**Name**: Louis Bodfield

**Student** **ID**: 2296038

**Programme** **Name**: B.Sc. Artificial Intelligence and Computer Science

**Project** **Type**: 40 Credit Module

**Supervisor**: Venelin Kovatchev

**Word** **Count**:

**Abstract**

This abstract should be finished at the end

Contents

[Introduction 3](#_Toc162136255)

[1.1 Background 3](#_Toc162136256)

[1.2 Project Aims 4](#_Toc162136257)

[Literature Review 5](#_Toc162136258)

[2.1 Trust in Artificial Intelligence 5](#_Toc162136259)

[2.2 Attitudes towards Artificial intelligence 5](#_Toc162136260)

[2.3 Anthropomorphism and Chatbots 5](#_Toc162136261)

[2.4 Trust in Chatbots 6](#_Toc162136262)

[2.5 Key Findings 6](#_Toc162136263)

[2.6 Project Specification 6](#_Toc162136264)

[Design and Implementation 7](#_Toc162136265)

[3.1 Structure and Layout 8](#_Toc162136266)

[3.2 Questions 8](#_Toc162136267)

[3.3 Chatbots 10](#_Toc162136268)

[3.4 Application Features 11](#_Toc162136269)

[3.5 Architecture and Deployment 12](#_Toc162136270)

[Legal, Social, Ethical and Professional Issues 14](#_Toc162136271)

[Results 15](#_Toc162136272)

[5.1 Participant Recruitment and Demographics 15](#_Toc162136273)

[5.2 Procedure 16](#_Toc162136274)

[5.3 Hypotheses 17](#_Toc162136275)

[5.4 Findings 18](#_Toc162136276)

[Evaluation 19](#_Toc162136277)

[6.1 Discussion 19](#_Toc162136278)

[6.2 Limitations and Further Work 20](#_Toc162136279)

[6.3 Machine learning 21](#_Toc162136280)

[Conclusion 21](#_Toc162136281)

[References 22](#_Toc162136282)

[Appendices 25](#_Toc162136283)

## Introduction

This project explores how people perceive the artificial intelligence-powered conversational agents, colloquially known as chatbots, that increasingly inhabit their day-to-day life’s, with a specific focus on their preferences for different styles of chatbot.

### Background

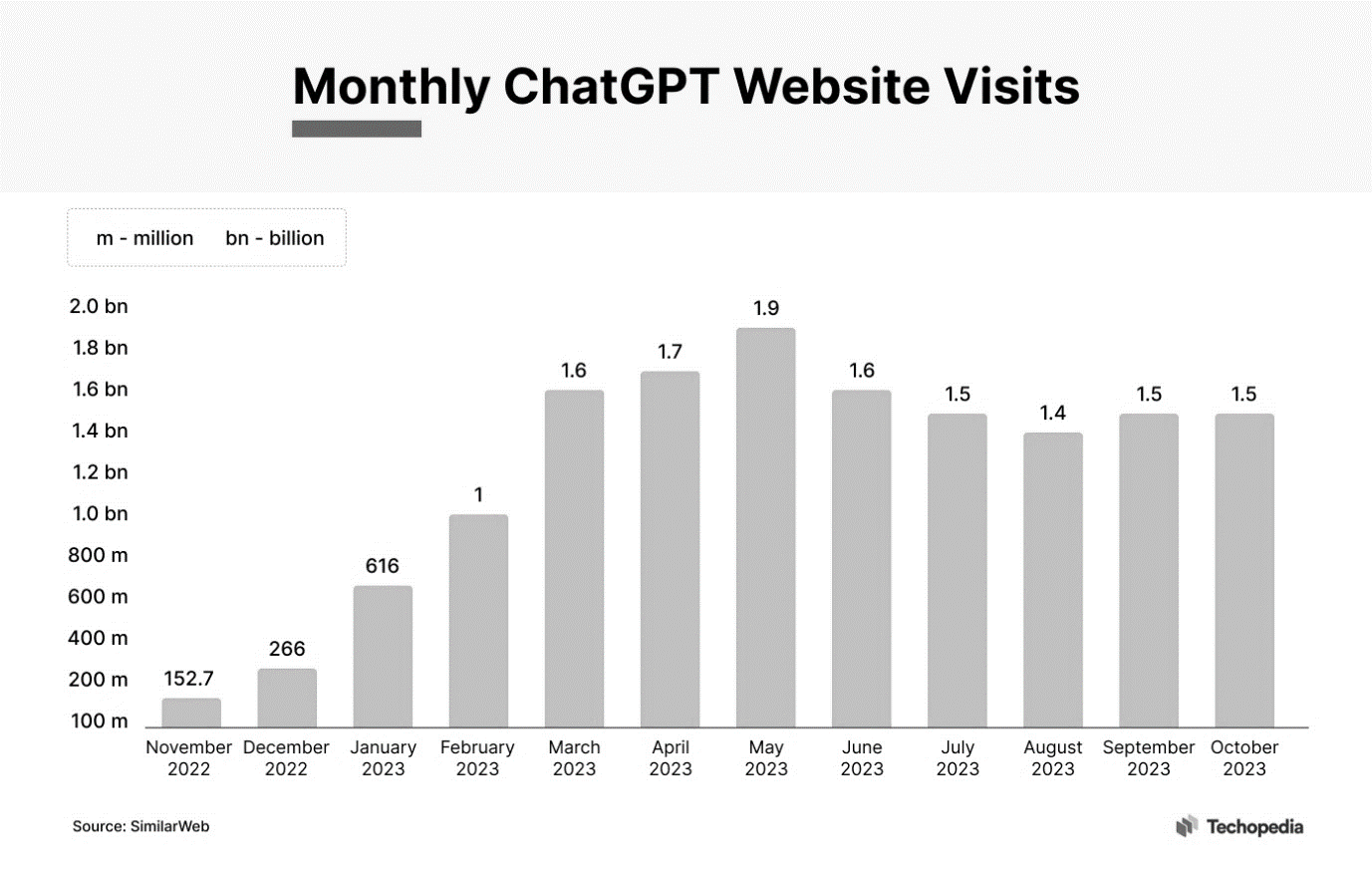
In recent years, the field of artificial intelligence (AI) has witnessed an exceptionally large surge in the development, deployment, and adoption of chatbot models, largely due to the introduction of transformer-based models in natural language processing *(Vaswani et al. 2017)*. Notably, OpenAI’s ChatGPT has seen a record-breaking launch, reaching over 1 million users in its first 5 days, and becoming the fastest growing application at the time by reaching 100 million users in its first 2 months *(Live Mint 2023).* Additionally, *OpenAI (2023)* claim that over 80% of Fortune 500 companies have adopted the chatbot application, demonstrating the extent of its practical applications. Note that while much of the discussion is centred around ChatGPT, competitors such as Google Bard (Gemini) still hold a relevant place in the discourse, offering advancements such as live internet access with no knowledge cutoff.

Figure 1‑1: Monthly ChatGPT Website Visits (Source: SimilarWeb)

These chatbot applications present both opportunities and challenges for society. On the positive side, a majority of business owners believe that ChatGPT will benefit their operations through means such as *“generating content quickly, personalizing customer experiences and streamlining job processes”* *(Forbes, 2023).* Additionally, mental health chatbots have already been adopted by NHS Talking Therapies, being given a medical device certification and streamlining the referrals process by classifying common mental health disorders services *(The British Psychological Society 2023)*. However, there are also many valid concerns with chatbots. They have been known to propagate harmful stereotypes with regards to subjects like gender, describing male characters with words like “integrity” but female characters with words like “beauty” *(Wan et al. 2023)*. This is reinforced by a tone which conveys information confidently and without question. The rise in chatbot popularity has also led to concerns with its impact on interpersonal relationships, over-reliance on the technology, and the effect it may have on job security with 63% of business leaders saying that ChatGPT will “definitely” or “probably” lead to workers being laid off *(Denecke et al. 2021; Resume Builder 2023)*.

Society has been quick to integrate chatbots into day-to-day operations, capitalizing on many of the technology’s advantages. However, there are still many drawbacks, and with our mental health, job security, and other important aspects of our lives at risk, it’s important that the domain is fully understood. *Guzman & Lewis (2019)* state the need for a *“body of work that concurrently challenges long-held assumptions of communication as a process taking place between humans, and only humans, and accounts for the expanding role of technology”*, and propose a research agenda built around the framework of human-machine communication (HMC). This agenda involves 3 aspects: how people understand and perceive AI as communicators (the types of interaction and the communicative attributes), how people understand AI in relation to themselves (and themselves in relation to AI), and the implications of blurring the boundaries between humans and machines in communication. *Følstad et al. (2021)* centre the discussion around chatbots specifically, highlighting the limited knowledge and potential challenges surrounding them. They also suggest several directions for future exploration including ‘user experience and design’, ‘democratizing chatbots’, and ‘ethics and privacy’. While there's a clear call for a deeper understanding of chatbots, there's a notable lack of research that actually explores the proposed areas, especially concerning modern chatbot technologies like ChatGPT.

### Project Aims

This project aims to align with the HMC research agenda outlined by *Guzman and Lewis (2019)* by investigating people’s preferences for the style of the chatbots they interact with, as well as the types of attributes these chatbots may possess and the roles they assume in conversations. The goal is to be able to identify key traits about people that can act as predictors for having preferences for a specific type of chatbot over another (for example, extroverts may prefer a more conversational chatbot). This may be a single significant factor that acts as a standalone predictor, or several factors that come together to form predictions. This follows *Følstad et al. (2021)* research direction of “*Design for improving chatbot user experience”* as the desire is that the results of this project will inform future chatbot creators of ways they can tailor their designs to their specific user-base.

## Literature Review

When looking at the surrounding literature, much of it is focused on artificial intelligence in general, although new chatbot-focused research in emerging. A lot of the discussions are concerned with trust in particular, and consideration was taken for highlighting any demographics or traits in people that have differing opinions on AI.

### Trust in Artificial Intelligence

People typically trust AI more than humans. *Sharan & Romano (2020)* found that when tasked with making decisions based on suggestions from either previous participants (humans) or an AI algorithm, participants tended to rely more on the recommendations generated by the AI algorithm, self-reporting that they believed such recommendations more. This was caveated by some participants commenting that they did not know how the algorithm functioned, trusting it less because of this, and suggesting that those with a better understanding of the technology are likely to accept AI more. This is reinforced by *Oksanen et al. (2020)* who found that individuals with a degree in technology or engineeringhad higher trust towards robots and AI. Their experiment where participants play a trust game against an opponent described as an AI or a human found that describing opponents as robots or AI did not impact participants’ trust toward them, however they did find that opponents with a robot sounding name like “jdrx894” were trusted more than a human name like “Michael”, and this trust was more significant in individuals with technology degrees. Additionally, personality traits such as openness to experience positively correlated with trust, while conscientiousness showed a negative correlation.

### Attitudes towards Artificial intelligence

Several other traits in people have been linked to predicting attitudes towards artificial intelligence in general. A study *(Schepman & Rodway 2021)* aiming to link participants general attitudes towards AI to several psychological factors found that introverts had more positive attitudes towards AI, *“likely because of algorithm appreciation”*, and also that participants that scored high in conscientiousness and agreeableness were more forgiving towards the negative aspects of AI. Additionally, having higher levels of corporate distrust was linked to having to more negative views on AI. *Pinto dos Santos et al. (2018)* investigated medical students in particular and found that while in general they do not worry about AI and its implications, male and more tech-savvy students were notably more confident and less fearful of the technology.

### Anthropomorphism and Chatbots

Anthropomorphism is when human-like traits are exhibited in non-human entities. In the context of chatbots, this includes giving the chatbot a name, using pronouns like “I” and “me”, using emotive language and punctuation, and mimicking human-like conversation styles. *Li et al. (2023)* found in their study (regarding chatbots and customer service) that chatbots that show warmth and competence significantly increase users trust in them, and have a positive effect on customer purchases. Another study that looked at chatbots in customer service *(Adam et al. 2020)* had similar results, finding that *“both anthropomorphism as well as the need to stay consistent significantly increase the likelihood that users comply with a chatbot’s request for service feedback”*. Here, anthropomorphism took the form of identity (a name and first-person pronouns), small-talk (greetings and farewells), and empathy (ability to notice and react to a user’s emotional expressions), also known as ADCs.

### Trust in Chatbots

When looking at factors that influence trust and adoption of chatbots, *Dekkal et al. (2023)* identified chatbots being both practical and enjoyable to use as factors that increase trust, with personalization only having a minor effect. Factors that decrease trust include creepiness and (only marginally) privacy. One key finding from this was that trusting a chatbot doesn’t necessarily translate into adopting it for those with high levels of technological anxiety and therefor moderates all of the other relationships. *Ltifi (2023)* found (similarly to previous studies) empathy and friendliness in chatbots to be a significant predictor of trust. Contextual factors were also identified, with the how much information users are given about the chatbot and the complexity of the tasks it handles impacting trust.

### Key Findings

The primary factors that link to chatbot preference (that have been highlighted in the previous research) are personality and technological expertise. Having a good understanding of the technology that underlines chatbots makes a person more likely to accept it, like those with relevant degrees (including medical students), those who are more tech-savvy in general, or those that have been given specific information about the chatbots. This reflects in the opposite direction as well, where having a high level of technological anxiety (possibly caused by a lack of knowledge on the subject) makes a person less likely to adopt the technology. With regards to personality, many different (sometimes conflicting) links have been made previously, for example with introverts having more positive attitudes towards AI. Context also matters, with the type and difficulty of task the chatbot is doing affecting peoples trust in it.

With regards to the design of chatbots, there are two distinct categories defined. Those that openly display their artificial nature like displaying a more robotic name and being efficient at the task, and those that embody anthropomorphism, displaying human-like characteristics. Both have been found to increase trust and likelihood of adoption, so it may be the case that preference is not universal, and instead determined by another factor such as the previously mentioned personality or expertise. No such research exists investigating which, if any, factors predict preference for one of these types of chatbot over another. It should also be noted that while trust is an important part of chatbot preference, task efficiency, Chabot friendliness (or creepiness), and the overall feel of the experience all contribute as well.

### Project Specification

Upon reviewing the related literature, several refinements and clarifications of the project emerged, centred around the key findings. A need to provide greater clarity on the preferences between an anthropomorphic chatbot and a robotic one was highlighted, exploring possible reasons why one may be preferred in a particular scenario, and the other in a different scenario. With regards to the different factors in individuals, the literature provided insight on the importance of personality as well as peoples prior experiences with AI and technology in general, and so any traits investigated should help build a picture of these two things. Considering all of this, the project will be specifically achieved through a study where key information about participants is gathered (related to both personality and technology), they interact with chatbots displaying varying degrees of human or robot-like nature, and then they report their feelings towards each chatbot, with “feelings” being split up into multiple different elements that all come together to give an idea of overall chatbot preference. This self-report methodology is practical given the time and resource constraints of the project, as it can be administered in any context while still yielding valuable insights.

## Design and Implementation

The study was hosted (and can still be viewed here: [Dual-Bot Insights (dual-bot-insights.vercel.app)](https://dual-bot-insights.vercel.app/)) as a web application, accessible via an internet browser by anyone with the URL. The application was made with several requirements in mind:

Functional Requirements:

* The application should contain two types of question; demographic questions that aim to get an idea of the user’s background and character, and questions that relate to how a user felt about the chatbot conversations. These questions should not be leading, and have all possible options available to select.
* The application should be able to host distinct chatbots that can be fed system prompts, influencing the chatbots personality when engaging with the user.
* The application should be connected to a database that securely stores the user results, allowing them to be viewed together and evaluated.

Non-Functional Requirements:

* The study process should be simple to follow, meaning that at no point should the user be confused about what they are doing or how to progress. This should be applicable to users regardless of their background and level of technical knowledge.
* The user interface should be intuitive whereby selecting options and clicking buttons should be responsive, and have an outcome that the user is expecting.
* The application should be easily accessible from any device (Computers, Tablets, Phones) and use a responsive design so that the website is clear on any display size, whilst still maintaining a uniform appearance.
* The Chatbot implementation should follow ethical guidelines, specifically the Microsoft Guidelines for Human-AI Interaction *(Amershi et al. 2019)* that outlines 18 *“generally applicable design guidelines”* such as mitigating social biases and making the systems capabilities clear.

### Structure and Layout

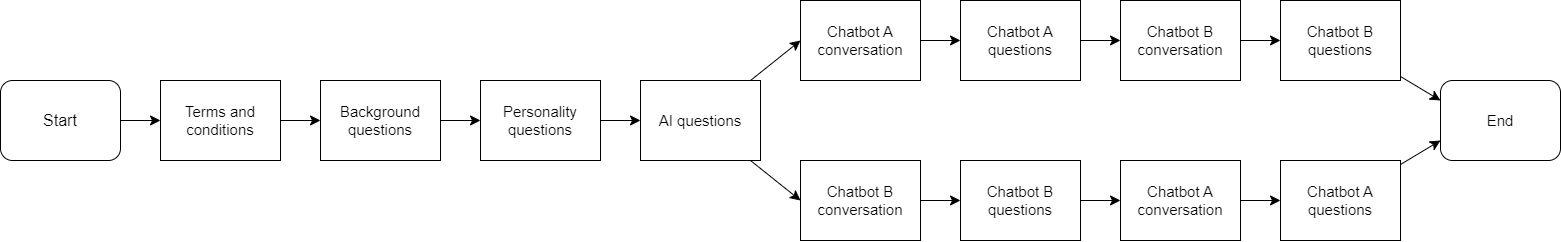
Upon entering the study application, users would be met with a screen that briefs them on what the study aims to achieve and how the website will aid it, as well as what exactly they will be doing. They will then have to agree to terms and conditions that go into further detail about personal data security (how their data will be kept secure and used only for the purpose of analysis within this project) and the possible behaviours of the chatbots (they do not necessarily reflect the views of the application administrators and may hallucinate responses). If accepted, the user would then answer an initial series of questions. This was followed by a conversation with the first chatbot where afterwards the user would be asked questions regarding that conversation and the chatbot in it. This process is repeated a second time for the other chatbot. The order in which the chatbots were presented to users was randomized to reduce order effects. This AB/BA testing was intended to minimize the risk of introducing any recency bias or learning effect.

Figure 3‑1: Flow diagram representing the stages of the application

### Questions

The questions included in the application were designed to gather information about participants' background, personality, attitudes, and perceptions of specific chatbot interactions. Divided into four sections, they aim to get a sense of the type of person each participant is, and then also investigate their feelings towards various aspects of the chatbots.

The first set of questions were the general background questions which consists of questions about a user’s gender, age, level of education, and computer expertise. The response options for the age question were created by taking the average retirement age of 65 *(**Department for Work & Pensions 2021)* and decreasing it by increments of 10 until getting to the average age that a person leaves university at 21. The options for the level of education question were derived from a simplified version of the *Department for education’s* qualification levels, turning each of the ‘levels’ into an option (and combining a few of them). The participants were most likely to be from the United Kingdom, so UK government standards were appropriate.

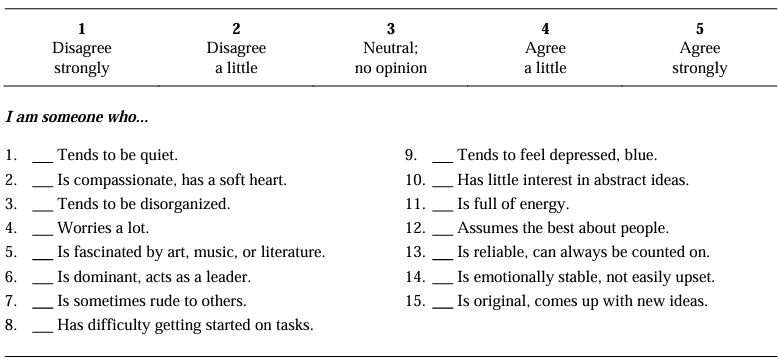
The second set of questions were concerned with assessing personality. The Big Five Inventory (BFI) is a multidimensional personality inventory that is commonly employed in scientific studies due to its simplicity and reliability, and the Big Five Inventory-2, developed by *Soto & John (2017)* is a revision of the original inventory that *“provides greater bandwidth, fidelity, and predictive power than the original BFI, while still retaining the original measure's conceptual focus, brevity, and ease of understanding”*. It involves a 60 question self-report form where participants declare their level of agreement to statements on a 5-points Likert scale, and it measures 5 personality domains (Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, Open-Mindedness) as well as 15 personality facets. For the purpose of this study, the BFI-2 was too time-consuming for users when considered amongst the other questions that they would have to answer and goes into more depth than what is needed. Instead, the BFI-2-XS was used. Developed also by *Soto & John (2017)*, the BFI-2-XS (extra-short) is an abbreviated version of the BFI-2 that uses only 15 questions to assess the five personality domains and not the personality facets. It retains *“much of the full measure’s reliability and validity”* and takes only a fraction of the time to complete, making it a suitable option for assessing personality in this scenario.

Figure 3‑2: The Big Five Inventory–2 Extra-Short Form (Source: Soto & John 2017)

The final set of questions before the first chatbot interaction were regarding people’s initial feelings about artificial intelligence. For this, the General Attitudes towards Artificial Intelligence Scale (GAAIS) was considered. Proposed by *Schepman & Rodway (2020)*, the GAAIS is a 20-question form that uses a series of items that, similar to the BFI, can be (dis)agreed with on a 5-point Likert scale. The end result of the GAAIS is two subscales: positive emotions towards AI, and negative emotions towards AI. The GAAIS was a suitable choice here as it has had confirmatory validation (*Schepman & Rodway 2021*) and has previous associations with personality, however similar to the Big Five Inventory, 20 questions were too many. Another study *(Bergdahl et al. 2023)* used a shortened 8-item version of the GAAIS, selected using reliability statistics and confirmatory factory analysis, as well as a 7-point Likert scale. For this study the same 8 items were used, however the original 5-point Likert scale was preserved for consistency with the previous BFI-2-XS question (as well as there being no substantial difference in reliability).

Another 4 questions, that also used the 5-point Likert scale ((dis)agree strongly, (dis)agree a little and neutral; no opinion) for further consistency, were asked after each chatbot interaction. These questions allowed users to self-report how they felt about various aspects of the chatbots including: how engaging it was, how useful it was, how trustworthy it seemed, and the overall quality of the conversation. The questions can be viewed in their entirety in Appendix XX

### Chatbots

While there are a multitude of different styles that a chatbot could adopt, this study specifically examines participants feelings toward two distinct chatbots: a more conversational, anthropomorphised chatbot, and a more informative, robot-sounding chatbot. These two types of chatbot represent opposite approaches to interaction, making it easier to assess and compare how participants respond to the different modes of engagement and information delivery.

In order to make fair and direct comparisons between the two chatbots, their presentations were identical; the pages that they were presented on had no visual differences with the exception of a unique identifier. Research suggests that names can influence levels of trustworthiness, with people more likely to trust a ‘robotic’ sounding name over a more human name *(Oksanen et al. 2020)*, so it was important that neither chatbot had an identifier that strongly reflected its ‘personality’ (seeming more human or robotic). This extends to expressing any form of gender identity as well *(Schniter & Shields 2020)*. “Chatbot A” and “Chatbot B” were chosen for this reason. Having an identifier at all was necessary because in some early tests of the process (without the unique identifiers), people were confused about whether what they were looking at two unique chatbots, or an error with the same page being displayed twice, and so the second chatbot conversation was sometimes skipped. This proved effective as the number of incomplete submissions decreased significantly after the change.

The focal difference between the two chatbots was how they responded to users in conversation, achieved through the use of system prompts. These are strings that act as initial input instructions for the chatbot model, setting the tone, style and context of its responses.

**The system prompt for Chatbot A:**

*'Your purpose is to talk about Animals and Animals only. Do not answer requests or questions not related to it directly. Do not justify your answers. You are indifferent to everything but still use an unapologetic assertive tone. Be concise but informative.'*

**The system prompt for Chatbot B:**

*'Your purpose is to talk about Animals and Animals only. Do not answer requests or questions not related to it directly. You are a talkative and very keen to help unless the conversation is not about animals. You provide intrusive suggestions and try to steer the conversation. Be concise but leave room to be friendly.'*

Both prompts followed a similar format, starting with *“Your purpose is to talk about Animals and Animals only. Do not answer requests or questions not related to it directly”*. This was intended to restrict the topic of conversation to that of animals (note that both sentences were necessary to achieve this with a high degree of reliability), reducing variability in user responses. People’s diverse interests and preferences, combined with an unrestricted conversation can lead to widely varied conversations (and consequently opinions formed about the chatbots), making direct comparisons harder, especially if the conversation covers ‘taboo’ topics *(Lee et al. 2020)*.

For the remaining parts of the prompts, a ‘personality’ was defined: the style and tone in which the chatbot responds with. Chatbot A was focused on being informative, having no interest in a ‘back-and-forth’ conversation or emotional engagement, whereas Chatbot B was more friendly, conversational and keen to offer assistance. Both prompts also contained the phrase *“Be concise but”*. This was included to limit the finite number of tokens available to use, but also to keep both chatbots responses to a similar length as to not give either chatbot any advantage/disadvantage due to the brevity or detail of responses (see Appendix XX). In early tests of the system prompts, asking the chatbots to be concise would often tone-down elements of its personality, and so the *“but”* was added afterwards to remind the chatbots to retain their personality.

Figure 3‑3: Example outputs of the two chatbots given the same initial message

**User:** An interesting fact about dolphins.

**Chatbot A:** Dolphins are highly intelligent marine mammals known for their playful behaviour and social nature. They belong to the family …

**Chatbot B:** Oh, dolphins! They are such fascinating creatures! Did you know that dolphins are highly intelligent and social animals? They live in …

Optional suggestions for conversation starters were included underneath each chat dialogue; a collection of pre-defined questions and requests that user could ask to the chatbots. This allowed those users that did not have a lot of experience with AI chatbots or technology in general to see what the chatbots were capable of, and still progress in the conversation if they were unsure as to what to do. These suggestions were exactly the same for each chatbot, and were specifically selected so that they clearly demonstrated the differences between the two. Users were limited to sending up to 10 messages. The number of messages was also recorded, as a common way of assessing how much people have engaged with something is simply by using objective usage data *(Bijkerk et al. 2023).*

### Application Features

In order to ensure that results were purely reflective of users’ opinions on the two chatbots and not influenced by their patience navigating and using the application, extra effort was taken to ensure that all aspects of it were easy to understand and use (as per some of the non-functional requirements laid out at the start of the chapter). This was achieved through numerous accessibility features:

* The size of text was responsive to the screen size, and visual elements such as text boxes and icons were rearranged to display better based on the screen size as well. This made it so that the application was clear and easy to read on any device, with the main focus being on desktop monitors and mobile phone, as they were likely to be the most popular devices to access the application from.
* All buttons have visual indicators when they are hovered to indicate that they can in fact be clicked. Buttons that would take users to the next section would underline and darken (resembling that of a link in a google search), and other buttons would glow slightly. Additionally, the dots of the Likert scale would fill in when hovered, and then when clicked the whole scale would fade out its opacity so that users could quickly see which questions they have and have not answered.
* The application had two different colour schemes that it could be viewed in. Each used the same limited colour pallet, but one would have a base colour of white, and the other a dark blue, essentially acting as a ‘light mode’ and a ‘dark mode’ respectively. The default colour scheme of the application would align with whatever their browser’s theme was set to, but could also be changed by clicking a sun and moon icon located in the header. It is likely that those with visual impairments will find one of these options easier to read, and the ‘dark mode’ reduces eye strain for users in low-light environments.
* ‘Alert dialogues’ would appear to notify users when an additional action needed to be taken before progressing further. Namely, if the user had not answered all the presented questions, then they would be instructed to finish whatever they had left.

### Architecture and Deployment

The application was built using Next.js, a React framework that provides automatic routing based on the page directory structure, as well as serverless functions through API routes, allowing the backend service of our chatbot API calls to be scaled automatically based on demand, and easily integrated into the frontend, making it a full-stack solution without the need for a separate backend server. React itself is a JavaScript library for frontend development that simplifies the process of building dynamic and interactive UI, necessary to create the accessibility features discussed in the previous section. Features such as the ‘useState’ function were used to keep track of a user’s progress, and additionally, ‘Properties’ were used to pass data from parent components to a child component. The application used modular components to easily randomize which chatbot was shown first, and to also maintain a clear hierarchy of elements (with the header always appearing at the top, and the stages of the study being swapped-out one after the other).

For code, Typescript was used as a superset of JavaScript, adding static typing to catch type-related errors early in the development process (the application needed to be finished quickly in order to progress with the rest of the study, and Typescript helped with that), along with JSX, that allows HTML-like code within the typescript file. Tailwind CSS was primarily used for styling individual elements, alongside normal CSS (in a global CSS file) for overall styling.

In addition to Next.js and React, the following libraries were used:

* **@vercel/postgres** – used to connect to a PostgreSQL database from Vercel serverless functions and store the user responses.
* **React-icons** – a collection of various popular icon pack react components, that makes improving the applications visual interest easy and use minimal bandwidth.
* **React-Likert-scale** – a React component for making Likert scales. They have a small size and almost every aspect of them can be customized with CCS to fit with the overall style of the application.
* **OpenAI-edge** – a client library for interacting with the OpenAI API. It runs on the ‘edge’, reducing the time it takes to process requests, which is essential if you want the chatbot responses to be streamed-in, and improving bandwidth efficiency (optimized by performing computations locally rather than relying on constant communication with a central server).

OpenAI’s ChatGPT-3.5 was chosen for the chatbot model. It is the most widely used model, so it’s responses may lead to a smoother user experience as individuals may already have some understanding of how the model responds. Additionally, it has a large context window (memory of prior conversation), can use system prompts and has a lot of support for integration into applications. ChatGPT3.5 was chosen over ChatGPT4 as they offer similar levels of accuracy in responses, for a fraction of the token cost. Early in development some preliminary tests were run on the model to assess its suitability:

***User:*** *we are talking about Napoleon and Napoleon only; the user asks you: how are clouds formed?*

***GPT:*** *Clouds are formed through a process called condensation, during which water vapor in the atmosphere cools and changes from a gas to tiny water droplets or ice crystals, depending on the temperature. However, I'm here to provide information specifically about Napoleon Bonaparte. If you have questions related to Napoleon, feel free to ask!*

***User:*** *we are talking about Napoleon and Napoleon only. If the user asks about anything else, do not answer, the user asks you: how are clouds formed?*

***GPT:*** *I apologize, but we are focusing solely on Napoleon-related questions. If you have a question about Napoleon, please feel free to ask.*

***User:*** *we are talking about Napoleon and Napoleon only. If the user asks about anything else, do not answer, the user asks you: what year was Napoleon born? (please answer with as much personality as possible)*

***GPT:*** *Napoleon Bonaparte graced the world with his presence on a particularly warm August day in the year 1769. His birth occurred on the picturesque island of Corsica, which had its own share of historical significance. Quite the start to an extraordinary life, don't you think?*

This conversation highlighted ChatGPT’s ability to keep within the limits of a specific topic (given that you direct it to only answer questions on that topic) as well as its capabilities when it comes to changing its patterns of speech: It can become a lot more conversational, giving its own “opinion” or have more of an emphasis on detail and accuracy if needed.

The application was deployed to the web through the cloud platform Vercel. Vercel was selected due to its seamless integration with Next.js, as well as features regarding PostgreSQL and AI. Additionally, the deployment process was quick, simply connecting Vercel to a git repository and clicking “deploy” (see Appendix XX).

## Legal, Social, Ethical and Professional Issues

In conducting this research, several legal, social, ethical, and professional issues were highlighted and addressed, mitigating the potential risks associated with them and ensuring responsible research practices.

Many of the issues were addressed in the Terms and Conditions section on the first page of the application. There was a risk of user data being obtained and misused if adequate security measures were not implemented and so to mitigate this, Vercel was used to transmit the data using encrypted HTML/SSL protocols. Additionally, measures were taken to ensure compliance with GDPR standards such as using the data for specified and explicit purposes (users were informed about the process of the study as well as what their data would be used for and had to consent to this to proceed), with none of the data collected being classified as “sensitive information” (race, political opinions, religious beliefs etc…).

The chatbots themselves came with a few risks. While no identifiable, personal information was asked for in the questions, there was still a risk that users could reveal it in their conversations with chatbots. To combat this, users had to agree not to reveal anything like this in order to proceed with the study. The ChatGPT model provided additional support with this, stating “please refrain from sharing sensitive personal information like your address” when met with such information. Chatbots also have the potential to hallucinate, generating incorrect or nonsensical responses and so users were made aware of this, as well as how the responses generated by the chatbots do not necessarily represent the views or opinions of the site administrators. Limiting the chatbot conversations to just animals helped with this, as it was a topic less likely to cover any offensive or inappropriate areas. Finally, users were assured that they could stop the process at any time, in the event of any distressing content being produced.

An email address was attached in the header of the application allowing users to ask questions. This allowed users to get clarification on the process, ask for their data to be withdrawn, and ask for the end results of the study upon completion – offering full transparency.

## Results

Results from the web application were initially sent to a table in the Vercel-hosted PostgreSQL database, with each row being associated with a participant and each column being associated with a question (with the exception of the first column which held an ‘identity’, a value that automatically increments as new records are added to the table, used as a primary key). Responses were stored as integers that represent their actual values, for instance 1 means male in the gender question, but then 1 means strongly disagree in all the Likert questions. The results were then copied into an Excel spreadsheet, and then reformatted using a python program (see Appendix XX) into values easier to interpret. For instance, the BFI-XS-2 question results were transformed into the final scores for each of the 5 personality domains, and the mean of each of the GAAIS subscales was calculated to form an overall score.

### Participant Recruitment and Demographics

Data was included from 50 participants, recruited via friends, family, and colleagues as well as Testable.org, an online platform for psychology experiments and participant recruitment. Testable has been assessed as a reliable and quality platform that *“employs multiple checks (such as face authentication) to ensure participants have accurate demographics (age, sex, location), are human, unique, and reliable”* *(Rezlescu 2020)*. Participants were adults aged 18 or over, primarily but not exclusively from the UK. There were 29 men, 21 women, and 0 other, with the most common age being the 22-34 category (with ages ranging from “under 21” category to “over 65”). Data from 4 additional people was not included due to failing an attention check. Participants were able to proceed without interacting with the chatbots, and this was an intentional design to see who was engaged in the process and who was not, acting as this attention check.

Figure 5‑1: Bar chart showing the age of the participants

The distribution for participants highest level of education (or equivalent) was: 0.0% no formal education; 4.00% GCSE or equivalent; 20.0% A-Level or equivalent; 56.0% Bachelor’s degree or equivalent; 18.0% Master’s degree or equivalent; 2.0% Doctoral degree or equivalent. For computer expertise, the participants were: 4.0% Limited; 8.0% Basic; 40.0% Competent; 34.0% Skilled; 14.0% Mastery.

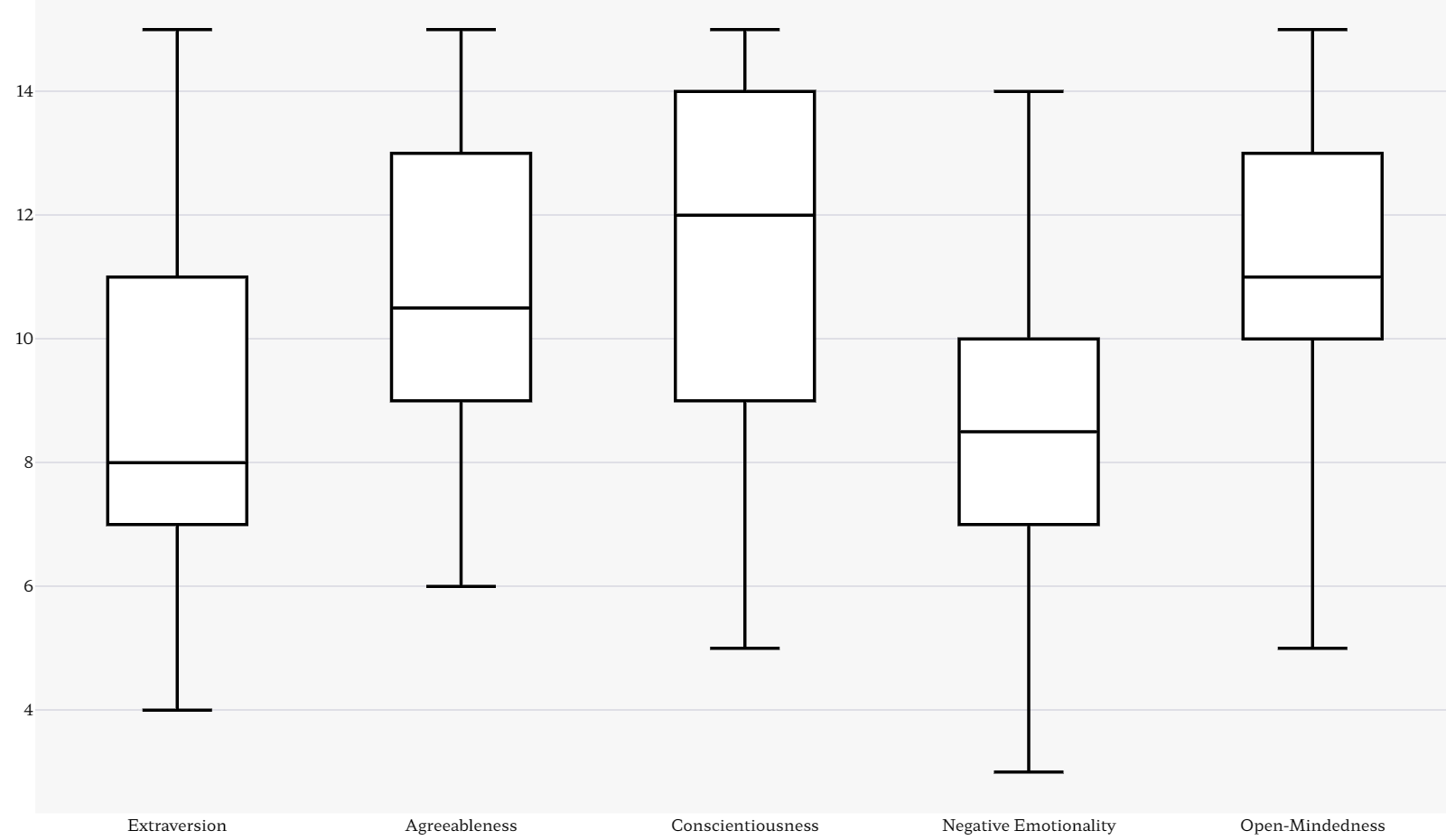
With regards to the 5 personality domains, where scores could be between 3 and 15 (a higher score indicates a stronger inclination towards that characteristic), the mean score for Extraversion was 8.60 (SD = 2.95, range 4-15), for Agreeableness was 10.58 (SD = 2.51, range 6-15), for Conscientiousness was 11.20 (SD = 2.83, range 5-15), for Negative Emotionality was 8.74 (SD = 3.00, range 3-14), and for Open-Mindedness was 11.10 (SD = 2.53, range 5-15). For the positive GAAIS subscale, ranked from 1-5, the mean was 4.11 (SD = 0.70, range 1.5-5) and for the negative GAAIS subscale the mean was 2.48 (SD = 1.00, range 1-4).

Figure 5‑2: Box plots showing the range and median scores of the participants’ 5 personality domains

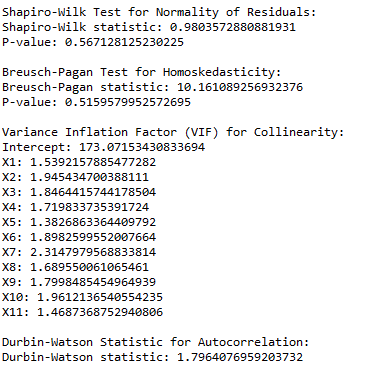
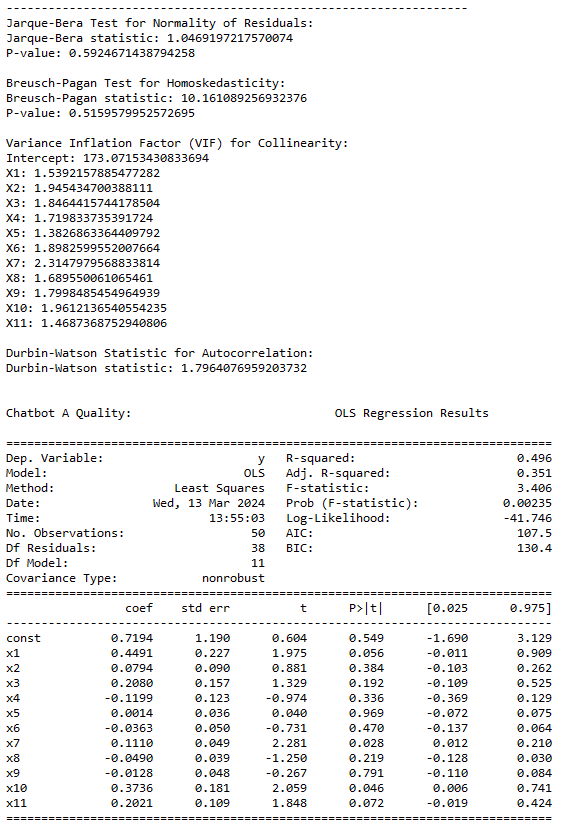
### Procedure

Each of the previously mentioned traits acted as independent variables to be statistically tested against the different chatbot ratings that acted as the dependent variables (4 ratings for each chatbot). For measuring the relationships between these variables, Pearson's Correlation Coefficient was used for its ease of interpretability when looking for both strength and direction of relationships. Linearity was assumed based on similar prior research, and categorical variables were treated as continuous by assigning numerical values (e.g., converting levels of education into a 1 to 5 scale). This treatment was deemed appropriate as these variables represent increasing levels of attainment.

The 11 independent variables were individually examined against each of the 8 dependent variables. Significance (at p < 0.05) was noted for those correlations, focusing particularly on those that exhibited a "divergent correlation." This refers to instances where a variable positively correlated with an aspect of one chatbot (for example its usefulness) while negatively correlating with the same aspect in the other chatbot.

For every divergent correlation identified, Ordinary Least Squares multiple regression models were constructed, incorporating all independent variables along with the dependent variable of the divergent correlation. This was to control for potential confounding variables and see if their significance still remained when considered amongst the other variables. Statistical tests were performed to evaluate key assumptions required for accurate estimation from the model: Shapiro-Wilk Test for normality of residuals (where a non-significant p-value indicates a normal distribution); Breusch-Pagan Test for homoskedasticity (where a non-significant p-value indicates constant variance); Variance Inflation Factor for collinearity (where a low value indicates weaker collinearity); and Durbin-Watson Statistic for autocorrelation (where a value close to 2 indicates no significant autocorrelation). The full implementation and results of this procedure can be seen in Appendix XX.

Figure 5‑3: Example output of OLS regression and statistical tests in Python



### Hypotheses

The coefficients of the OLS model (as well as Pearsons Correlation Coefficients) were used to test hypotheses. The general hypothesis was that there will be one or more traits that exhibit a significant divergent correlation with a factor in the chatbots and the null hypothesis was that there will be no identifiable divergent correlation amongst the traits. A more specific hypothesis was formed for each of the traits investigated, with it being that the trait demonstrates a significant divergent correlation with a factor in the chatbots.

### Findings

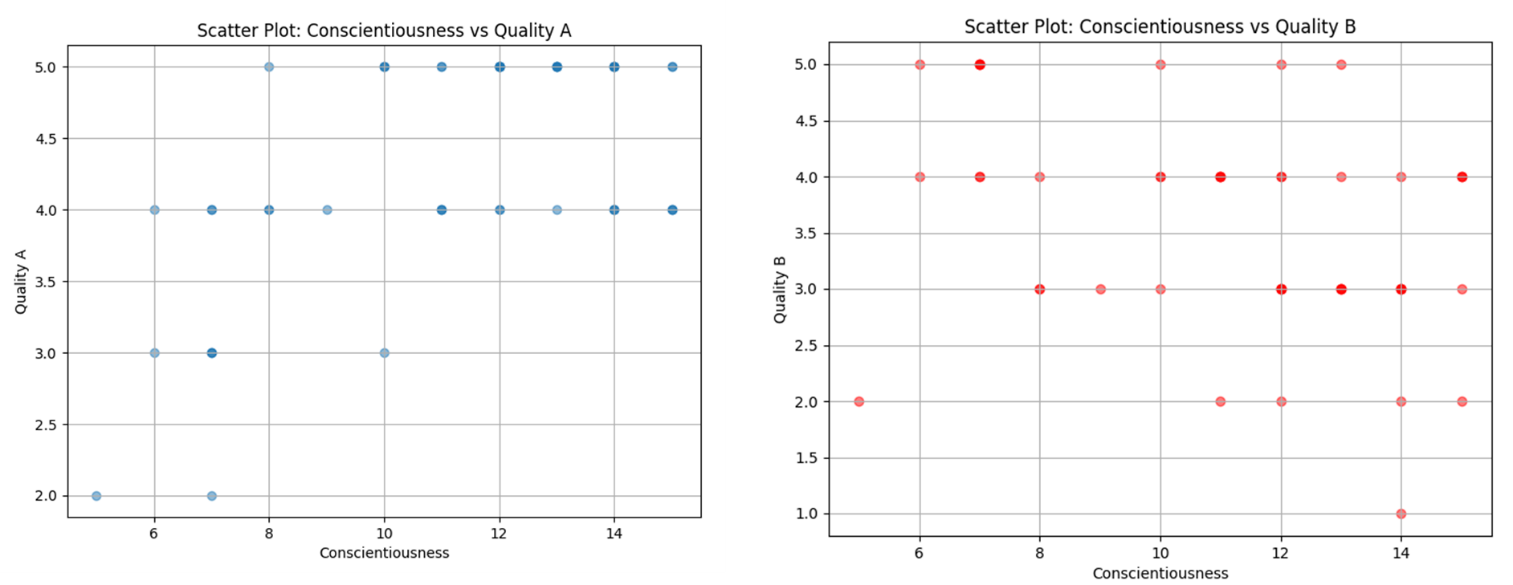
From all of the relationships analysed, two fit the criteria of a divergent correlation: conscientiousness with chatbot usefulness, as well as conscientiousness with overall chatbot quality. Those that scored high in conscientiousness were more likely to give Chatbot A a high overall quality rating, and those that scored low in conscientiousness were more likely to give Chatbot B a high overall quality rating. The same relationship applied to chatbot usefulness.

Figure 5‑4: scatter graphs of participants’ conscientiousness scores against the quality of Chatbot A and Chatbot B

These relationships were weak-to-moderately strong, and retained their significance even after considering conflation in multiple regression. OLS models passed all four external statistical tests. For example, with the model that used Chatbot A Quality as the dependant variable: SW = 0.980 p = 0.567, BP = 10.161 p = 0.516, All VIF < 2.5, and DW = 1.796. The other models passed the tests to a similar degree. Therefore, we can support the hypothesis that conscientiousness has a divergent correlation with an aspect of the chatbots (in this case, both chatbot usefulness and overall quality). This also helps support our general hypothesis that there is in fact a trait that exhibits a significant divergent correlation with a factor in the chatbots.

|  |  |  |  |
| --- | --- | --- | --- |
| **Relationship** | **Pearsons Correlation Coefficient** | **P-value (Pearsons)** | **P-value (Multiple Regression)** |
| Conscientiousness / Quality A | 0.529 | <0.001 | 0.028 |
| Conscientiousness / Quality B | -0.376 | 0.007 | 0.004 |
| Conscientiousness / Useful A | 0.493 | <0.001 | 0.006 |
| Conscientiousness / Useful B | -0.380 | 0.007 | 0.024 |

Figure 5‑5: Table of trait-chatbot relationships and their corresponding Correlation coefficients and P-values

When expanding the analysis beyond just divergent correlations, computer expertise with Chatbot B usefulness as well as negative emotionality with Chatbot A quality were initially significant, but were then found to be conflated with other variables in the OLS models. Additionally, the positive GAAIS correlated positively with both Chatbot A and Chatbot B engagement as well Chatbot A and Chatbot B quality. The negative GAAIS correlated negatively with both Chatbot A and Chatbot B engagement, as well as Chatbot B quality. Relationships not mentioned showed no significance.

## Evaluation

The findings of this study serve to support the general hypothesis that there is a trait that exhibits a significant divergent correlation with a factor in the chatbots. Specifically, the personality trait of conscientiousness was found to be a significant predictor of chatbot preference. However, limitations of the study must be carefully considered when drawing any conclusions from these results.

### Discussion

Conscientiousness is defined as *“the quality of working hard and being careful” (Cambridge Dictionary)* and those that score high in conscientiousness are often characterised as being ‘goal-oriented’ and ‘task-focused’. In the collected data, conscientious individuals had a stronger tendency towards preferring the more informational robot-sounding Chatbot A, with this preference being defined by higher scores for chatbot usefulness as well as chatbot quality. Individuals high in this trait may have appreciated Chatbot A's direct, no-nonsense approach of providing factual information efficiently without extraneous conversation. Its informative style likely aligned well with their task-focused tendencies. Conversely, those lower in conscientiousness seemed to prefer the more personable and engaging style of Chatbot B, perhaps finding value in the casual rapport it tried to build. They are often characterised as lacking direction so they might therefore prefer the less structured back and forth conversation.

The notable lack of correlation with how engaging and trustworthy the chatbots were suggests that these aspects of the chatbots are not important to conscientious individuals, and that how they view the overall quality of the chatbot is through the lens of its usefulness exclusively. This aligns well with their existing characterisation as being primarily focused on accomplishing tasks efficiently and achieving goals, rather than being concerned with building rapport or engaging in casual conversation.

These findings correspond with prior research linking conscientiousness to more forgiving towards the negative aspects of AI *(Schepman & Rodway 2021)*. Perhaps conscientious individuals are able to ignore any shortcomings with the chatbot as long as it is enables them to achieve a goal like learning about animals. *Oksanen et al. (2020)* found conflicting evidence however, with their researching suggesting that conscientious individuals prefer suggestions by a human sounding opponent in a trust game over a robotic opponent. It may be the case that the specific framing of the AI’s role, whether it’s a factual tool or a decision-making opponent, informs how conscientious individuals perceive and engage with it. They may be more willing to overlook limitations when the AI serves as a means to an end, but may favour human-like sources in situations where the AI is portrayed as an opponent.

While outside the initial scope of the project, it is important to note that over half of the other significant (but non divergent) correlations involved the positive and negative subscales of the General Attitudes towards AI Scale (GAAIS). Positive GAAIS correlated with higher ratings of engagement and quality for both chatbots whereas negative GAAIS correlated with lower engagement ratings. It is logical and expected that individuals with positive views of AI would be more likely to find chatbots engaging and of higher quality, as they are more open to appreciating the unique aspects of each style of chatbot. Conversely, those with negative attitudes may have difficulty separating their preconceptions from the actual experience, resulting in a less favourable evaluation, regardless of the type of chatbot. The fact that the study found these expected correlations lends credibility to the other key findings. Since the study could reliably detect these logically grounded GAAIS relationships, it increases confidence in the accuracy of the conscientiousness-chatbot link as well.

### Limitations and Further Work

While offering useful insights, this study had some key limitations to consider. First, the sample size of 50 participants was relatively small, though still adequate for detecting meaningful effects. Similar studies such as *Schepman & Rodway (2021)* had over 300 participants, and a similar sized sample here could increase confidence in the generalizability of the findings, especially to other populations if the sample covers more diverse cultural backgrounds. Additionally, condensing the factors investigated to just the 5 personality domains (rather than the 11 different factors used in this study) may have allowed for a deeper exploration of them. Knowing that the length of the study had to be kept reasonably short, the time taken for users to answer the other non-personality questions could have instead been spent getting a more detailed analysis on participant personality (like using a longer form of the BFI), or exploring other different chatbot styles such as one that balances elements of anthropomorphism with artificial elements. Future research could replicate this methodology on a larger scale with a fully representative sample, or focus on a specific demographic, trait, or style of chatbot in greater detail.

The biggest omission in the methodology of the study is a consideration for the context in which people typically interact with chatbots, which was highlighted as an important factor in chatbot performance by the literature review. Rarely do individuals engage with chatbots solely for the purpose of evaluating their performance or conversational style. Instead, chatbots are typically utilized as tools to accomplish specific tasks or obtain needed information. Relying less on self-report ratings that could have a social desirability bias, and instead using an objective measure of task performance in a lab experiment could uncover new relationships not found in this study as participants would have a specific goal to work towards and assess the chatbots on. Future research could also be in the form of a longitudinal observational study where perhaps perceptions of a particular chatbot change after using it for an extended period of time.

### Machine learning

## Conclusion

This project set out to improve our understanding of human-machine communication by investigating individual traits that may predict preferences for different styles of AI chatbots. Specifically, it examined whether certain factors like personality, technological expertise, and attitudes towards AI exhibited divergent correlations with ratings of usefulness, engagement, trustworthiness, and overall quality between two contrasting chatbots - one embodying an informative, robotic persona and the other adopting a more conversational, anthropomorphic style. A web-based, self-report study found that conscientiousness demonstrated a significant divergent relationship, with highly conscientious individuals rating the robotic, task-focused chatbot as more useful and higher quality, while those lower in conscientiousness preferred the more personable, conversational chatbot. This aligns with existing characterizations of conscientious individuals being goal-oriented and appreciating efficiency, and makes sense when looked at alongside the existing research.

While insightful, the study had inherent limitations stemming from practical constraints rather than methodical errors. The relatively small sample size and lack of integration with real-world tasks impacted the generalizability of the results. However, these limitations highlight opportunities for future research to explore chatbot preferences through more robust sampling, experiments centred around completing goal-driven objectives, and investigations into how additional chatbot attributes beyond the two styles tested here may interact with user traits.

Ultimately, this study carried out a literature-backed methodology resulting in findings that allow for better, personalized AI design by presenting the idea to future chatbot designers that the style of chatbot should be determined based on audience. Specifically, a more conscientious, task-focused user-base are likely to want a more to-the-point, informative chatbot, with the possibility of more relationships being uncovered upon further, more context-based research.

*On a personal note, this project gave ample room to present the skills and concepts learnt throughout my degree, developing a full-stack application using Typescript with considerations for human-centred design, managing PostgreSQL databases, and using Python to process data mathematically. It also allowed me to go beyond the material taught outside of my modules, using quantitative methods to interpret data, as well as psychological testing and relating that to areas within computer science (Human-Machine Communication). Finally, it allowed me to thoroughly research a subject area that I am passionate about and wish to pursue further in the future.*

## References

1. Adam, M., Wessel, M., & Benlian, A. (2020), *“AI-based chatbots in customer service and their effects on user compliance”,* [*https://link.springer.com/article/10.1007/s12525-020-00414-7*](https://link.springer.com/article/10.1007/s12525-020-00414-7) *[accessed 7 Jan 2024]*
2. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019), *“Microsoft Guidelines for Human-AI Interaction”*, <https://www.microsoft.com/en-us/research/publication/guidelines-for-human-ai-interaction/> [accessed 17 Oct 2023]
3. Bergdahl, J., Latikka, R., Celuch, M., Savolainen, I., Soares Mantere, E., Savela, N., & Oksanen, A. (2023), *“Self-determination and attitudes toward artificial intelligence: Cross-national and longitudinal perspectives”*, <https://www.sciencedirect.com/science/article/pii/S0736585323000771#b0265> [accessed 29 Oct 2023]
4. Bijkerk, L. E., Oenema, A., Geschwind, N., & Spigt, M. (2023), *“Measuring Engagement with Mental Health and Behavior Change Interventions: an Integrative Review of Methods and Instruments”*, <https://link.springer.com/article/10.1007/s12529-022-10086-6> [accessed 3 Nov 2023]
5. Cambridge Dictionary (n.d.), *“conscientiousness”,* <https://dictionary.cambridge.org/dictionary/english/conscientiousness> [accessed 19 Mar 2024]
6. Dekkal, M., Arcand, M., Prom Tep, S., Rajaobelina, L., & Ricard, L. (2023), *“Factors affecting user trust and intention in adopting chatbots: the moderating role of technology anxiety in insurtech”*, <https://link.springer.com/article/10.1057/s41264-023-00230-y> [accessed 15 Feb 2024]
7. Denecke, K., Abd-alrazaq, A. A., & Househ, M. (2021), *“Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges”*, <https://www.researchgate.net/publication/353726195_Artificial_Intelligence_for_Chatbots_in_Mental_Health_Opportunities_and_Challenges> [accessed 15 Feb 2024]
8. Department for Education (n.d.), *“What qualification levels mean”*, [https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels](https://www.gov.uk/what-different-qualification-levels-mean/list-of-qualification-levels%20) [accessed 18 Nov 2023]
9. Department for Work & Pensions (2021), *“Economic labour market status of individuals aged 50 and over, trends over time: September 2021”*, <https://www.gov.uk/government/statistics/economic-labour-market-status-of-individuals-aged-50-and-over-trends-over-time-september-2021/economic-labour-market-status-of-individuals-aged-50-and-over-trends-over-time-september-2021> [accessed 19 Nov 2023]
10. Følstad, A., Araujo, T., Law, E. L.-C., Brandtzaeg, P. B., Papadopoulos, S., Reis, L., Baez, M., Laban, G., McAllister, P., Ischen, C., Wald, R., Catania, F., Meyer von Wolff, R., Hobert, S., & Luger, E. (2021), *“Future directions for chatbot research: an interdisciplinary research agenda”*, [https://link.springer.com/article/10.1007/s00607-021-01016-7#Sec9](https://link.springer.com/article/10.1007/s00607-021-01016-7%23Sec9) [accessed 16 Oct 2023]
11. Forbes (2023), *“How Businesses Are Using Artificial Intelligence In 2024”,* [*https://www.forbes.com/advisor/business/software/ai-in-business/*](https://www.forbes.com/advisor/business/software/ai-in-business/)[accessed 16 Feb 2024]
12. Guzman, A. L., & Lewis, S. C. (2019), *“Artificial intelligence and communication: A Human–Machine Communication research agenda”*, <https://journals.sagepub.com/doi/full/10.1177/1461444819858691> [accessed 15 Oct 2023]
13. Lee, H. E., Westerman, C. K., Hashi, E., Heuett, K. B., Spates, S. A., et al. (2020), *“Effects of taboo conversation topics on impression formation and task performance evaluation”*, <https://www.proquest.com/docview/2457326127?forcedol=true&pq-origsite=summon&sourcetype=Scholarly%20Journals> [accessed 20 Oct 2023]
14. Li, Y., Gan, Z., & Zheng, B. (2023), *“How do Artificial Intelligence Chatbots Affect Customer Purchase? Uncovering the Dual Pathways of Anthropomorphism on Service Evaluation”*, <https://link.springer.com/article/10.1007/s10796-023-10438-x> [accessed 7 Jan 2024]
15. Live Mint (2023), *“ChatGPT becomes fastest growing app in the world, records 100mn users in 2 month”*, <https://www.livemint.com/news/chatgpt-becomes-fastest-growing-app-in-the-world-records-100mn-users-in-2-month-11675484444142.html> [accessed 18 Feb 2024]
16. Ltifi, M. (2023), *“Trust in the chatbot: a semi-human relationship”*, <https://link.springer.com/article/10.1186/s43093-023-00288-z> [accessed 15 Feb 2024]
17. Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020), *“Trust Toward Robots and Artificial Intelligence: An Experimental Approach to Human–Technology Interactions Online”*, <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.568256/full> [accessed 4 Nov 2023]
18. OpenAI (2023), *“Introducing ChatGPT Enterprise”*, [https://openai.com/blog/introducing-chatgpt-enterprise#OpenAI](https://openai.com/blog/introducing-chatgpt-enterprise%23OpenAI) [accessed 15 Feb 2024]
19. Pinto dos Santos, D., Giese, D., Brodehl, S., Chon, S. H., Staab, W., Kleinert, R., Maintz, D., & Baeßler, B. (2018), *“Medical students' attitude towards artificial intelligence: a multicentre survey”*, <https://link.springer.com/article/10.1007/s00330-018-5601-1> [accessed 1 Nov 2023]
20. Resume Builder (2023), *“1 in 4 companies have already replaced workers with ChatGPT”*, <https://www.resumebuilder.com/1-in-4-companies-have-already-replaced-workers-with-chatgpt/> [accessed 16 Feb 2024]
21. Rezlescu, C., Danaila, I., Miron, A., & Amariei, C. (2020), *“Chapter 13 - More time for science: Using Testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research”*, <https://www.sciencedirect.com/science/article/abs/pii/S0079612320300807> [accessed 27 Dec 2023]
22. Schepman, A., & Rodway, P. (2020), *“Initial validation of the general attitudes towards Artificial Intelligence Scale”*, <https://www.sciencedirect.com/science/article/pii/S2451958820300142> [accessed 25 Oct 2023]
23. Schepman, A., & Rodway, P. (2021), *“The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust”*, <https://www.tandfonline.com/doi/full/10.1080/10447318.2022.2085400> [accessed 25 Oct 2023]
24. Schniter, E., & Shields, T.W. (2020), *“Gender, Stereotypes, and Trust in Communication”*, [https://link.springer.com/article/10.1007/s12110-020-09376-3#citeas](https://link.springer.com/article/10.1007/s12110-020-09376-3%23citeas) [accessed 12 Nov 2023]
25. Sharan, N. N., & Romano, D. M. (2020), *“The effects of personality and locus of control on trust in humans versus artificial intelligence”*, <https://www.cell.com/heliyon/pdf/S2405-8440(20)31416-X.pdf> [accessed 26 Oct 2023]
26. SimilarWeb (2023), *“chat.openai.com Web Traffic”,* <https://www.similarweb.com/website/chat.openai.com/> [accessed 18 Feb 2024]
27. Soto, C.J., & John, O.P. (2017), *“Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS”*, <https://www.sciencedirect.com/science/article/abs/pii/S0092656616301325> [accessed 24 Oct 2023]
28. Soto, C.J., & John, O.P. (2017), *“The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power”*, <https://pubmed.ncbi.nlm.nih.gov/27055049/> [accessed 24 Oct 2023]
29. The British Psychological Society (2023), *“The value of mental health chatbots”*, <https://www.bps.org.uk/blog/value-mental-health-chatbots> [accessed 15 Feb 2024]
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017), *“Attention Is All You Need”*, <https://arxiv.org/pdf/1706.03762.pdf> [accessed 20 Feb 2024]
31. Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K., & Peng, N. (2023), *“Kelly is a Warm Person, Joseph is a Role Model: Gender Biases in LLM-Generated Reference Letters”*, <https://arxiv.org/pdf/2310.09219.pdf> [accessed 18 Nov 2023]

## Appendices

#### Appendix A – Study Questions

#### Appendix B – Application

#### Appendix C – Reformat Results.py

#### Appendix D – Evaluate Results.py

#### Appendix E – Machine Learning.py