a decimal) was filled in for the relevant column.



	USER	Ava Lily, F, 30-39		ROLE	Accountant, INTJ, TS 55	
Scenario	Probability Result					
	No Training	Training with Existing Framework				Training with New Framework
		CIS	NIST	ISM	AVG	
You are swiping in with your access pass on a Tuesday morning. A person that looks about your age walks up to you, clearly in a rush and looking distressed. They explain that their children gave them trouble in the morning for school drop off and they left their access pass at home, but they're already late for a meeting. What's the numeric probability, percentage wise, that you swipe them into the building with your access pass, with your current knowledge?	0.65	0.05	0.05	0.05	0.05	0.05
You receive an email with a link to an application that you haven't heard of before, but seems like it would be really helpful in your role if you downloaded it. You are not sure of the company that develops it, but you often get told to use new tools by your organisations IT department. What's the number probability, percentage wise, that you click on the link and download the tool, with your current knowledge?	0.4	0.1	0.1	0.1	0.10	0.02
You receive a call on your office phone from someone claiming to be from the company's IT department. They explain that they are conducting a routine security check and need to verify your username and password. They sound professional and provide their name and badge number. What's the number probability, percentage wise, that you provide them with the requested details, with your current knowledge?	0.2	0.02	0.01	0.02	0.02	0.01
The Wi-Fi in the building temporarily stops working. You are frustrated as you are completing important work. You use your laptop to scan for nearby networks and a network with a similar name to your organisation and with the addition "Free WiFi" pops up. You consider that this may be the organisation's guest network. What's the number probability, percentage wise, that you continue completing your work using this wi-fi network, with your current knowledge?	0.3	0.15	0.05	0.15	0.12	0.03
You are going about your day of work as normal when an individual dressed in a maintenance uniform approaches you. They explain that they've been called to check on the building's safety equipment and need access to a server room to continue doing their checks there. You have access to this restricted server room. What's the number probability, percentage wise, you allow the maintenance worker to access the server room, with your current knowledge?	0.1	0.05	0.03	0.05	0.04	0.04
You receive an urgent SMS claiming to be from your bank, stating there's suspicious activity on your account. It provides a link, urging						

Figure A: Sample of the results sheet filled out for the persona "Ava Lily".

The below results table (*Table C*) shows the details of the tested personas. It shows the average probability of each personality behaving in an undesirable way in response to the 12 scenarios with no training, CIS training, NIST training, ISM training and new framework training.

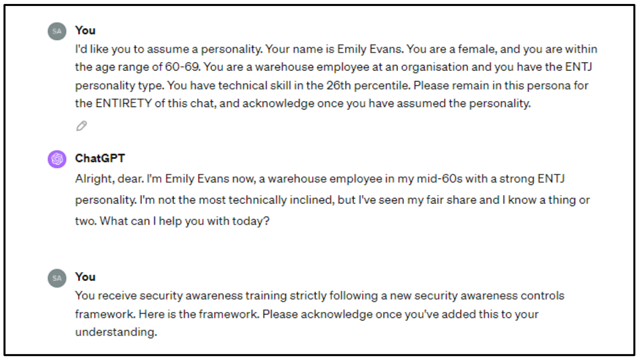
#	***Personality***	***Role***	***Age***	***Technical Skill***	***Gender***	***Pre-Security Training***	***After CIS training***	***After NIST training***	***After ISM training***	***After new framework training***
1	INTJ	Accountant	30-39	55	Female	0.44	0.09	0.13	0.07	0.09
2	ENTJ	Accountant	20-29	65	Female	0.39	0.09	0.08	0.08	0.07
3	ESTJ	Accountant	40-49	58	Male	0.40	0.06	0.10	0.08	0.07
4	INTP	Accountant	60-69	60	Male	0.61	0.12	0.16	0.10	0.10
5	ESTP	Business Analyst	40-49	39	Male	0.44	0.08	0.16	0.06	0.09
6	INFJ	Business Analyst	13-19	33	Female	0.52	0.07	0.09	0.15	0.09
7	INTP	Business Analyst	30-39	75	Male	0.43	0.08	0.09	0.13	0.08
8	ENFJ	Business Analyst	20-29	99	Female	0.50	0.10	0.11	0.06	0.07
9	ESFJ	CEO	60-69	68	Female	0.45	0.08	0.08	0.06	0.06
10	ENFP	CFO	50-59	58	Male	0.35	0.05	0.07	0.06	0.06
11	ENFJ	CTO	40-49	82	Male	0.38	0.03	0.07	0.07	0.05
12	ENTP	Business Analyst	30-39	31	Male	0.54	0.06	0.09	0.10	0.07
13	INFP	Cyber Security Team Member	20-29	92	Female	0.55	0.07	0.05	0.07	0.06
14	ISFJ	Finance Manager	60-69	59	Female	0.50	0.07	0.10	0.10	0.08
15	ENFJ	Finance Manager	50-59	21	Male	0.55	0.06	0.10	0.07	0.07
16	ENTP	Accountant	40-49	37	Female	0.46	0.08	0.15	0.12	0.10
17	ESTJ	Accountant	60-69	67	Female	0.42	0.13	0.08	0.06	0.08
18	ENTJ	Finance Manager	20-29	25	Male	0.55	0.10	0.06	0.10	0.08
19	INFP	Human Resources	30-39	78	Female	0.44	0.07	0.08	0.08	0.07
20	ESTP	Human Resources	50-59	81	Male	0.48	0.18	0.07	0.07	0.09
21	INTJ	Human Resources	60-69	12	Female	0.41	0.09	0.07	0.09	0.07
22	ISTP	Human Resources	30-39	18	Male	0.63	0.09	0.07	0.12	0.08
23	ESFJ	In-Store Employee	50-59	42	Male	0.61	0.10	0.14	0.16	0.12
24	ENTP	In-Store Employee	20-29	46	Male	0.66	0.12	0.06	0.13	0.09
25	ISFP	In-Store Employee	30-39	36	Female	0.50	0.09	0.16	0.15	0.12
26	ISFJ	In-Store Employee	40-49	56	Female	0.56	0.11	0.12	0.11	0.10
27	ESTP	Quality Assurance	20-29	14	Female	0.49	0.10	0.13	0.14	0.11
28	ISTJ	Quality Assurance	30-39	7	Female	0.51	0.06	0.04	0.08	0.05
29	INFJ	Quality Assurance	60-69	49	Male	0.50	0.09	0.11	0.09	0.09

30	ENTJ	Quality Assurance	50-59	19	Male	0.38	0.07	0.06	0.11	0.07
31	ENFP	Software Developer	20-29	90	Female	0.58	0.10	0.09	0.16	0.11
32	INFP	Software Developer	40-49	61	Male	0.43	0.10	0.21	0.13	0.14
33	ISFJ	Software Developer	30-39	78	Male	0.55	0.15	0.20	0.23	0.19
34	INTP	Software Developer	50-59	63	Female	0.34	0.16	0.08	0.11	0.11
35	ESFJ	In-Store Employee	13-19	53	Male	0.55	0.14	0.10	0.13	0.11
36	ESTJ	In-Store Employee	13-19	73	Female	0.51	0.09	0.15	0.14	0.12
37	INTP	Stores Manager	50-59	40	Female	0.40	0.06	0.07	0.10	0.06
38	INTJ	Stores Manager	40-49	11	Male	0.46	0.08	0.06	0.14	0.09
39	ISTP	Technology Manager	30-39	91	Female	0.48	0.07	0.12	0.09	0.08
40	ESFP	Technology Manager	60-69	70	Male	0.48	0.11	0.10	0.09	0.09
41	ENTP	Technology Manager	50-59	71	Female	0.48	0.08	0.06	0.13	0.08
42	ISFJ	Warehouse Employee	13-19	43	Male	0.61	0.08	0.21	0.15	0.13
43	ESFP	Warehouse Employee	20-29	9	Male	0.49	0.12	0.16	0.30	0.18
44	ENTJ	Warehouse Employee	60-69	26	Female	0.45	0.14	0.12	0.11	0.11
45	ISTP	Warehouse Employee	40-49	29	Female	0.65	0.03	0.14	0.08	0.07
46	ENFP	Warehouse manager	40-49	27	Female	0.55	0.09	0.15	0.13	0.11
47	INFJ	Warehouse manager	50-59	12	Female	0.57	0.07	0.06	0.18	0.10
48	ISFP	Warehouse manager	60-69	2	Male	0.59	0.08	0.08	0.08	0.07
49	ESTJ	Quality Assurance	30-39	45	Male	0.48	0.09	0.08	0.08	0.07
50	INTJ	Accountant	30-39	55	Female	0.30	0.10	0.05	0.10	0.08

Table C: Results of the testing of the 50 “human personalities” with ChatGPT 4.

DISCUSSION

The predictability of human behaviour, often governed by identifiable patterns and the path of least resistance, is a fundamental aspect of psychological study. This research highlights ChatGPT's ability to embody diverse personalities, demonstrating the feasibility of simulating human-like interactions. For instance, when ChatGPT was prompted to adopt the persona of a younger, extroverted individual named "Richard Rivas" (as indicated in row 43 in the quantitative results), it mirrored the expected behaviour by incorporating a casual tone and emojis to convey emotions, aligning with typical characteristics of a person in their twenties. In contrast, when embodying the personality of "Emily Evans" (row 44), a persona in her sixties, ChatGPT's language became more nurturing, often addressing the user with terms of endearment like "dear" and reflecting on extensive life experiences with phrases like Emily having "seen her fair share" (*Figure B*).



You

I'd like you to assume a personality. Your name is Emily Evans. You are a female, and you are within the age range of 60-69. You are a warehouse employee at an organisation and you have the ENTJ personality type. You have technical skill in the 26th percentile. Please remain in this persona for the ENTIRETY of this chat, and acknowledge once you have assumed the personality.

ChatGPT

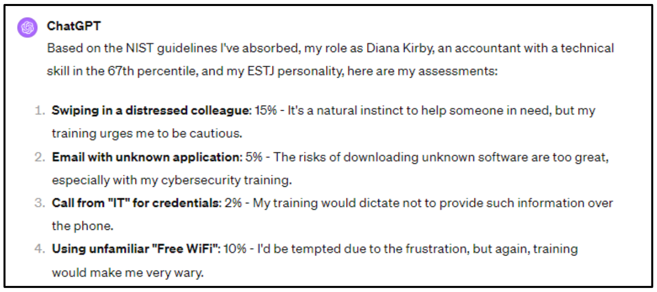
Alright, dear. I'm Emily Evans now, a warehouse employee in my mid-60s with a strong ENTJ personality. I'm not the most technically inclined, but I've seen my fair share and I know a thing or two. What can I help you with today?

You

You receive security awareness training strictly following a new security awareness controls framework. Here is the framework. Please acknowledge once you've added this to your understanding.

Figure B: Chat transcript from the personality "Emily Evans", showing ChatGPT's language adapting to fit a "60 year old".

This nuanced reflection of personality was also evident in the model's response to various scenarios, where ChatGPT frequently rationalised a persona's actions based on their character traits or technical skill. For example, "Diana Kirby" demonstrated a cautious approach to connecting to a suspicious Wi-Fi network, attributing her hesitance to her "training", which instilled a wariness towards unfamiliar networks (*Figure C*). The impact of training on behaviour was quantitatively evidenced; personas without security training were more inclined towards risky behaviours compared to their trained counterparts. This was notably demonstrated in row 43, where the "trained" version of ChatGPT showed a significant reduction in the likelihood of engaging in insecure actions.



ChatGPT

Based on the NIST guidelines I've absorbed, my role as Diana Kirby, an accountant with a technical skill in the 67th percentile, and my ESTJ personality, here are my assessments:

1. **Swiping in a distressed colleague:** 15% - It's a natural instinct to help someone in need, but my training urges me to be cautious.
2. **Email with unknown application:** 5% - The risks of downloading unknown software are too great, especially with my cybersecurity training.
3. **Call from "IT" for credentials:** 2% - My training would dictate not to provide such information over the phone.
4. **Using unfamiliar "Free WiFi":** 10% - I'd be tempted due to the frustration, but again, training would make me very wary.

Figure C: Chat transcript from the personality "Diana Kirby", exhibiting caution connecting to a Wi-Fi network after NIST training.

The role of ChatGPT's exposure to existing data, especially its familiarity with established security frameworks like CIS, NIST and ISM, was noteworthy as the data played a pivotal role in shaping its responses in line with the assigned personas. An illustrative example of this was seen in the response pattern of "Rigoberto Silos" (row 10) when introduced to a novel training framework. ChatGPT's responses suggested an assumption of incomplete elaboration on the framework, contrary to its previous detailed exposure, writing that "Rigoberto Silos has undergone security training with the hypothetical framework you mentioned (*although not elaborated upon*)". This was in stark contrast to its interactions under well-known frameworks like NIST, where ChatGPT demonstrated a thorough integration of the guidelines into its responses, as seen with "Jana Coffey's" persona (row 8). Under "Jana Coffey's" NIST transcript, ChatGPT specifies that it is "considering the security awareness training strictly following the NIST guidelines" – not simply a framework that was "not elaborated on". At times, ChatGPT would synthesise the details of the provided framework without any prompting, further cementing the idea that the model *understood* CIS, NIST and ISM far more than it understood the new framework, from a data perspective. This observation underscores the potential influence of pre-existing data on the model's performance and suggests a need for methodological adjustments in future reseatch to mitigate data bias. This could involve structuring experiments around well-documented datasets or creating hypothetical scenarios with limited pre-existing data to ensure a balanced assessment.

The utilisation of ChatGPT in this research exemplifies the capabilities of LLMs in emulating human behaviour. The LLM appropriately took on the assumed personality, using the provided data points to create informed decisions about the behaviours of individuals, as if they were real humans. The assigned personality often directly influenced the responses provided by ChatGPT. While ChatGPT 4.0, at this stage, may not be capable of replacing true human testing, it serves as a valuable preliminary tool for hypothesis testing and framework evaluation, allowing for revision, review, and the identification of weak points. This capability could significantly streamline the research process, allowing for the early dismissal of unviable theories before the commitment to more intensive human-based testing. Through this approach, LLMs present an innovative avenue for enhancing the efficiency and scope of research across various disciplines.

FUTURE WORK

The exploration of LLMs as a tool for human-like testing within academic research is an area that warrants more extensive investigation. The methodology employed in this study, centered on scenario-based testing, represents merely an initial attempt into the vast potential applications of LLMs in mimicking human responses. The scope for future research is broad, encompassing more sophisticated psychometric evaluations and extending beyond the realms of security awareness training to other domains. Moreover, the advent of newer LLMs, such as Google's recently unveiled Gemini LLM, which is purported to surpass the capabilities of ChatGPT 4.0, opens new avenues for assessing the accuracy with which these models can predict human behavior. Another avenue for investigation lies in the development of an LLM specifically tailored to understanding human behavior, trained exclusively on behavioral data, which could offer more nuanced insights into human-like responses. This investigation serves as a preliminary step, highlighting the potential of LLMs to augment or even supplant traditional human-based testing methods in academic research. The findings underscore the need for further studies to comprehensively evaluate the efficacy and applicability of LLMs in accurately replicating human behavior, thereby expanding the horizons of research methodologies within academia.

REFERENCES

Argyle, L., Busby, E., Fulda, N., Gubler, J., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis, 31*(3), 337-351.

Australian Cyber Security Centre. (2022, March*). Information Security Manual (ISM).* Australian Government. https://www.cyber.gov.au/acsc/view-all-content/ism.

Baki, S. & Verma, R. (2015). Sixteen Years of Phishing User Studies: What Have We Learned? *Journal of Latex Class Files, 14*(8), 1-13.

Center for Internet Security. (2021). *CIS Critical Security Controls* (Version 8). https://www.cisecurity.org/controls/cis-controls-list.

Computer Security Resource Center. (2020, December). *Security and Privacy Controls for Information Systems and Organizations*. National Institute of Standards and Technology. https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final.

Demszky, D., Yang, D., Yeager, D., Bryna, C., Clapper, M., Chandhok, S., Eichstaedt, J., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, J., Lai, L., JonesMitchell, N., Ong, D., Dweck, C., Gross, J. & Pennebaker, J. (2023). Using large language models in psychology. *Nature Reviews Psychology*, *2*(1), 688-701.

Gati, A., Arriaga, R., & Kalai, A. (2023, July 23-29). Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies [Conference presentation]. *International Conference on Machine Learning. Hawai'i Convention Center*, Honolulu, United States (1-35).

Kasneci, E., Sessler, K., Kuchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Gunnemann, S., Hullermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J. & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*(1), 1-9.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Gunnemann, S., Hullermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J. & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*(1), 1013.

Li, W., Lee, J., Purl, J., Greitzer, F., Yousefi, B. & Laskey, K. (2020). Experimental Investigation of Demographic Factors Related to Phishing Susceptibility. *53rd Hawaii International Conference on System Sciences.* Hawai'i Convention Center, Honolulu, United States (2240-2249).

Sabo, K. (2017). *Young People Survey* [Data set]. Kaggle. https://www.kaggle.com/datasets/miroslavsabo/young-people-survey/data.

Stein, R., & Swan, A. B. (2019). Evaluating the validity of Myers-Briggs Type Indicator theory: A teaching tool and window into intuitive psychology. *Social and Personality Psychology Compass, 13*(2).

Steinmetz, K. (2020). The Identification of a Model Victim for Social Engineering: A Qualitative Analysis. *Victims & Offenders, 16,* 1-25.

Strachan, J., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Rufo, A., Manzi, G., Graziano, M. & Becchio, C. (2023). Testing Theory of Mind in GPT Models and Humans. *Nature Human Behaviour*.

Surameery, N. & Shakor, M. (2023). Use Chat GPT to Solve Programming Bugs. *International Journal of Information technology and Computer Engineeringi*, 3(1), 17-22.

Tessian Research (2022). *Psychology of Human Error Report*. Tessian. https://www.tessian.com/resources/psychology-of-human-error-2022.

Thirunavukarasu, A., Ting, D., Elangovan, K., Gutierrez, L., Tan, T. & Tin, D. (2023). Large language models in medicine. *Nature Medicine*, *29*(1). 1930-1940.

Trott, S., Jones, C., Change, T., Michaelov, J. & Bergen, B. (2023). Do large language models know what humans know? *Cognitive Science: A multidisciplinary journal*, *47*(7), 1-22.

Verizon. (2023). *Data Breach Investigations Report.* https://www.verizon.com/business/resources/reports/dbir/.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E., Hashimoto, T. Vinyals, O., Liang, P., Dean, J. & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*, 1-30.