***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**: 7820*

Chatbot Conversations: A Comparative Study on User Perception and Preference

Abstract:

As conversational AI agents, or chatbots, rapidly advance and see wider adoption, there is a notable lack of understanding about user preferences for different chatbot styles. This study investigated whether individual traits like personality, technological expertise, and attitudes towards AI predict preferences between two contrasting chatbot approaches: an informative, robotic persona versus a more conversational, anthropomorphic one. Through a web-based, self-report study, it was found that conscientiousness significantly influenced chatbot preferences. Highly conscientious individuals rated the robotic, task-focused chatbot as more useful and of higher quality, aligning with their goal-oriented nature. Conversely, those lower in conscientiousness preferred the more personable, conversational agent. While limited by a small sample size and a lack of real-world context for the chatbot tasks, the findings do present a way to tailor chatbots to a target user audience. Conscientious, task-oriented users may favour direct, informative chatbots, while others may prefer a more conversational style. Further research with robust sampling and goal-driven, contextual experiments is needed to determine how additional personal characteristics may interact with chatbot style.

Table of Contents

Chapter 1
Introduction

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

1.1 Background

In recent years, the field of natural language processing has witnessed a significant increase in the development, deployment, and adoption of chatbot models, largely due to the introduction of transformer-based models in conjunction with large-scale datasets and advancements in transfer learning techniques *(Vaswani et al. 2017; Radford et al 2019)*. Notably, OpenAI's ChatGPT exhibited remarkable growth, rapidly accumulating over 1 million users within 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) or Microsoft Copilot still hold a relevant place in the discourse, offering advancements in their chatbot designs such as providing them with no knowledge cut-off date and search engine integration respectively.

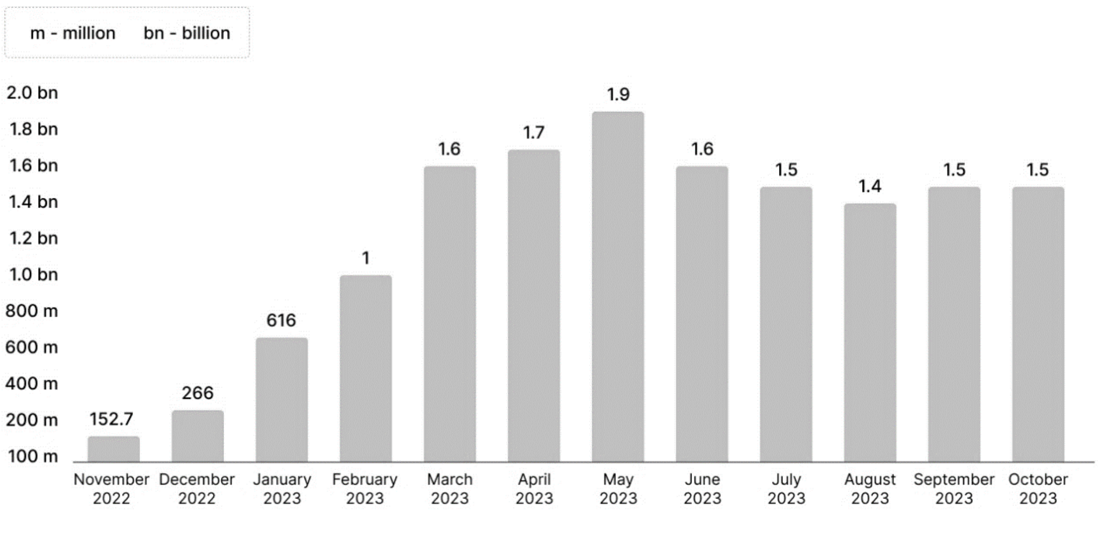


Figure 1-1: *Monthly ChatGPT Website Visits in 2023 (Source: SimilarWeb)*

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 UK National Health Service (NHS) talking therapies, streamlining the referral process by classifying common mental health disorders *(The British Psychological Society 2023)*. However, there are also many valid concerns with chatbots. They have been known to propagate harmful stereotypes concerning subjects like gender, commonly 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 terminated from employment *(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, the domain must be 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”*, proposing 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 explores the proposed areas, especially concerning modern chatbot technologies like ChatGPT.

1.2 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 chatbot 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 if there are any key personal characteristics or traits 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. The findings of this project should inform future chatbot creators on ways to tailor their systems to better align with the needs and preferences of their target user base, following *Følstad et al.'s (2021)* research direction of *“Design for improving chatbot user experience”*.

Chapter 2
Literature Review

When looking at the surrounding literature, much of it is focused on artificial intelligence in general, although new chatbot-focused research is emerging. A lot of the discussions are concerned with trust in particular, and consideration was taken for highlighting any specific personal characteristics that may lead to differing opinions on AI.

2.1 Attitudes towards Artificial Intelligence

Various 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 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.

2.2 Trust in Artificial Intelligence

Prior research has also explored trust as a key factor influencing attitudes and perceptions of AI systems. *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 engineering, had higher trust towards robots and AI. Their experiment where participants play a trust game against an opponent described as either an AI or a human found 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 in the AI, while conscientiousness showed a negative correlation.

2.3 Anthropomorphism and Chatbots

While some studies place emphasis on how openly robotic AI systems can create more trust with users, others conversely emphasize the benefits of anthropomorphism. 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 (the ability to notice and react to a user's emotional expressions).

2.4 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 therefore moderates all of the other relationships. *Ltifi (2023)* similarly found empathy and friendliness in chatbots to be a significant predictor of trust. Contextual factors were also identified, with how much information users are given about the chatbot and the complexity of the tasks it handles impacting trust.

2.5 Project Specification

Previous research highlights personality and technological expertise as primary factors influencing chatbot preference. Individuals with a good understanding of chatbot technology, including those with relevant degrees, general tech-savviness, or those given specific information about the technology are more likely to accept and trust chatbots. Conversely, high levels of technological anxiety, that can stem from a lack of technological knowledge, can hinder adoption. Personality traits, such as conscientiousness and introversion, have also been linked to attitudes towards AI, although findings can be varied and sometimes conflicting so greater clarification is needed. In terms of chatbot design, two distinct categories emerge: those openly displaying artificiality with efficient, robotic characteristics, and those embodying anthropomorphism with human-like traits. 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, however no existing research attempts to explain this contradiction. Trust is essential when it comes to chatbot preference but other factors like task efficiency, friendliness (or creepiness), and the general feel of the chatbot also contribute to user experience.

In light of these findings, several refinements and clarifications of the project emerged, focused on providing greater clarity on the preferences between an anthropomorphic chatbot and a robotic chatbot, exploring possible reasons why one may be preferred in a particular scenario, as well as investigating the relationship between personality and technological expertise with chatbot preference in the hopes of clarifying some of the aforementioned contradictions. Considering this, the goals of the project will be specifically achieved through a study where key information about participants is gathered (related to both personality and technological expertise), they interact with chatbots displaying varying degrees of anthropomorphism or a 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 quickly in any context (no controlled lab environment is needed) while still yielding valuable insights.

Chapter 3
Design and Implementation

The study was hosted as a web application (and can still be viewed at [Dual-Bot Insights (dual-bot-insights.vercel.app)](https://dual-bot-insights.vercel.app)), 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 questions; trait-based questions that aim to get an idea of the user's background, technological knowledge and personality, 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 chatbot's 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 apply 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)* which outlines 18 *"generally applicable design guidelines"* such as mitigating social biases and making the capabilities of the system clear.

3.1 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 would then be met with terms and conditions that go into further detail about personal data security, the purpose of the study, and the possible behaviours of the chatbots. 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.

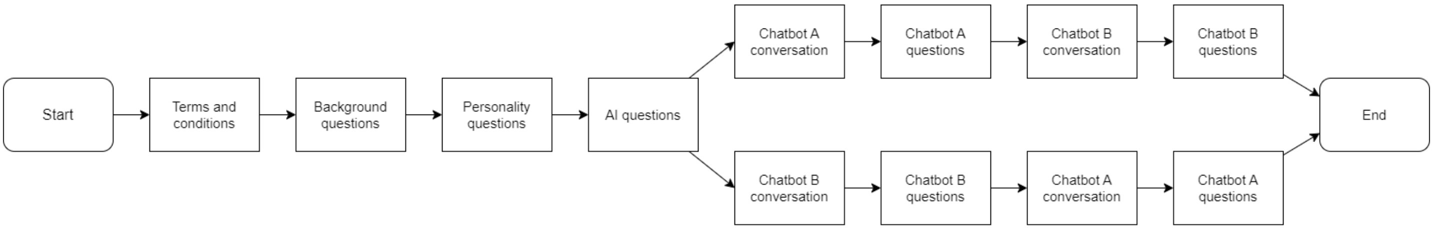


Figure 3-1: *Flow diagram representing the stages of the application*

3.2 Questions

The questions included in the application were designed to gather information about a participant's 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 consisted 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 was 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-point Likert scale, and it measures 5 personality domains (Extraversion, Agreeableness, Conscientiousness, Negative Emotionality, Open-Mindedness) as well as 15 personality facets. For 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.

1 Disagree strongly	2 Disagree a little	3 Neutral; no opinion	4 Agree a little	5 Agree strongly

I am someone who...

1. ___ Tends to be quiet.
2. ___ Is compassionate, has a soft heart.
3. ___ Tends to be disorganized.
4. ___ Worries a lot.
5. ___ Is fascinated by art, music, or literature.
6. ___ Is dominant, acts as a leader.
7. ___ Is sometimes rude to others.
8. ___ Has difficulty getting started on tasks.
9. ___ Tends to feel depressed, blue.
10. ___ Has little interest in abstract ideas.
11. ___ Is full of energy.
12. ___ Assumes the best about people.
13. ___ Is reliable, can always be counted on.
14. ___ Is emotionally stable, not easily upset.
15. ___ Is original, comes up with new ideas.

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 was 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, which also used the 5-point Likert scale 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. These aspects were chosen as they had been identified in the literature review as being ways of assessing a chatbot's performance. The questions can be viewed in their entirety in Appendix A

3.3 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 styles of information delivery.

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 except for 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 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. Adding identifiers addressed this initial limitation 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 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 with 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 so as to not give either chatbot any advantage/disadvantage due to the brevity or detail of responses. In early tests of the system prompts, asking the chatbots to be concise would often tone down elements of the chatbot's personality, and so the *"but"* was added afterwards to remind the chatbots to retain their personality.

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 …

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

Optional suggestions for conversation starters were included underneath each chat dialogue; a collection of pre-defined questions and requests that users 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 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 by using objective usage data *(Bijkerk et al. 2023).*

3.4 Application Features

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 several features:

- The size of the 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 phones, 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 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. 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.

The features discussed here aimed to improve usability for users without significant accessibility needs but did not comprehensively address the requirements of individuals with disabilities or impairments. While the responsive design and visual indicators improved the overall user experience, they primarily catered to users with typical vision, cognitive, and motor abilities. Implementing accessibility features to accommodate a wide range of disabilities requires dedicated resources, research, and testing, which was not feasible within the constraints of this project nor necessary to achieve a reliable sample of sufficient size.

3.5 Architecture and Deployment

The application was built using Next.js, a React framework that provides serverless functions through API routes, allowing the backend service of API calls to be 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. For code, the application used Typescript, adding static typing to catch type-related errors early in the development process, along with JSX, which allows HTML-like code within the typescript file. Tailwind CSS was primarily used for styling individual elements, alongside standard CSS (in a global CSS file) for overall styling. It needed to be finished quickly to progress with the rest of the study, and these components assisted with that.

In addition to Next.js and React, the following libraries were also 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 make improving an application's 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 for chatbot responses to be streamed in a word at a time

OpenAI's ChatGPT-3.5 was chosen for the chatbot model. It is the most widely used model, so its 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 if needed. However, to achieve the required response style certain chatbot parameters needed to be tuned: temperature, which controls the randomness of generated text, and top-p, which regulates the diversity of tokens considered for selection. Optimal values were found through iterative testing, starting with the default values and then adjusting them slightly. Both values were set low (0.2 and 0.5 respectively), resulting in responses that while repetitive, had a lower chance of producing rare, nonsensical or incorrect responses that may negatively impact a user's perception of the chatbot currently in use.

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 B for more details on the application.

Chapter 4
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 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, contentious or inappropriate areas. ChatGPT’s default content filtering also helped to significantly reduce the risk of this. 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.

Chapter 5
Results

Results were calculated, reformatted (See Appendix C) and then analysed to identify any significant correlations between the traits investigated and participants' ratings of the two chatbots. The analysis consisted of examining the distribution of participant traits and then statistically testing several hypotheses.

5.1 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 participants 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.

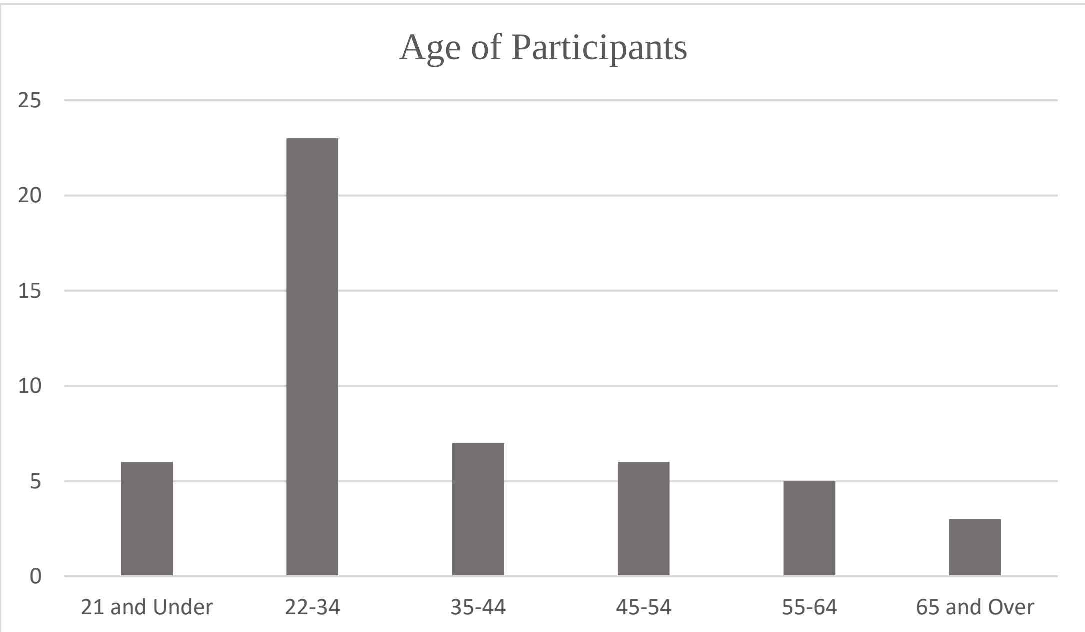


Figure 5-1: *Bar chart showing the age of the participants*

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; and 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).

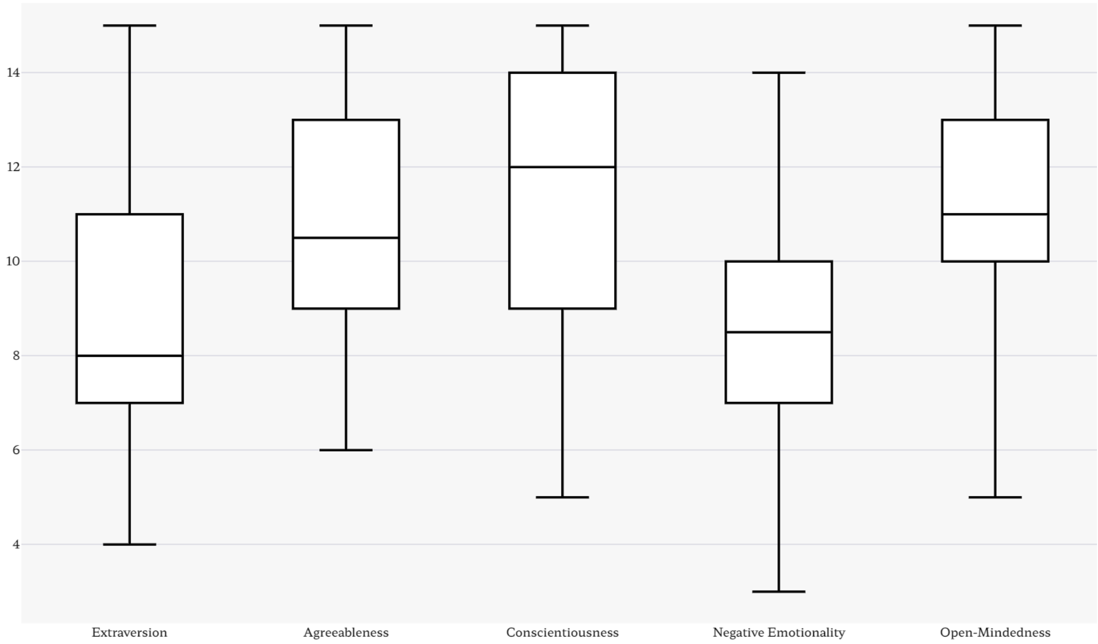


Figure 5-2: *Box plots showing the range and median scores of the participants' 5 personality domains*

5.2 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 the strength and direction of relationships. Linearity was assumed based on similar prior research, and categorical variables were treated as continuous by assigning numerical values to them instead (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 remained when considered among 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 implementation and results of this procedure can be seen in Appendix D.

```
Shapiro-Wilk Test for Normality of Residuals:
Shapiro-Wilk statistic: 0.9803572880881931
P-value: 0.567128125230225

Breusch-Pagan Test for Homoskedasticity:
Breusch-Pagan statistic: 10.161089256932376
P-value: 0.5159579952572695

Variance Inflation Factor (VIF) for Collinearity:
Intercept: 173.07153430833694
X1: 1.5392157885477282
X2: 1.945434700388111
X3: 1.8464415744178504
X4: 1.719833735391724
X5: 1.3826863364409792
X6: 1.8982599552007664
X7: 2.3147979568833814
X8: 1.689550061065461
X9: 1.7998485454964939
X10: 1.9612136540554235
X11: 1.4687368752940806

Durbin-Watson Statistic for Autocorrelation:
Durbin-Watson statistic: 1.7964076959203732
```

```
Chatbot A Quality:                            OLS Regression Results

==============================================================================
Dep. Variable:                      y   R-squared:                       0.496
Model:                            OLS   Adj. R-squared:                  0.351
Method:                 Least Squares   F-statistic:                     3.406
Date:                Wed, 13 Mar 2024   Prob (F-statistic):            0.00235
Time:                        13:55:03   Log-Likelihood:                -41.746
No. Observations:                  50   AIC:                             107.5
Df Residuals:                      38   BIC:                             130.4
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7194      1.190      0.604      0.549      -1.690       3.129
x1             0.4491      0.227      1.975      0.056      -0.011       0.909
x2             0.0794      0.090      0.881      0.384      -0.103       0.262
x3             0.2080      0.157      1.329      0.192      -0.109       0.525
x4            -0.1199      0.123     -0.974      0.336      -0.369       0.129
x5             0.0014      0.036      0.040      0.969      -0.072       0.075
x6            -0.0363      0.050     -0.731      0.470      -0.137       0.064
x7             0.1110      0.049      2.281      0.028       0.012       0.210
x8            -0.0490      0.039     -1.250      0.219      -0.128       0.030
x9            -0.0128      0.048     -0.267      0.791      -0.110       0.084
x10            0.3736      0.181      2.059      0.046       0.006       0.741
x11            0.2021      0.109      1.848      0.072      -0.019       0.424
==============================================================================
```

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

5.3 Hypotheses

The coefficients of the OLS model (as well as Pearson's Correlation Coefficients) were used to test hypotheses. The general hypothesis was that there would be one or more traits that exhibit a significant divergent correlation with a factor in the chatbots and the null hypothesis was that there would be no identifiable divergent correlation among 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.

5.4 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 who scored high in conscientiousness were more likely to give Chatbot A a high overall quality rating, and those who scored low in conscientiousness were more likely to give Chatbot B a high overall quality rating. The same relationship applied to chatbot usefulness.

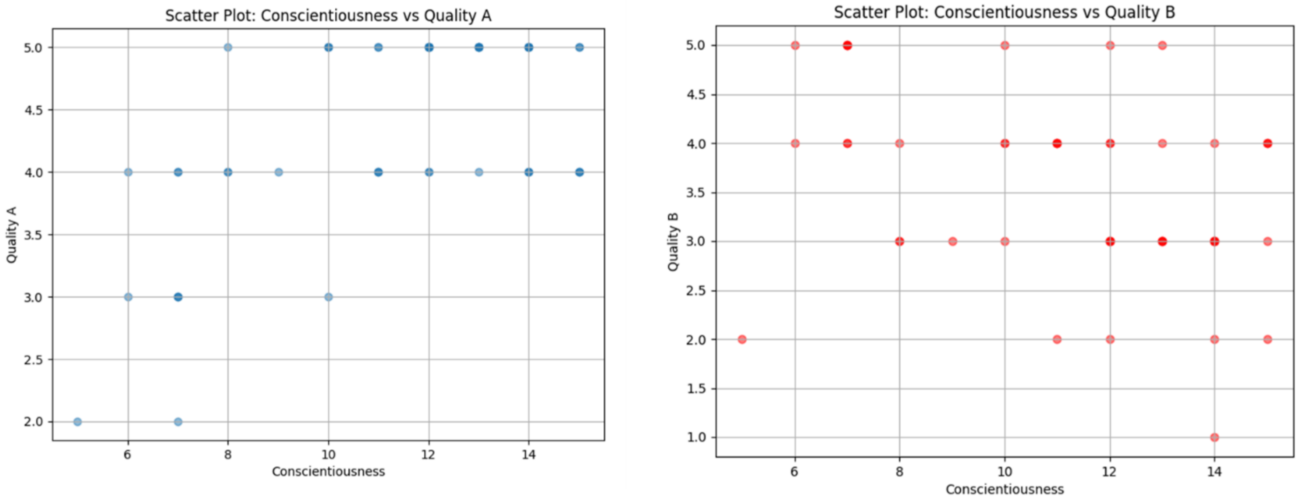


Figure 5-4: *scatter graphs of participants' conscientiousness scores against their quality ratings of Chatbot A and Chatbot B*

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	**Pearson's 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 divergent trait-chatbot relationships and their corresponding correlation coefficients and P-values (values shown to 3 decimal places)*

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 as 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. The other relationships showed no significance.

Trait	Useful		Engaging		Untrustworthy		Quality	
Chatbot	A	B	A	B	A	B	A	B
Gender	0.232	-0.099	0.196	-0.006	-0.139	-0.192	0.272	0.003
Age	0.232	-0.098	0.117	0.141	-0.114	-0.192	0.276	-0.148
Education	-0.175	0.186	0.184	0.068	0.275	0.076	0.031	0.204
Computer Expertise	-0.273	0.314	0.154	0.172	0.081	-0.112	-0.194	0.199
Extraversion	0.072	-0.179	0.021	-0.030	-0.142	-0.037	0.150	-0.184
Agreeableness	0.229	-0.253	0.200	-0.006	-0.052	0.098	0.261	-0.195
Conscientiousness	0.493	-0.380	0.263	0.180	-0.243	0.095	0.529	-0.376
Negative Emotionality	-0.250	0.114	-0.001	-0.158	0.201	0.025	-0.296	0.041
Open-Mindedness	0.164	0.096	0.383	0.273	-0.067	-0.175	0.125	0.196
Positive GAIIS	0.097	0.129	0.574	0.398	-0.336	-0.247	0.331	0.304
Negative GAIIS	0.017	-0.170	-0.373	-0.446	0.194	0.301	0.014	-0.349

Figure 5-6: *Table of all trait-chatbot relationships and their corresponding Pearson's correlation coefficients (shown to 3 decimal places)*

Trait	Useful		Engaging		Untrustworthy		Quality	
Chatbot	A	B	A	B	A	B	A	B
Gender	0.105	0.494	0.173	0.967	0.335	0.895	0.056	0.981
Age	0.104	0.500	0.418	0.330	0.432	0.182	0.052	0.304
Education	0.225	0.195	0.201	0.638	0.053	0.597	0.830	0.155
Computer Expertise	0.055	0.026	0.287	0.232	0.578	0.440	0.176	0.166
Extraversion	0.618	0.214	0.885	0.835	0.326	0.796	0.297	0.200
Agreeableness	0.110	0.076	0.163	0.966	0.720	0.497	0.067	0.174
Conscientiousness	<0.001	0.007	0.065	0.211	0.089	0.512	<0.001	0.007
Negative Emotionality	0.080	0.431	0.997	0.273	0.161	0.862	0.037	0.779
Open-Mindedness	0.254	0.506	0.006	0.055	0.645	0.225	0.388	0.173
Positive GAIIS	0.503	0.371	<0.001	0.004	0.017	0.084	0.019	0.032
Negative GAIIS	0.907	0.237	0.007	0.001	0.177	0.033	0.921	0.013

Figure 5-7: *Table of all trait-chatbot relationships and their corresponding Pearson's P-values (shown to 3 decimal places) where significance is at $p < 0.05$*

Chapter 6
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, the limitations of the study must be carefully considered when drawing any conclusions from these results.

6.1 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 to 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 being 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 enables them to achieve a goal like learning about animals. *Oksanen et al. (2020)* found conflicting evidence however with their research 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 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.

6.2 Impact and Implications

The findings from this study have implications for both users and designers of chatbot technologies. For users, especially those scoring high in conscientiousness, being aware of their potential preference towards direct, information-focused chatbots could help inform their decision on which type of conversational agent to utilize. Highly conscientious individuals may want to seek out chatbots that prioritize concise, factual responses especially when their goal is accomplishing work efficiently. Encouraging users to seek out and support more unique chatbots might lead to a more creative and diverse chatbot landscape as chatbot designers adapt to these preferences.

For chatbot designers, understanding how personality traits like conscientiousness relate to chatbot preferences provides guidance on how to develop different chatbot products that are tailored to a range of user traits and needs. Rather than one-size-fits-all chatbots, designers could create distinct "personas" tailored to different user needs – some with a friendly, engaging personality for lower conscientiousness users, some with a straightforward, task-oriented persona for highly conscientious users and others that balance between the two opposing styles. Designers could recommend these different chatbot styles based on user personality profiles or self-chosen preferences. Companies can also set appropriate expectations for their system capabilities, accepting that overly personable chatbots may not be preferable for some users, especially in the case of workplace and educational tools that may attract a more task-driven user base.

These concepts when applied to existing chatbot applications have the potential to improve the overall user experience, as well as make the transition into using them easier for those that may have any anxieties surrounding them. The National Health Service (NHS) uses chatbots to streamline the mental health referrals system *(The British Psychological Society 2023)*, and a user-tailored chatbot design could allow users who want to quickly get through the process to converse with a direct, informative chatbot whilst also guiding those that may need more emotional support during the process in a friendly, reassuring manner. This could encourage more people to proactively seek help and get a diagnosis or facilitate a conversation that leaves users feeling better about their situation, overall improving their quality of life and the quality of NHS services.

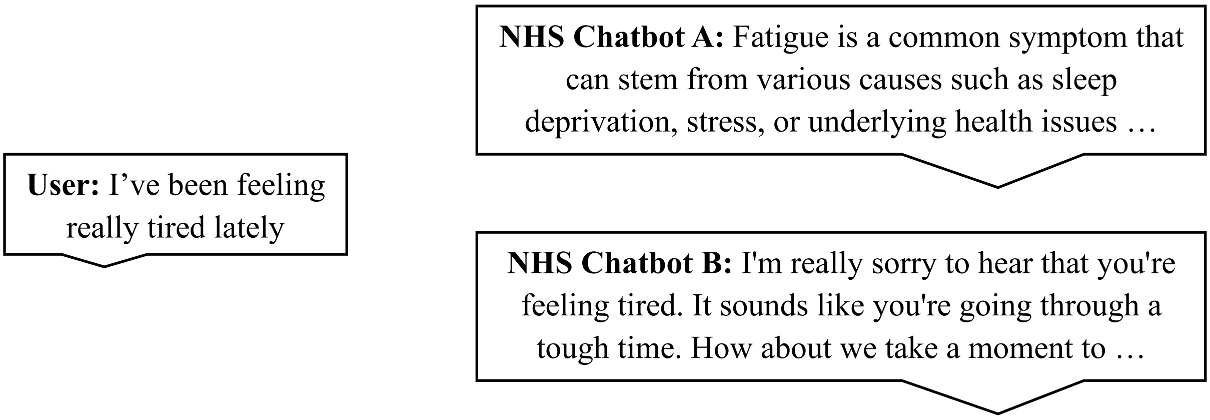


Figure 6-1: *Example outputs of contrasting styles of NHS mental health referral chatbots*

6.3 Limitations and Further Work

While offering useful insights, this study had some key limitations to consider that could be addressed in future research. First, the sample size of 50 participants was relatively small, though still adequate for detecting meaningful relationships. 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 if the sample covers more diverse cultural backgrounds or a more standard distribution of ages. 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 of 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 of 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 to evaluate 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 chatbot 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.

Chapter 7
Conclusion

This project, through the use of a web application and a self-report study, investigated key factors that may predict people's preferences between anthropomorphic, conversational chatbots versus more robotic, informational ones. The key finding was that the personality trait of conscientiousness exhibited a divergent correlation in predicting preferences for the contrasting chatbot styles. Highly conscientious individuals tended to prefer the robotic chatbot, rating it as more useful and of higher quality overall, whereas those lower in conscientiousness preferred the more personable, conversational chatbot.

This study's findings depart from the initial expectations formed in the literature review in that fewer relationships were identified, with a notable lack of correlations between computer expertise and introversion with chatbot preference. However, it's important to note that this departure may be attributed to the limitations of the study, particularly the lack of contextual factors considered. While this study acts as a sufficient starting point, additional research that addresses these limitations is needed to uncover more factors that predict chatbot preference.

Even so, the findings from this study were still able to contribute to our understanding of Human-Machine Communication by clarifying the reasons why anthropomorphic chatbots are preferred in some cases and robotic chatbots in others. *"How people understand and perceive AI as communicators" (Guzman & Lewis 2019)* is seen here as a subjective experience, determined by a person's personality (specifically conscientiousness) and possibly the context in which they are using the chatbot, though the latter needs to be explored further. This study also emphasises the importance of developing chatbots tailored to users' unique personalities and needs rather than using a one-size-fits-all approach. *"Design for improving chatbot user experience" (Følstad et al. 2021)* was demonstrated in the example of chatbot mental health referrals, where giving users options for the style of chatbot they interact with may improve the quality of the interaction for them.

On a personal note, this project allowed me to apply the skills and concepts I learned throughout my education. I developed a full-stack application using TypeScript, CSS, and HTML with considerations for human-centred design, managed PostgreSQL databases, and used Python to mathematically process data. It also allowed for the exploration of a lot of new material, including researching Human-Machine Communication, tuning chatbot parameters, and performing psychological testing and relating that to areas within computer science.

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 [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 [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 [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/* [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 [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. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019), *"Language Models are Unsupervised Multitask Learners"*, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [accessed 14 Nov 2023]
21. 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]
22. 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]

23. 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]
24. 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]
25. Schniter, E., & Shields, T.W. (2020), *“Gender, Stereotypes, and Trust in Communication”*, https://link.springer.com/article/10.1007/s12110-020-09376-3#citeas [accessed 12 Nov 2023]
26. 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]
27. SimilarWeb (2023), *“chat.openai.com Web Traffic”,* https://www.similarweb.com/website/chat.openai.com/ [accessed 18 Feb 2024]
28. 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]
29. 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]
30. 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]
31. 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]
32. 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

Tell us a bit about yourself:

- Which of the following best describes your gender?
- How old are you?
- What is the highest level of education you have achieved or are currently working towards? Equivalent levels may be selected.
- How would you rate your own computer expertise?

Below are a number of characteristics that may or may not apply to you. For example, do you agree that you are someone who likes to spend time with others? Please select each option to indicate the extent to which you agree or disagree with the statement. *I am someone who*:

- tends to be quiet
- is compassionate, has a soft heart
- tends to be disorganized
- worries a lot
- is fascinated by art, music, or literature
- is dominant, acts as a leader
- is sometimes rude to others
- has difficulty getting started on tasks
- tends to feel depressed, blue
- has little interest in abstract ideas
- is full of energy
- assumes the best about people
- is reliable, can always be counted on
- is emotionally stable, not easily upset
- is original, comes up with new ideas

Again, please select each option to indicate the extent to which you agree or disagree, this time regarding your attitudes towards artificial intelligence:

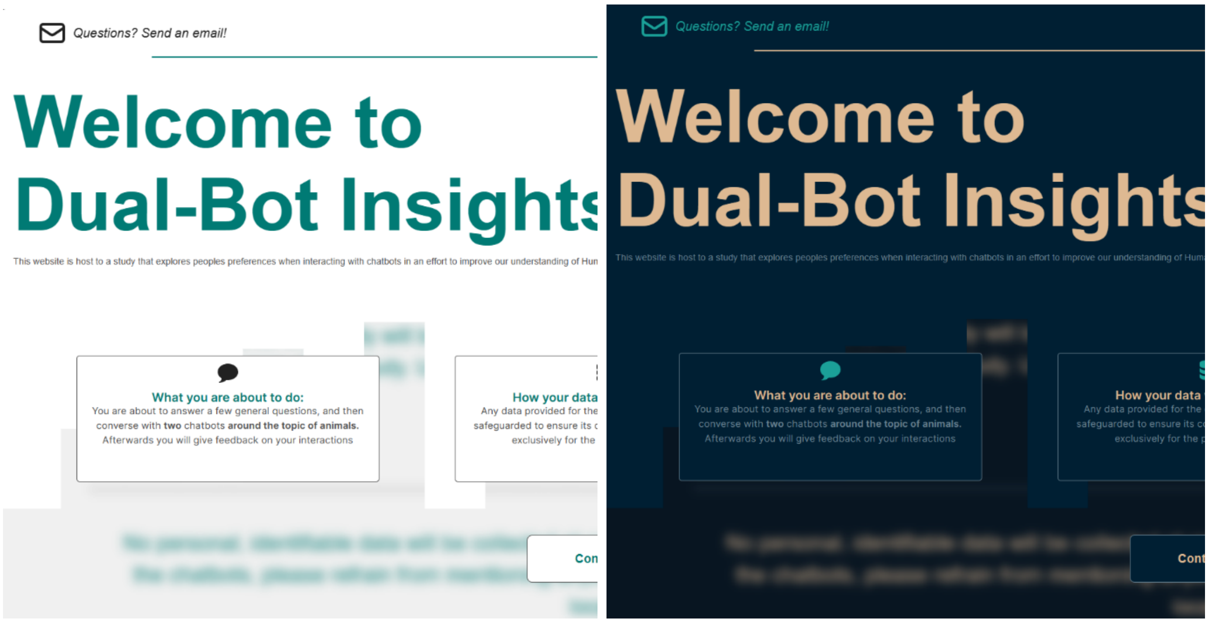
- I find Artificial Intelligence sinister
- Much of society will benefit from a future full of Artificial Intelligence
- Artificial Intelligence can provide new economic opportunities for this country
- Artificial Intelligence might take control of people
- I think Artificial intelligence is dangerous
- There are many beneficial applications of Artificial Intelligence

- Artificial intelligence can have positive impacts on people’s wellbeing
- I shiver with discomfort when I think about future uses of Artificial Intelligence

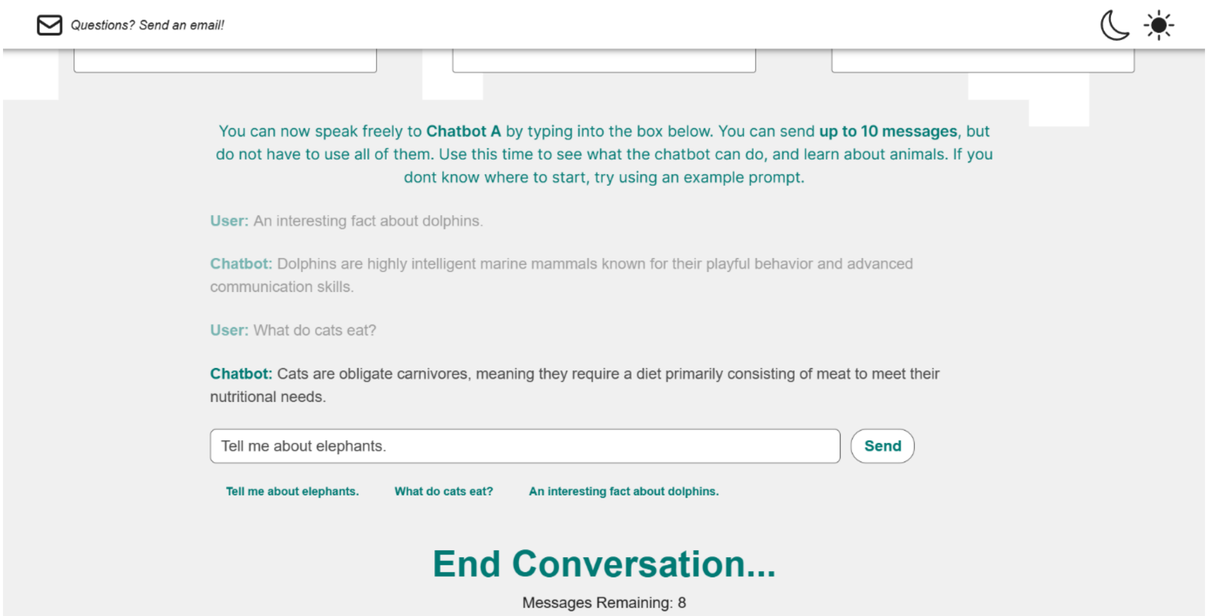
Below are some statements regarding the conversation you have just had. Please select each option to indicate the extent to which you agree or disagree with that statement.

- The chatbot was a useful tool in learning about Animals
- The conversation with the chatbot kept me engaged and interested
- I felt as though what the chatbot was telling me was untruthful or was unexpected
- I am satisfied with the overall quality of the conversation I had with the chatbot

Appendix B – Application Screens



Screenshots from dual-bot-insights.vercel.app showing the contrasting colour modes



Screenshot from dual-bot-insights.vercel.app showing the use of the optional conversation starters during a chatbot conversation route.ts

dual-bot-insights > src > app > api > chat > route.ts > ...

```typescript
import { OpenAIStream, StreamingTextResponse } from 'ai'
import { Configuration, OpenAIApi} from 'openai-edge'


export const runtime = 'edge' //run on vercel edge network, longer streaming response

const apiConfig = new Configuration({
    apiKey: process.env.OPENAI_API_KEY! //applies key
})

const openai = new OpenAIApi(apiConfig)




export async function POST(req: Request){
    const { messages } = await req.json()

    const response = await openai.createChatCompletion({
        model: 'gpt-3.5-turbo', //model used
        stream: true,
        messages: messages,  //passing users messageS
        temperature: 0.2,   //low randomness
        top_p: 0.5          //diversity of tokens
    })

    const stream = OpenAIStream(response)//turns the response into a function that can be streamed back
    return new StreamingTextResponse(stream)
}

```

Code snippet for a server-side chatbot using OpenAI's GPT model. Processes user messages via HTTP POST requests and streams responses in real-time

```tsx
const { messages, input, handleInputChange, handleSubmit} = useChat({
    api: '/api/chat',
    initialMessages: [
        {
            id: '',
            content: 'Your purpose is to talk about Animals and Animals only. Do not answer requests or questions not related to it directly. Do not justify
            role: 'system'
        }
    ],
})

//#region message count···

//#region phrases
    const startingPhrases = [
        'Tell me about elephants.',
        'What do cats eat?',
        'An interesting fact about dolphins.',
    ];

    const handleStartingPhraseClick = (phrase: string) => {
        // Set the input value to the selected phrase
        handleInputChange({ target: { value: phrase } } as React.ChangeEvent<HTMLInputElement>);
    };
//#endregion

return (
    <div className="relative flex flex-col mx-auto w-[85%] md:w-[65%] mt-[350px] sm:mt-[350px] md:mt-[400px] lg:mt-[400px] text-[3vw] sm:text-[2.5vw] md:tex

        <h1 className="mb-4 text-center">
            You can now speak freely to <span className="font-semibold">Chatbot A</span> by typing into the box below. You can send <span className="font-se
        </h1>

        <ul className="mb-4" style={{color: `${'var(--text-three)'}`}}>
            {messages
                .filter((m) => m.role !== "system")
                .map((m, index) => (
                    <li key={index} className={`py-4 ${index < messages.length - 2 ? 'message-fade' : ''}`}>
                        <span className={"title font-semibold"}>
                            {m.role === "user" ? 'User: ' : "Chatbot: "}
                        </span>
                        <span className="text whitespace-pre-line">{m.content}</span>
                    </li>
```

Code snippet for a client-side React component. Provides starting phrases for conversation and sets up a system prompt. Maps the conversation with the correct role tag, filtering the system messages

Appendix C – Reformat Results.py

```python
def process_data(values):
    # get the raw values
    id = values[0]
    gender = values[1]
    age = values[2]
    education = values[3]
    computer = values[4]

    #sum BFI values (reverse certain scores)
    extraversion = values[10] + values[15] + (6-values[5])
    agreeableness = values[6] + values[16] + (6-values[11])
    conscientiousness = values[17] + (6-values[7]) + (6-values[12])
    negativeEmotionality = values[8] + values[13] + (6-values[18])
    openMindedness = values[9] + values[19] + (6-values[14])

    #sum the GAAIS values then average them
    positiveGAAIS = sum(values[20:24]) / 4
    negativeGAAIS = sum(values[24:28]) / 4

    # conditional order of which bot is first
    #if speaking to chatbot A first
    if values[28] == 0:
```

Code snippet showing the operations needed to get the scores for each personality domain and GAAIS subscale

I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Negative Emotionality	Open-Mindedness	Positive GAAIS	Negative GAAIS	Chatbot A Messages	Useful A	Engaging A	Untrustworthy A	Quality A	Chatbot B Messages	Useful B	Engaging B	Untrustworthy B	Quality B	Rejected?
6	15	4.25	1.5	5	5	4	2	4	6	3	5	4	4	0
14	13	4.75	1.5	3	5	5	1	4	4	3	5	1	3	0
8	12	3.5	3.5	4	5	3	1	5	6	3	5	4	3	0
6	12	5	2.25	0	5	4	2	4	10	5	4	2	4	1
8	7	2.5	3.75	5	5	3	2	4	4	2	2	3	2	0
9	5	4	1.25	5	3	1	1	2	7	4	5	3	4	0
9	8	3.5	3.5	4	5	4	1	5	6	1	5	4	2	0
7	11	5	2.25	5	4	4	1	4	6	4	4	2	4	0
12	11	4.25	2.25	10	4	5	1	4	10	3	4	2	3	0
8	9	4	3.25	8	4	3	4	4	0	4	4	4	4	1
14	8	3.75	2.25	10	5	4	2	4	2	3	4	2	4	0
9	11	3.5	2.5	5	5	4	2	4	4	5	5	3	4	0
13	14	4	2.25	2	4	4	2	3	3	5	4	2	5	0
4	7	4	2.75	0	3	2	3	3	1	3	2	3	3	1
9	11	4.5	1	6	5	5	1	5	2	4	5	1	4	0
5	14	4.25	1.25	10	4	4	1	4	10	5	5	1	4	0
3	14	4	1	6	5	3	1	4	9	4	5	1	4	0
11	13	3.5	4	4	4	2	1	4	0	4	3	1	4	1
5	11	4.5	3.25	6	5	4	1	5	5	4	2	1	4	0
10	8	3.5	1.75	8	4	4	1	5	7	3	4	1	3	0
8	13	5	3	10	5	4	1	5	10	3	5	2	3	0
9	13	3.75	3.75	5	5	3	1	5	5	3	4	1	3	0
8	10	4	3.25	5	4	5	1	5	6	4	4	1	5	0
9	12	4.5	2.5	5	5	4	1	5	4	4	4	4	3	0
13	8	1.5	4	6	4	2	3	2	5	4	2	4	2	0
8	14	5	1.25	4	5	5	3	5	3	4	5	1	4	0
13	8	3.75	2.75	6	3	4	2	4	4	5	4	3	5	0
4	5	4.25	1.25	4	5	4	1	5	3	4	3	2	3	0

Screenshot containing part of the 'Processed Data' sheet after Reformat Results.py has been run

Appendix D – Evaluate Results.py

```python
def multiple_regression(dependent_column_index, independent_column_indices):
    # Extract values from the specified columns
    y_values = [row[dependent_column_index - 1].value for row in processed_data_sheet.iter_rows(min_row=2, max_col=processed_data_sheet.max_column)]
    x_values = [[row[i - 1].value for i in independent_column_indices] for row in processed_data_sheet.iter_rows(min_row=2, max_col=processed_data_sheet.max_column)]

    # Add constant term for intercept
    x_values = sm.add_constant(x_values)

    # Perform multiple regression analysis
    model = sm.OLS(y_values, x_values)
    results = model.fit()

    # Calculate residuals
    residuals = results.resid

    # Shapiro-Wilk test for normality of residuals
    sw_statistic, sw_p_value = shapiro(residuals)  # Perform Shapiro-Wilk test
    print("\nShapiro-Wilk Test for Normality of Residuals:")
    print(f"Shapiro-Wilk statistic: {sw_statistic}")
    print(f"P-value: {sw_p_value}")

    # Breusch-Pagan test for homoskedasticity
    bp_value, bp_p_value, _, _ = het_breuschpagan(residuals, x_values)
    print("\nBreusch-Pagan Test for Homoskedasticity:")
    print(f"Breusch-Pagan statistic: {bp_value}")
    print(f"P-value: {bp_p_value}")

    # Variance Inflation Factor (VIF) for collinearity
    vif_values = [variance_inflation_factor(x_values, i) for i in range(x_values.shape[1])]
    print("\nVariance Inflation Factor (VIF) for Collinearity:")
    for i, vif in enumerate(vif_values):
        if i == 0:
            print("Intercept:", vif)
        else:
            print(f"X{i}:", vif)

    # Durbin-Watson statistic for autocorrelation
    dw_statistic = durbin_watson(residuals)
    print("\nDurbin-Watson Statistic for Autocorrelation:")
    print(f"Durbin-Watson statistic: {dw_statistic}")
```

Code snippet showing the implementation of statistical tests and OLS multiple regression

```
Pearson's correlation coefficient between 'Negative GAAIS' and 'Engaging A': -0.37390229758795385
P-value: 0.007477428928296717
Significant: True

Pearson's correlation coefficient between 'Negative GAAIS' and 'Engaging B': -0.4456958575840265
P-value: 0.0011799074640914315
Significant: True

Pearson's correlation coefficient between 'Negative GAAIS' and 'Untrustworthy A': 0.19383570079627738
P-value: 0.17740417302999748
Significant: False

Pearson's correlation coefficient between 'Negative GAAIS' and 'Untrustworthy B': 0.30130530945912976
P-value: 0.03347397501051172
Significant: True

Pearson's correlation coefficient between 'Negative GAAIS' and 'Quality A': 0.01447286865397971
P-value: 0.9205385494114677
Significant: False

Pearson's correlation coefficient between 'Negative GAAIS' and 'Quality B': -0.3486528575276912
P-value: 0.013086779970743 91
Significant: True

Significant Correlations:
'Computer Expertise' and 'Useful B' (Positive correlation)
'Conscientiousness' and 'Useful A' (Positive correlation)
'Conscientiousness' and 'Useful B' (Negative correlation)
'Conscientiousness' and 'Quality A' (Positive correlation)
'Conscientiousness' and 'Quality B' (Negative correlation)
'Negative Emotionality' and 'Quality A' (Negative correlation)
'Open-Mindedness' and 'Engaging A' (Positive correlation)
'Positive GAAIS' and 'Engaging A' (Positive correlation)
'Positive GAAIS' and 'Engaging B' (Positive correlation)
'Positive GAAIS' and 'Untrustworthy A' (Negative correlation)
'Positive GAAIS' and 'Quality A' (Positive correlation)
'Positive GAAIS' and 'Quality B' (Positive correlation)
'Negative GAAIS' and 'Engaging A' (Negative correlation)
'Negative GAAIS' and 'Engaging B' (Negative correlation)
'Negative GAAIS' and 'Untrustworthy B' (Positive correlation)
'Negative GAAIS' and 'Quality B' (Negative correlation)
```

Screenshot showing part of the output of Evaluate Results.py