Universiteit Leiden

Conformity, Comfortability, and Chatbots: How Virtual Avatars Affect the Likeability of an Agent

CJ Reitter (s3794490)

Supervisors:
Peter van der Putten & Maarten Lamers

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 9, 2026

Abstract

Large Language Models (LLMs) have been becoming more powerful and more popular in recent years, and with that, there has been a significant rise in the usage of personal and casual-use chatbots. Many of them have avatars attached to encourage users to anthropomorphize the chatbot. This raises the question of how precisely the human likeness of these avatars affects the chatbot's likeability. Previous research has mainly focused on the human likeness of the chatbot's messages, excluding a chatbot avatar entirely from the interface, which we propose may significantly affect user perception. This thesis is comprised of two main experiments, experiment 1 and 2. Experiment 1 consisted of participants ranking the human likeness of the profile pictures used within the study. This resulted in HuLA-50, an open soruce database consisting of 83 unique participants' 2339 total rankings. Experiment 2 consisted of a correlational study that analyses the relationship between the human likeness of a chatbot's avatar and its likeability. The relationship found between these variables from the results is plotted against Mori's Uncanny Valley model using hierarchical regression. A cubic relationship was not found between the two variables, with all null hypotheses being rejected. This indicates that there was no evidence found in the study that supports the relationship between the avatar's human likeness and the chatbot's likeability, in comparison to Mori's Uncanny Valley model. These findings suggest that in comparison to other features, such as message content and tone, the avatar of a chatbot during a short interaction does not have a substantial effect on the likeability of the chatbot. Analysis of how participants ranked the likeability of the other chatbot features seems to support a stronger correlation; how participants ranked the likeability of the chatbot's message contents appeared to have the strongest correlation with how the participant ranked the likeability of the chatbot overall.

Contents

1 Introduction

Large language models (LLMs) have increasingly been used as personal-use agents [5], with categories like "companionship" topping the use-case list in the Harvard Business Review's 2026 study [33]. A common form that these "companionship" agents take is "character chatbots" or "AI companions" [21]. In a paper by Zhang et al. [34], they attribute this rise to the chatbot's rapid response times, its ability to create a feeling of emotional vulnerability, and its overall likeability attributed to the anthropomorphisation of the chatbot's features, so they resemble a specific "companion". These features may include their tone of speech, a built-in "backstory", and a digital avatar.

For comparing anthropomorphisation and likeability, the current standard in academia is using Masahiro Mori's Uncanny Valley model [17, 18] as a reference model [11]. However, for the academic studies found related to the anthropomorphisation of chatbot interactions (e.g. those done by Tan et al. [30], Skjuve et al. [26], and Ma et al. [13]), the studies frequently did not include an avatar as a feature being studied within the paper, or use different dimensions besides likeability and visual human likeness.

In the case of the the papers found to be closest in scope to this thesis, Tan et al. and Ma et al., the variable of human likeness was self-defined and only consisted of two states, "high" and "low". This then presents a natural expansion of their research where the human likeness variable is expanded to a larger interval scale. Furthermore, as human likeness tends to be self-defined in these papers, this indicates the field lacks benchmark databases that rank the human likeness of digital avatars.

The field, in general, also does not have much empirical research within the scope of chatbots, either examining a different scope (e.g. [12, 32, 16]), or not attaching a specific scope to the interaction. For the latter papers, participants in these studies are instead presented an agent, commonly a physical robot (e.g. [2, 10]) or a digital 3D model (e.g. [4, 22, 12]) and asked to rank the likeability of the agent solely on its appearance. This establishes that there are a sufficient amount of papers that examine the relationship proposed in Mori's Uncanny model at a baseline without a realistic scope attached to them. Therefore, applying the model to more realistic scopes potentially fills a larger gap within academia.

Based on previous research within this field, there are 2 main purposes this thesis seeks to fulfil with 2 experiments. The first is to address the lack of sufficient benchmark databases related to the ranking of avatar human likeness by creating an open source database. The second is to expand upon previous research related to Mori's Uncanny Model within the scope of chatbots by focusing on the visual anthropomorphisation of a chatbot avatar on a larger interval scale than what was previously done.

The latter specifically attempts to answer the question "How does the level of anthropomorphisation of an LLM-based chatbot's digital avatar affect the chatbot's overall likeability?". Experiment 2 attempts to answer this question using the Uncanny Valley model, specifically, as a reference model.

The first experiment of this study was conducted by asking participants to rank the human likeness of the images used for the chatbot's profile picture. These rankings were used to assign human likeness rankings to each of the 50 images. The second experiment provided a new set of participants with a chatbot to interact with that had one of the images ranked in the first experiment attached to it. Each participant was then asked how features of the chatbot, including the avatar, affected the likeability of the chatbot, so the perceived likeability and human likeness

could be measured against one another.

The first experiment of the study resulted in an database containing 2339 rankings split between 83 unique participants. This database will be referred to as HuLA-50 (Human Likeness Avatar) throughout the paper. HuLA-50 is open-source and can now be used as a resource for other academics within the field of human-agent interaction. For access to HuLA-50 (found in the "avatar_ranking_imageranking" table of the "chatbot_database.sqlite3" database), it can be found in the "cj-reitter/lu_uncannyvalleystudy_chatbot" repository. This repository also contains the code for the websites that hosted both experiment 1 and 2.

The second experiment did not find any evidence that supported a relationship between the chatbot avatar's human likeness and the chatbot's likeability, in relation to Mori's Uncanny Valley model. Experiment 2 can therefore serve as a foundation for further academic research. For example, certain parts of the research question/design could be modified, comparing the modified study's results to the ones found in this experiment.

2 Background

The usage of LLMs as chatbots has only recently gained widespread popularity and is currently in the hundreds of millions [1, 8, 20], meaning that the academic study of chatbots is both topical and relatively novel. One of the most common ways that users interact with these chatbots is for casual/social interactions [5], specifically as companions [33]. This thesis attempts to analyse this type of interaction, specifically how the human likeness of the avatar of these chatbots might affect their perceived likeability. As human likeness and likeability are the two variables being compared, this thesis uses Mashario Mori's Uncanny Valley model [17, 18] as a reference model to compare this relationship to.

2.1 Mori's Uncanny Valley Model

The Uncanny Valley model developed by Masahiro Mori [17, 18] in 1970 has been used as a theoretical model throughout the human-agent interaction field of academia and research. The main conceit of the model is that the "affinity" of an agent is influenced by how much its anthropomorphised. Affinity, as described by Mori, is essentially the likeability of an agent, specifically on an inherent level due to the agent exhibiting familiar traits. Anthropomorphisation in Mori's paper is defined by a person finding human-like characteristics in an agent that is not necessarily human themselves. He measures it on a 0% to 100% scale, where 0% is, for instance, an industrial robot and 100% is, for instance, a healthy human.

The hypothesized relationship between human likeness and affinity in the paper can be shown in Figure 1, where there's a non-linear relationship represented between the two variables. Specifically, the affinity of the agent increases as the human likeness of it increases, before peaking. After peaking, there is a significant dip in the affinity of the agent as its human likeness is near-identical to a human, but not perfectly aligned. The likeability then increases rapidly again when the human likeness of the agent is almost exactly the same as a human.

That valley in the graph is known at the "Uncanny Valley". It is described as when a person interacting with an agent experiences a "loss of affinity" towards the agent, instead experiencing an "eerie sensation". Mori proposes that movement heightens this effect, but this won't be analysed as

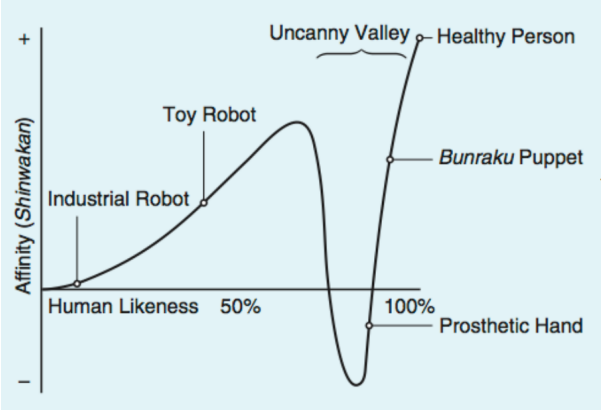

Figure 1: Uncanny Valley model in Masahiro Mori's The Uncanny Valley [18], representing the proposed relationship between the human likeness of an agent and its affinity.

profile pictures, for the most part, are still.

There are several proposed explanations for why humans experience this loss of affinity. Mori himself proposes this is a biological response of some kind, as objects like corpses fall into this valley. From this perspective, it is in the best survival interest of a human to have a desire avoid death and decay. Therefore, humans might have evolved to feel this uncanniness when presented with such imagery, motivating them to avoid it.

Another explanation is a mismatch of expectations. In Kätsyri et al. 2015 paper [11], they aggregated and reviewed many papers that analyse Mori's Uncanny Valley model. They found many papers propose that uncanniness comes from a mismatch between the expected and actual aesthetic of an agent.

Agents with very low human likeness (i.e. industrial robots or stuffed animals) are not expected to look/behave like a human, and agents with perfect human likeness (i.e. healthy humans) are expected to look/behave like a human. In both cases, the actual aesthetic of the agent ends up matching its expected aesthetic. Conversely, when agents have a human likeness in the Uncanny Valley range, they tend to be expected and/or designed to look/behave like an actual human (i.e. realistic computer generated avatars or androids) but fail to deliver on that expectation.

2.2 Character Chatbots

This thesis studies the research question through the variables in Mori's Uncanny Valley model, as it is a standard reference model in academia related to the anthropomorphisation of human agents [11]. Specifically, the study applies the model to the form factor of a chatbot due to the rise of LLMs being used for personal interaction or as "character chatbots". A working paper funded by OpenAI [5] reporting that approximately 70% of users use ChatGPT for non-work related topics.

Furthermore, looking at OpenRouter rankings [21] at the time of writing this paper, 2 of the top 10 external apps that report their usage of OpenRouter are entirely dedicated to chatbots with digital avatars (JanitorAI and ISEKAI ZERO). These character chatbots, or AI companions as they are frequently called, have seen a dramatic increase of usage. OpenRouter rankings in the "Entertainment" category, which is predominately apps dedicated to character chatbots, have risen from 121 billion to 680 billion tokens spent daily in the past year.

To get a better understanding of how people use these "character chatbots", so that they can be best replicated for the study, a paper by Zhang et al. [34] will be referred to. According to the paper, the most common use for these "AI companions" is "Emotional and Social Support" (80.33%), which they define as "Users seek empathy, advice, or companionship by sharing challenges, experiences, or health concerns for emotional support and connection" which falls into the "companionship" category, of the 4 categories they believe chatbot users fall into. This category represents users who use chatbots to interact with them in a way that they believe has social value, i.e. the way they would interact interpersonally with another person.

Note that for the Zhang et al. paper, it was found that many users of these chatbots see them as "companions" and discuss intimate or personal topics with them, specifically seeing them as a friend or romantic partner instead of an acquaintance. However, there are many potential ethical concerns with encouraging human participants to engage in that type of interaction, some of which are described in the Zhang et al. paper, that do not fall within the scope of this thesis. Therefore, for this thesis to study a relational interaction between a chatbot and user while limiting potential ethical concerns, the chatbot in this study will instead anthropomorphise itself as having an acquaintance-like relationship with the participant for the studies in this thesis.

3 Related Work

Up until the 21st century, the Uncanny Valley model generally remained unsubstantiated by empirical evidence. Currently, there are now many academic papers attempting to substantiate the model from different academic scopes. The majority of these papers do not examine chatbot interactions. Many of them do not examine a specific human-agent interaction, instead asking a participant to rank the likeability of the agent solely on its appearance. For example, Bartneck et al. [2] and Kim et al. [10] asked participants to rank the likeability of various physical robotics. Burleigh et al [4], Piweck et al [22], and Katsyri et al. [11] ran similar studies using digital 3D models, though they did not resemble profile pictures.

There are not as many papers within the scope of this thesis, which makes sense as LLM-based chatbots and "character chatbots" are a relatively novel use-case for AI. Most of this thesis' related work instead closely related to one of three categories; avatar analysis not specifically related to chatbots, chatbots without avatars associated with them, or chatbots with avatars associated with them.

3.1 Digital Avatars

Digital avatars have long been associated with other fields such as social media and gaming, meaning there is a lot of research already studying digital avatars as a feature. However, the form factor they are examining these avatars in is different than the one in this thesis. In the context of social media, a study by Shah et al. [25] examines the effect influencers with an AI avatar (DHAI as they describe it) have on user perception. They found that there was a positive correlation between the human likeness of these AI avatars and consumer engagement.

In the context of gaming/simulations, Mathur et al. [16] examined how the mean perceived likeability of an agent's avatar affects the agent's trustworthiness. The paper found that when an avatar had low likeability and high but not perfect human likeness, therefore falling into the

Uncanny Valley, it led to participant confusion on whether they should trust the decisions of the agent or not. Another paper with the gaming form factor, Visser et al. [32] examines how, within a gaming interaction, certain features of an avatar (i.e. presence of the avatar, gender, and AI disclosure) affect the intensity of a player's engagement within the game. The paper found that the the presence of an avatar and disclosure of the avatar as AI-heightened increased participant's play intensity, while an avatar with a masculine appearance decreases the intensity.

As all of these papers found some correlation between the avatar and a certain aspect of user perception, it is possible that avatars as a feature affect user perception. So, in the context of this thesis, it is possible that, within the form factor of a chatbot interaction, the chatbot's avatar affects the likeability of the chatbot.

3.2 Chatbots

Of the papers found that pertain to chatbots, related work within this scope does not appear to cover the exact independent and dependent variables (a chatbot avatar's human likeness and its likeability respectively) being used and measured in this thesis. The papers found to be closest to this thesis were Ma et al. [13], which included perceived intelligence as an independent variable, and Song et al. [27], which include both motion and celebrity-likeness as additional independent variables. Ma et al. did not find a significant relationship between perceived intelligence/anthropomorphisation and user experiences, while Song et al. found that user trust decreased and perceived eeriness increased when the chatbot was more humanized and did not imitate a celebrity.

Both of these papers used a 2 x 2 dimensional space for these independent variables, as in each of the 2 independent variables for both of these studies only had 2 states. This could potentially suggest why Song et al. found a significant relationship in their study, while Ma et al. did not, as what each study considered "high" or "low" anthropomorphisation for a chatbot avatar could have differed. Therefore, it would be reasonable to expand on their studies by putting human likeness on an interval scale rather than a binary one. Similarly, as they had self-defined anthropomorphisation levels and differing results, it would likely be beneficial for this thesis to crowdsource its ranking of chatbot avatar human likeness to mitigate potential bias.

There were also papers within the scope of chatbots that did not use the same independent variable. Skjuve et al. [26], for example, did not have an avatar attached to their chatbot, instead the measured the uncanniness of conversational transparency. They found that conversational transparency had no significant effect on the uncanniness of the chatbot. Instead, uncanniness appeared to increase when the message content of the chatbot became more ambiguous. Those results indicate that message content and tone do have a role in user perception. As this thesis does not have the goal of analysing these features, they should be controlled within the experimental design.

Finally, the UI for chatbot used within this thesis was informed by Chek Tien Tan et al. [30]. This paper examined the relationship between human likeness and a user's learning experience. They found that participants reported a better learning experience when the chatbot had an avatar compared to when it did not. As the Tan et al. paper found a significant relationship between the studied variables within this form factor, this thesis will use a comparable UI to the one they implemented.

4 Research Context

The basis for this study was ultimately to address two potential research gaps, as described in Section 3. The first was that many existing papers self-defined the human likeness ranking of the digital avatars used in their studies. This indicates a potential gap in benchmark data.

As studying the anthropomorphisation of digital avatars is an incredibly popular topic in academia across many scopes, it is worthwhile to create an open-source database for that purpose. This database will be created by crowdsourcing participants to rank the human likeness of an image dataset in Experiment 1.

The second potential research gap is an expansion of research related to the anthropomorphisation of chatbot digital avatars. This specific form factor was chosen as LLM-based chatbots have only become popular in recent years, making the analysis of their interactions with humans relatively novel within academia compared to other form factors.

The papers done by Ma et al. [13] and Song et al. [27] examined the relationship between the human likeness of a chatbot's avatar and the user experience. As mentioned previously, both papers described chatbot avatar human likeness as either "low" or "high". This lends itself to a potential expansion where human likeness is measured on a higher interval scale. As the proposed graph of Mori's Uncanny Valley model [17, 18] is continuous, placing human likeness on an interval scale rather than a boolean would allow for more comparable analysis between the results of the experiment and the model.

Therefore, experiment 2 attempts to expand on previous academia by analysing the relationship between a chatbot avatar's human likeness and the likeability of a chatbot. However, unlike previous research, this thesis expands human likeness to an interval scale. Furthermore, so the correlational relationship can be isolated for this thesis, no other independent variables will be analysed.

For experiment 2's relationship, the measured variable, likeability, was specifically chosen, opposed to the other variable "eeriness" discussed in Mori's original paper [17, 18], as it appeared to be absent from much of the related work that used Mori's model as a reference model. In general, most of the related work was measuring a trait that relates closely to how Mori described "eeriness". However, this was only one of the two concepts that Mori discussed originally for the dependent variable of his proposed model. Mori used the Japanese terms bukimi and shinwakan in relation to the affinity dimension. Many papers seem fixated on the bukimi, which, according to several Japanese translation dictionaries such as JapanDict [9], translates to "weird", "creepy", "uncanny", or "eerie", which ended up being the term the 2012 translation used.

It was much rarer to find ones related to shinwakan. This is understandable, as shinwakan does not have a direct English translation, though it appears to be translated to terms like "familiarity", "likeability" and "affinity" for papers related to the Uncanny Valley [11]. The reason shinwakan does not have a direct English translation is because it is a relatively new addition to the Japanese lexicon. Shinwakan is a compound word from "shinwa", which is directly translated into "harmony", "mutual friendliness", or "affinity", and the suffix, "kan", which translates to "a sense of" when associated with a feeling. Though it appears that shinwakan was used before Mori's paper, many papers, such as [6], believe that Mori coined the word, so it is now widely associated almost entirely with the field of artificial intelligence.

"Affinity" was the term they used within the 2012 English translation, which is likely why many people refer to it as the affinity dimension. This, as just established, is a common translation of "shinwa", which is only part of the full word. Furthermore, affinity is generally defined as

the inherent likeability of an agent or object, normally due to shared characteristics. This is not in direct opposition to eeriness, since eeriness refers to mysterious and frightening feelings, not specifically due to different characteristics. This is supported by many of papers cited above that refer to "eeriness" and "trust" as the two ends of the spectrum for their dependent variables, with trust being a much more appropriate antonym to eeriness.

4.1 Experiment 2 Research Question

Together with the described variables and form factor, experiment 2 aims to find if there is a relationship between the human likeness of a chatbot's avatar and the likeability of the chatbot. Specifically, it attempts to answer the research question: "How does the level of anthropomorphisation of an LLM-based chatbot's digital avatar affect the chatbot's overall likeability?".

This research question is specifically in regards to the shape of the relationship. The expansion of human likeness to an interval scale is one of the main features of experiment 2 that differs from previous research in regards to chatbots. Therefore the shape of the correlational relationship will be the most novel aspect of the results.

As the basis of this question is inspired by Mori's Uncanny Valley model [17, 18], it will be used as a reference model for mathematical comparison of experiment 2's results. Specifically, a positive cubic relationship will be used for comparison, as it mathematically approximates the Uncanny Valley model to a reasonable accuracy.

Experiment 2 will attempt to compare the results with Mori's Uncanny Valley model by using hierarchical regression, which attempts to train a model to match a specific relationship. With hierarchical regression, three models are trained on three possible relationship that could be found between human likeness and likeability (linear, quadratic, and cubic polynomial relationships), potentially providing evidence that answers the research question.

Note that the Uncanny Valley model is a theoretical graph that Mori did not create with a mathematical basis to it. A positive cubic polynomial relationship would theoretically approximate the proposed graph, but it is not a perfect replication of the model. However, regression is extremely biased towards over-fitting when the number of hyperparameters increases. Therefore, for experiment 2, regression will be capped at a cubic relationship to reduce the chance of potential false positives in the results.

To test if a positive cubic polynomial relationship is found, this study breaks the question down into 3 separate null hypotheses. Hypothesis 1 and 2 tests if a cubic relationship better explains the relationship between the two variables than a linear or quadratic one, and Hypothesis 3 tests if the cubic relationship is positive. Specifically, Hypothesis 1 is that there is no significant difference between the coefficient of determination (R^2) of the linear and cubic polynomial regression ($\Delta R^2_{1,3} = 0$). Hypothesis 2 is that there is no significant difference between R^2 of the quadratic and cubic polynomial regression ($\Delta R^2_{2,3} = 0$). Hypothesis 3 is that there is either no significant, or a negative cubic beta coefficient present in the cubic regression of the human-likeness of the chatbot avatar versus the likeability of the chatbot ($\beta_3 \leq 0$).

4.2 Ethical Considerations

Throughout the design and testing of experiment 1 and 2, ethical practices were upheld and potential ethical considerations taken into account. These ethical practices generally followed

Leiden University Ethics Review Committee's Code of Conduct for Research Integrity [31]. The thesis' experimental design was not presented to the Ethics Review Committee before deployment as, according to their checklist, neither experiment involved testing that required approval from the committee.

For both experiments, participants were given a consent form, as seen in Figure A. Experiment 1 participants and Experiment 2 pre-testing participants also received the "Pre-testing Notes" in addition to the rest of the consent form. Participants would digitally sign the consent form by selecting a check box confirming they read and agree to the terms in the consent form.

The consent form details the purpose of the study, how the experiment will proceed, ethical considerations the participant should take into account, and contact information for the researcher and primary supervisor. This section details the procedures each set of participants would follow to complete Experiment 1 or pre-testing for Experiment 2. It also states that no responses to the questionnaire, nor any demographic data would be recorded.

The ethical considerations stated in the consent form contained all of the ethical implications of the research that was taken into account for this thesis. This was to ensure participants were aware of all aspects of the experiment before completing the experiment. The only exception to this was with the experiment 2 participants, where the relationship being analysed was obfuscated to limit bias in the responses. Experiment 2 participants were told the full purpose of the study after completing the study.

The ethical considerations listed were related to confidentiality, voluntary participation, and potential concerns that may arise from conversing with a chatbot. The confidentiality section states that none of the parties involved (i.e. the researchers, Ollama, and DeepSeek) would log the content of prompts and chatbot responses. Ollama was specifically chosen as the API the chatbot calls DeepSeek-v3.2 from as Ollama runs the model locally from their cloud servers. The only data Ollama logs are token and prompt counts, as in the length and number of messages respectively, which was stated to the participants.

Despite not logging any of the message content, participants were asked interact with the chatbot as if a stranger would read the content of the prompts and messages. Furthermore, it was made clear to the participant that everything the chatbot generates is made up, and that the chatbot cannot substitute human-to-human interaction. This was to limit participants engaging in potentially intimate or dangerous conversations with the chatbot. The exact manner that LLM-based chatbots affect the mental state of users is a heavily debated topic at the time of writing this thesis. Therefore, it would be potentially unethical to encourage intimate interaction in experiment 2 without approval from an ethics committee. So even though many users of "companion chatbots" engage in intimate conversations, it was determined that intimate interactions with the chatbot should be discouraged for this thesis.

Finally, it was explained to participants that participation in the study, regardless of which experiment they were apart of, was completely voluntary. Participants were allowed to leave the study at any time and responses to the questionnaires/rankings were not logged unless they clicked submit. This was explicitly stated in relation to the chatbot interaction on multiple occasions to the participant. This was to encourage the participant to be in check with their feelings during the chatbot interaction, so they would disengage from the interaction should they feel unsafe or uncomfortable.

5 Experimental Design

This thesis is comprised of two experiments, henceforth known as experiment 1 and 2. During experiment 1, participants ranked the human likeness of an avatar dataset consisting of 50 avatar images that modelled the appearance of a generic profile picture (see Figure 3). The results from experiment 1 formed the HuLA-50 database.

As HuLA-50 can serve as a benchmark database for many other experiments outside of experiment 2, this thesis isolates experiment 1 from experiment 2. For this thesis, experiment 2 acts as a case study for a specific use of experiment 1, rather than experiment 1 acting solely as pre-testing for experiment 2.

The rankings of each image in HuLA-50 were aggregated to find the mean human likeness of each image in the dataset. This mean human likeness was assigned to all 50 images in the dataset, so that the avatar image dataset could be used for experiment 2.

During experiment 2, participants completed the study by interacting with a chatbot and then ranking their experience with the chatbot. Before experiment 2, there was also pre-testing for the experiment, where participants simulated participating in a preliminary design of experiment 2's study. Participants would provide feedback and potential improvements to the experimental design. This feedback was taken into account and implemented before testing began for experiment 2. The specifics of each of these phases is continued in the subsections 5.2, 5.1 and 5.3.

During both experiments, including experiment 2's pre-testing, the website was deployed on PythonAnywere under the domain name "lu-thesis-study.com" and shared with participants. All participants were aware of what data of theirs was being logged, when/where it would be logged, and what their data was used for. Each session was assigned an anonymous id and no individual demographic information will be shared throughout the thesis. Each participant voluntarily and it was made clear to them that they could leave at any time without any consequences.

5.1 Experiment 1: Rating Experiment

During experiment 1, participants were asked to rank the human likeness of the images within the avatar dataset. The avatar dataset is a spread of 50 images sourced and manually selected from Magnific (formerly Freepik at the time of sourcing) [15]. Figure 3 showcases 10 of these 50 images, with the other 40 having comparable formatting and/or art styles.

Images were sourced from Magnific as its stock image library offered a diverse range of art styles for their profile pictures compared to other alternatives considered. Particularly, the site had many clearly AI-generated images, allowing for the selection of images that would ideally be perceived as uncanny. Profile pictures of actual humans were not chosen as the majority of the profile pictures on some of the most popular sites for "companion chatbots" are artistic renderings rather than real photographs.

Participants were mainly comprised of academic and professional peers, along with online crowdsourcing via LinkedIn [23]. None of the participants during this phase were offered or received any incentives for completing either task. Experiment 1 ran from 23 April 2026 to 28 April 2026.

Experiment 1 was designed for participants to complete in approximately 3-5 minutes on average. For the task, participants were taken to a form (shown in Figure 2) where they were given 1 of the 50 images from the avatar dataset at random and asked to rank the human likeness of the avatar shown from 0%-100% on an interval scale of 10%, with 0% being an industrial robot and 100%

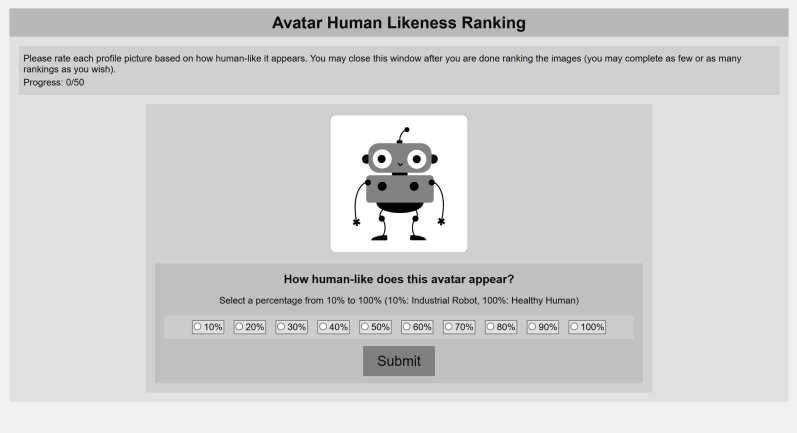

Figure 2: UI of the ranking page for participants performing Experiment 1. Images displayed were uniformly randomly selected from the subset of the 50 avatar images that had not yet been ranked by the participants.

being a healthy human. After submitting their rank, they were then shown another avatar, this cycle continuing until the participant ranked all 50 avatars. After all 50 avatars were ranked, the rating task was complete.

For experiment 1, participant data was analysed to determine if any participants should be filtered out. This was done by calculating the mean, median, and standard deviation of each participant's human likeness rankings, along with the count of how many rankings they completed. The mean/median was then plotted against the standard deviation to determine if there was any visual outliers.

Then, the Intraclass Correlation Coefficient (ICC) for the dataset was calculated, which provides statistics on the inter-rater reliability of the participants on average. As the results of experiment 1 are aggregated into an open-source database and used for the the ranking labels of the profile pictures in experiment 2, reporting the reliability of the participants is necessary for this thesis. ICC was used specifically as experiment 1 is a within-participant experiment with a measurable variable on an interval scale, making ICC one of the only reliable statistics for participant agreement and consistency within reliability statistics.

Furthermore, the mean human likeness ranking of each image was plotted against the standard deviation of each image to determine if there were any outlier images to be removed from the dataset. After filtering out any raters/images from the data, the data was standardized by calculating the participant's z scores, and then compared with the raw results to determine if standardization was needed.

Based on the aggregated results obtained from experiment 1, the mean human likeness was obtained for each image. These human likeness values were added as a label to each image in the dataset for experiment 2. This allows for the chatbot avatar's human likeness to be tracked between participants so the human likeness levels can be accurately compared.

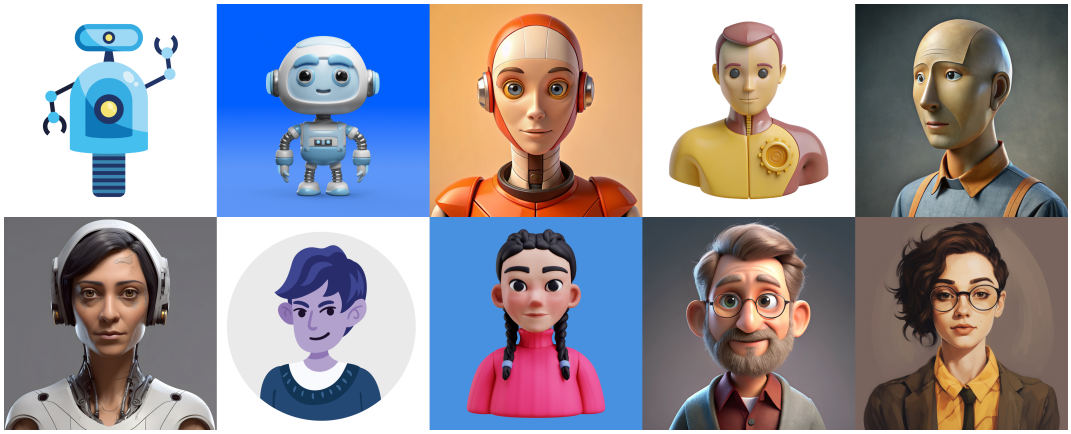

Figure 3: 10 of the 50 images used for the chatbot's avatar throughout the thesis. Each image belongs to one of the 10 human likeness rankings an image was assigned after experiment 1 (1-10), ordered horizontally from top left (human likeness of 1) to bottom right (human likeness of 10). All other 40 images were similar in format/art style to these 10 images.

5.2 Experiment 2 Pre-testing Simulation

During experiment 2's pre-testing, participants were asked to simulate the study of experiment 2 and provide feedback on the study's preliminary design and UI. This feedback was used to improve the experimental design of the study before testing for experiment 2 began. The specific improvements made in response to the pre-testing feedback are discussed in Section 5.3

Participants were comprised of academic and professional peers. None of the participants during this phase were offered or received any incentives for completing the simulation. Experiment 2's pre-testing ran from 23 April 2026 to 26 April 2026.

The simulation of experiment 2 took participants approximately 10-15 minutes to complete. For the task, participants were given a consent form to read before continuing (see Appendix for Figure A). The consent form was extremely similar to the one given to participants during the testing phase, with the addition of a section for pre-testers, so that they could give feedback on the consent form if they had any. After agreeing to having read the consent and being over the age of 18, participants were able to begin the simulation of the study.

In the first phase of the study, participants were given a chatbot interface with one of two UIs, as shown in Figure 4, inspired partially by the interfaces in Figure 1 of [30]. The picture used for the chatbot's avatar was randomized uniformly between the 50 profile pictures within the image dataset.

The chatbot's responses when the participant prompted it were generated using DeepSeek-v3.2, selected based on its popularity of use on OpenRouter [21] in the "roleplay" category, the most analogous category they have to a casual use chatbot. The DeepSeek model [7] used for the study was run on Ollama's Cloud servers via the Ollama Cloud API, a service provided by Ollama [19] that allows models to be localized on their servers so the model provider cannot access prompt requests or chat logs.

When called, the DeepSeek model was given the system message: "You are a friendly chatbot talking to a human user. Act like a friendly stranger. Keep answers brief, formatted like small talk, and keep the conversation flowing. Avoid asking for personal information from the user. Divert

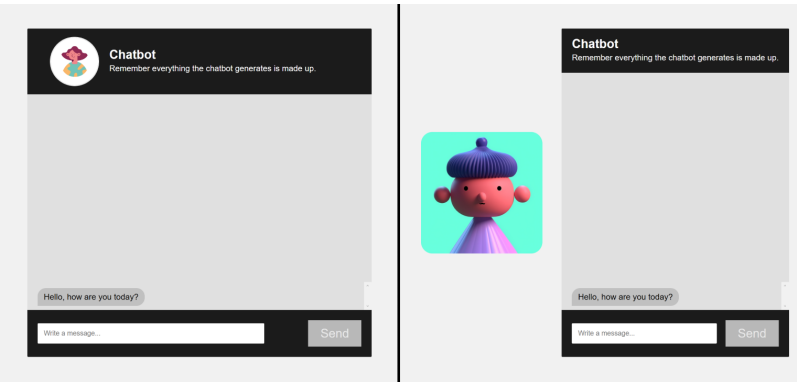

Figure 4: From Left to Right: UI of chatbot page option 1 & UI of chatbot page option 2. Each pre-testing participant during experiment 2's pre-testing had an equal chance of each UI being displayed. The chatbot's profile picture displayed was uniformly randomly selected from a set of 50 possible profile pictures.

the conversation if it becomes too intimate.". The model was also given the chat history of the conversation, including the participant's latest prompt. No other hyperparameter adjustments, such as temperature, were made to the model throughout either phase of the experiment. Specifically, the model's controlled hyperparameters were set to DeepSeek-v3.2's defaults (i.e. the temperature parameter was set to 1.0).

After interacting with the chatbot for 5 minutes, the participant is automatically redirected to a questionnaire (see Appendix for Figure B), consisting of 10 ranking questions on a Likert scale of 1-5 (1 = Strongly Disagree, 5 = Strongly Agree) based on the "Likeability" section of the Godspeed Questionnaire [3], 5 open-ended questions, and 2 demographic questions on the participant's age and gender. Both the ranking and open-ended questions asked the participant about the chatbot's avatar, the UI, message tone, message length, and overall experience. This was done to obfuscate the exact relationship being studied while the participant was completing the survey, limiting potential bias from the participant knowing the purpose of the study.

Participants were allowed to fill in the questions to better simulate the study, if they desired, despite all questions being optional and none of their responses being logged. Participants then continued onto a form asking them to provide feedback on certain features of the study, specifically technical problems, the questions asked on the questionnaire page, if the purpose of the study was successfully obfuscated, and the chatbot interface.

After submitting their feedback, their responses were logged and the simulation was complete. Participants' feedback was reviewed manually and taken into account when updating the experimental design of experiment 2.

5.3 Experiment 2: Avatar Human Likeness vs. Chatbot Likeability

During experiment 2's pre-testing (see Section 5.2), 9 participants responded. Based on these participants' feedback, several changes were made to the layout of the experiment. Unless otherwise stated in this section, the study ran the same way as the simulation ran during experiment 2's pre-testing, as it was described in section 5.2. The UI of chatbot page option 1 in Figure 4. was selected. Pre-testers reported this version of the UI was more analogous to a chatbot interaction a

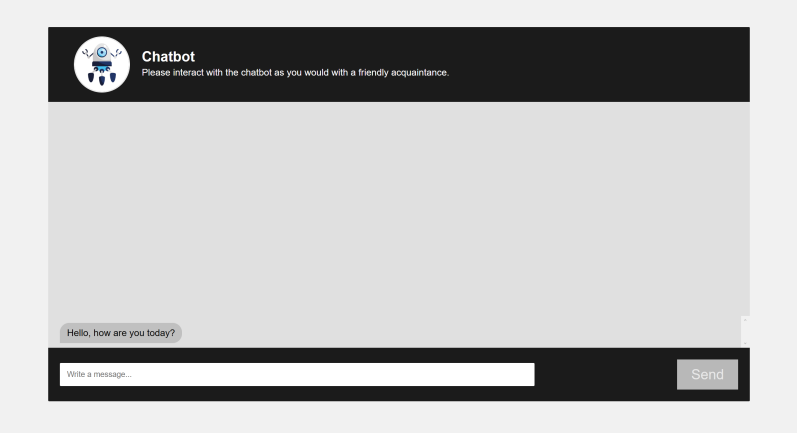

Figure 5: UI of the chatbot page for participants completing the study during the testing phase. Images displayed were randomly selected from a set of 50 possible profile pictures, the randomness weighted by the images' human likeness label (1-10), so each subset of images with a given human likeness had a uniform chance of being selected.

user would encounter in real life. Furthermore, pre-testers using mobile devices reported having trouble with the UI of option 2.

Slight modifications were also made to the UI of option 1, specifically to the size of the profile picture and the header subtext (shown in Figure 5. This was to increase the visibility of the profile picture and clarify the context of the interaction respectively.

As mentioned previously, the aggregated results from experiment 1 were used to label the human likeness of each image in the dataset used for experiment 2. The spread of the mean human likeness of each image is not uniform (see Figure 9 in Section 6). So, rather than the uniform randomisation used to generate the chatbot avatar's profile picture during pre-testing, the randomization will be weighted based on this spread. Thus, for all participants, there is an equal chance of the chatbot having an avatar with a given human likeness. This is to reduce potential bias in the data that may come from under/over-represented human likeness values.

The questionnaire was also modified for the testing phase (see Figure C). When asked if the questionnaire accurately obfuscated the purpose of the study for the participant, most of the pre-testers gave feedback that it did not. Therefore, modifications were made to better obfuscate the purpose.

With the modification to the questionnaire, the participants are now asked 20 ranking questions, still based on the same Likert scale during pretesting, now separated into 4 sections: Overall Chatbot Interaction, Chatbot Message Content, Chatbot Message Tone, and Chatbot Avatar. Each section contained a 2/3 or 3/2 split of 5 questions inspired by the "Likeability" and "Perceived Intelligence" sections of the Godspeed Questionnaire [3] respectively. The questions related to "Perceived Intelligence" act as distractor questions and the questions related to "Likeability" act as experimental questions, resulting in an even 10/10 split. The open-ended questions corresponded to the 4 ranking sections, with an additional question asking about the participant's experience with the UI. The demographic questions remained the same. All ranking questions were mandatory for participants, while the open-ended and demographic questions were optional.

During testing, digital crowdsourcing was used, the site shared on several channels, including

SurveyCircle [28], SurveySwap [29], R/SampleSize on Reddit [24], and LinkedIn [23]. None of the participants during this phase were offered or received any monetary incentives for the study. Participants from SurveyCircle and SurveySwap received a small amount of points and karma respectively for completing the study, which are used on each site to boost engagement for the participant's own study on the site. However, it was made clear to said participants how many/much points/karma they received, and that how they responded to the study would not affect the quantity of points/karma received after completion.

Responses were filtered solely by responses to the open-ended questions, with no response removed based on their answers to the ranking and demographic questions, as all ranking questions were mandatory. The criteria for filtering out a response was if their responses made it clear that they were not taking the study seriously to a certain capacity (e.g. insulting the experiment/person running the experiment in the responses), or their responses made it clear that their session was biased in some way that the average session wasn't (e.g. the chatbot was down during their session, or they refreshed the page and the timer for the page reset).

After filtering the data, data analysis occurred. This analysis consisted of hierarchical regression analysis, specifically linear, quadratic, and cubic, to see if the mean levels of likeability fit to a formula similar to the graph shown in the Uncanny Valley model. Analysis was performed on the mean responses of all experimental question within each of sections, then the mean responses of all experimental questions within each section, then the mean responses for all 10 experimental questions. Open-ended questions were analysed manually to see any patterns in the responses that may affect the results. The beta coefficient (β_3) will be obtained from the cubic regression and the change in the coefficient of determination (ΔR^2) will be calculated to determine whether or not the null hypotheses can be rejected or not rejected.

6 Experiment 1 Results

For experiment 1, 81 participants took part, with each participant on average ranking approximately 28 images, for a total of 2283 rankings. Each image had, on average, approximately 47 rankings. These rankings were used to compile the HuLA-50 database, which can now be found in this repository.

6.1 Participant Filtering

After plotting the mean/median human likeness ranking participants gave to the images vs. the standard deviation, most of the visual outliers had ranked significantly fewer images than the total 50 they could rank, as shown in Figure 6. Conversely, the participants who ranked all 50 images generally clustered within the 30-60 range for their mean ranking. This indicates that, by seeing the entire image set, participants who ranked all 50 images compared the images against each other more fairly and provided more accurately distributed rankings.

To test the inter-rater reliability beyond standard deviation, the six common single-measure and average-measure ICC estimates are calculated. These estimates are shown in Figure 7. Note that the one-way models (ICC(1,1) and ICC(1,K)) are not exceptionally relevant statistics compared to the other ICC estimates, as they measure the rater reliability assuming a between-subject experimental set-up. The average amount of rankings for each image, and therefore the average k,

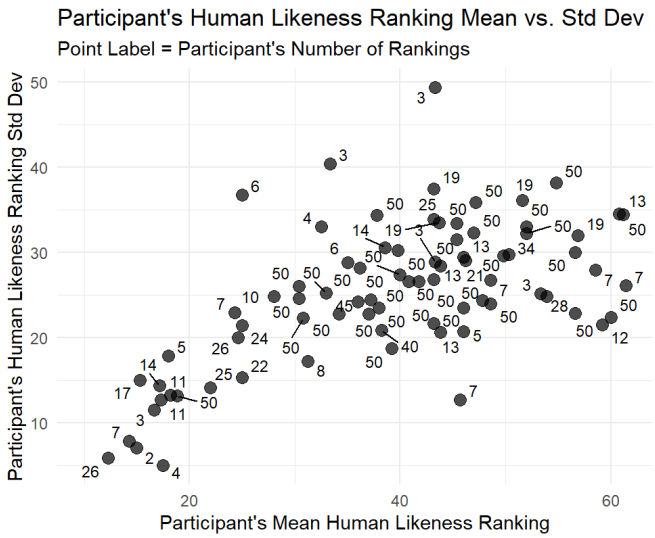

Figure 6: Scatter plot of participant's mean human likeness ranking vs. the standard deviation of their rankings during the rating task of the pretesting phase. Each data point represents a unique participant, with each point labelled with the number of images the participant ranked (maximum of 50 images).

ICC Table	
ICC Type	ICC Value
ICC(1,1)	0.5572
ICC(2,1)	0.5582
ICC(3,1)	0.6906
ICC(1,k)	0.9905
ICC(2,k)	0.9906
ICC(3,k)	0.9946

Figure 7: Table of the mean ICC values for each participant of Experiment 1. ICC models ICC(1,1), ICC(2,1), ICC(3,1), ICC(1,k), ICC(2,k), ICC(3,k) are shown.

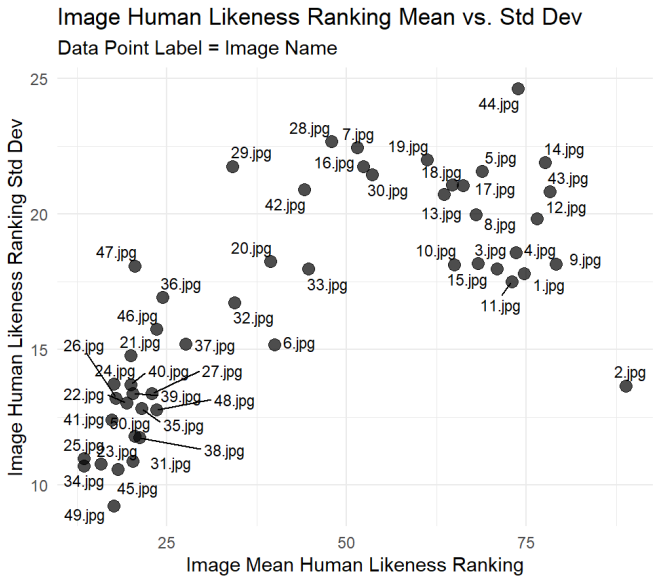

Figure 8: Scatter plot of each image's mean human likeness ranking vs. the standard deviation of its rankings, obtained during Experiment 1. Each data point represents an image, labelled with the image name.

was approximately 47, and the 33 participants ranked all 50 images. Therefore, most of the raters for each image encountered all 50.

The two-way model estimates are therefore the ICC values that were taken into account when interpreting rater reliability. First, the ICC(3,k) model value is 0.9946. This value indicates that for all participants, without generalizing the sample to the whole population, their average ranking for every image is extremely reliable. The ICC(2,k) model value value is also a near perfect estimate at 0.9906. This indicates that the reliability of all participant's average image ranking is also extremely high under the assumption that the sample is representative of the entire population.

As both ICC(3,k) and ICC(2,k) are near perfect estimates and there is a large enough sample size of participants, ICC(3,1) and ICC(2,1) model estimates can also be ignored for the purposes of this thesis. Both model estimates relate to the reliability of a single rater's given rankings, which is less relevant when the rankings of all participants are being aggregated.

As the estimated reliability for participants on average was extremely high, it was determined per-rater ICC would not be used as a factor for filtering out participants. Conversely, the number of rankings a participant had appeared to be correlated to deviation from the cluster of all participants. Therefore, any participant that had ranked less than 50 images would be filtered out.

6.2 Image Filtering

After filtering out participants that ranked less than 50 images, the mean human likeness ranking of each image was plotted against its standard deviation. This can be seen in Figure 8. As shown visually, there are no images with an exceptionally high standard deviations compared to any other image. Therefore, no images were filtered out before deployment. With the filtered out participants, this resulted in a filtered data set with 33 participants and 1650 rankings.

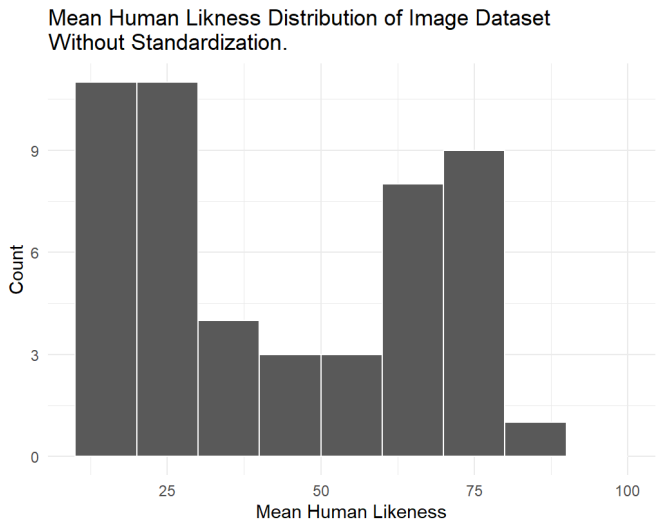

Figure 9: Distribution of the mean human likeness rating the filtered participants gave the 50 images during Experiment 1. Participants were filtered based on if they ranked all 50 images within the dataset. No standardization was used.

6.3 Result Standardization

These filtered rankings were standardized by each participant's z-scores, resulting in the spread of mean human likeness rankings shown in Figure 9. Comparing this to the spread of human likeness rankings for the filtered results without standardization (see Figure 10), the spread of results is not exceptionally more uniform after standardization. Standard deviation in relation to the z-score also did not decrease exceptionally. However, by standardizing the results, the images are placed in all 10 human likeness bins, unlike the non-standardized filtered results, where no image is given a human likeness ranking of 9 or 10. Since experiment 2 is measuring the likeability of the chatbot across the full scale of human likeness (10%-100%), for the purposes of this thesis, the image dataset will be assigned human likeness rankings based on the standardized average human likeness rankings.

7 Experiment 2 Results

In total, the study received 128 responses, 3 of which were filtered out due to the high likelihood they were duplicate responses, resulting in a total of 125 participants. Of the participants who disclosed their age, participants ranged in age from 19-60, with the mean age being 28 years old and the median age being 24 years old (see Figure 11). The age skewed younger as many of the participants were obtained from sites whose largest demographic are in their twenties. Of the participants who disclosed their gender, 63 identified as "female", 47 as "male", and 2 as "non-binary" or "other". A further 3 responses were filtered out, 2 due to the responses being written in another language and 1 due to their responses making it clear their session's experimental integrity had been compromised. This resulted in a total of 122 responses used for the regression analysis.

For the experimental questions (ranking questions related to likeability), Cronbach's raw alpha was 0.9 and standardized alpha was 0.91, indicating the sample of participants is internally consistent.

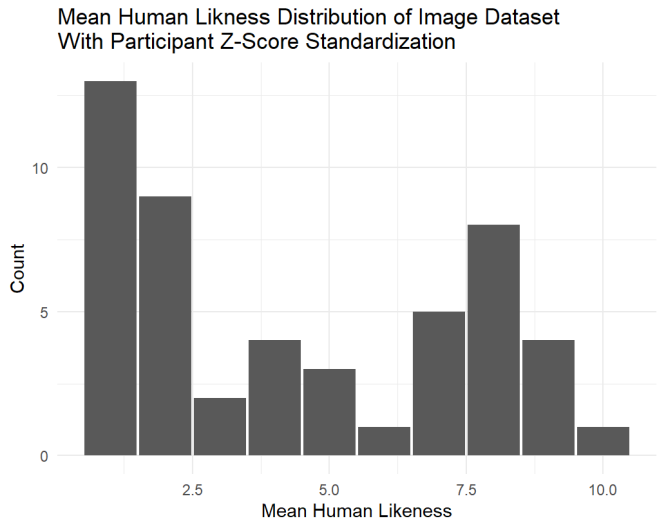

Figure 10: Distribution of the mean human likeness rating the filtered participants gave the 50 images during Experiment 1. Participants were filtered based on if they ranked all 50 images within the dataset. Standardized by participant z-score and placed into 10 equal width bins that represent the 10 intervals of human likeness used for the thesis (10%-100%).

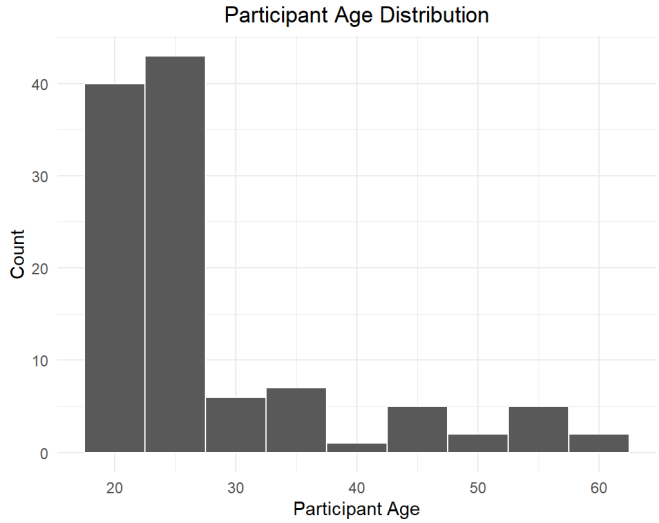

Figure 11: Histogram of the ages of participants who disclosed their age during Experiment 2.

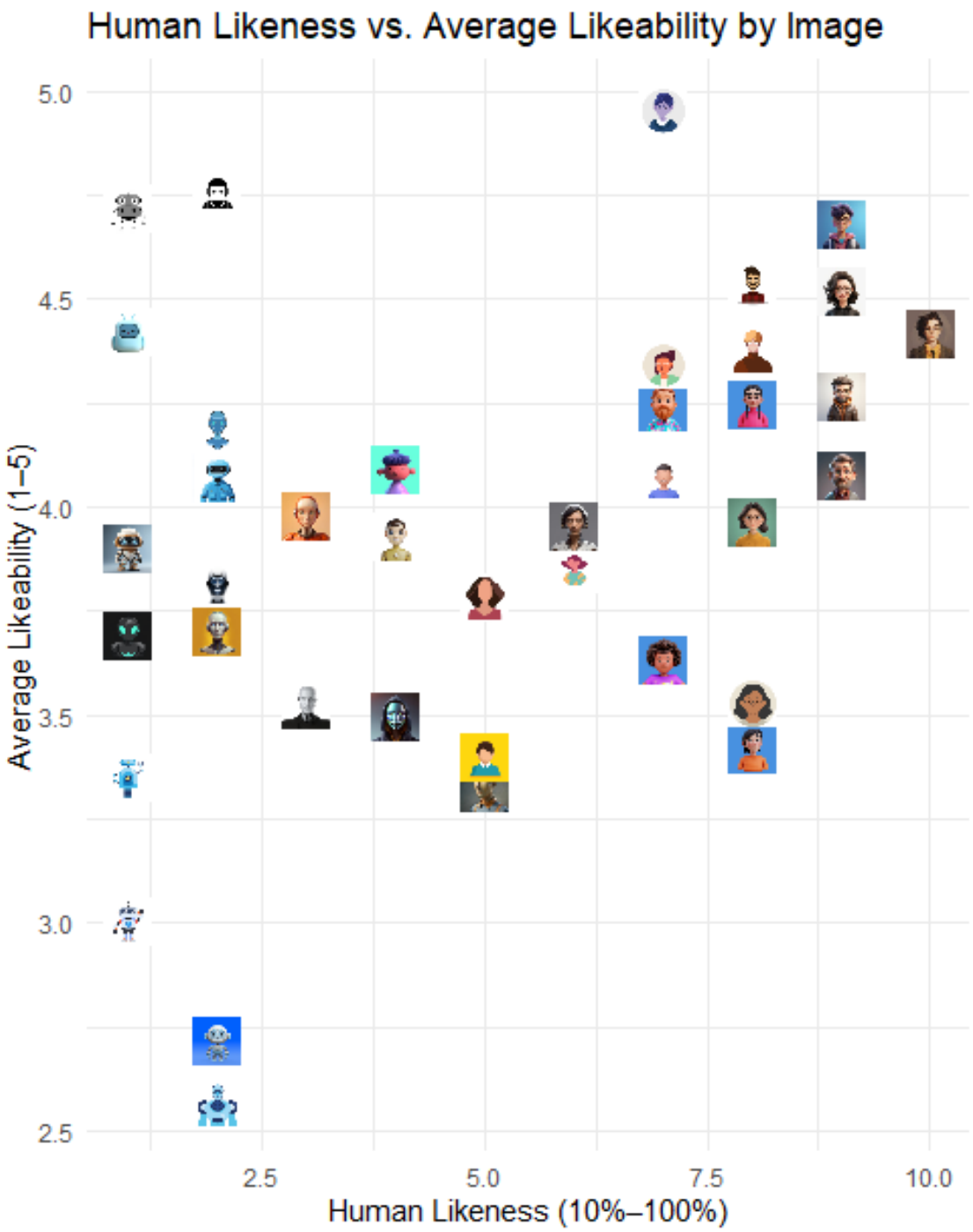

Figure 12: Scatter of the human likeness of an image from experiment 2's dataset versus the mean likeability ranking of the image. Mean likeability was calculated by taking the average chatbot likeability ranking for all sessions where the chatbot had the given image id attached. Each data point represents an image within the dataset. Note that, due to overlap, the y-axis interval was condensed and a small jitter on the y-axis was added to visually separate the images, so actual mean likeability might differ slightly than what is presented visually.

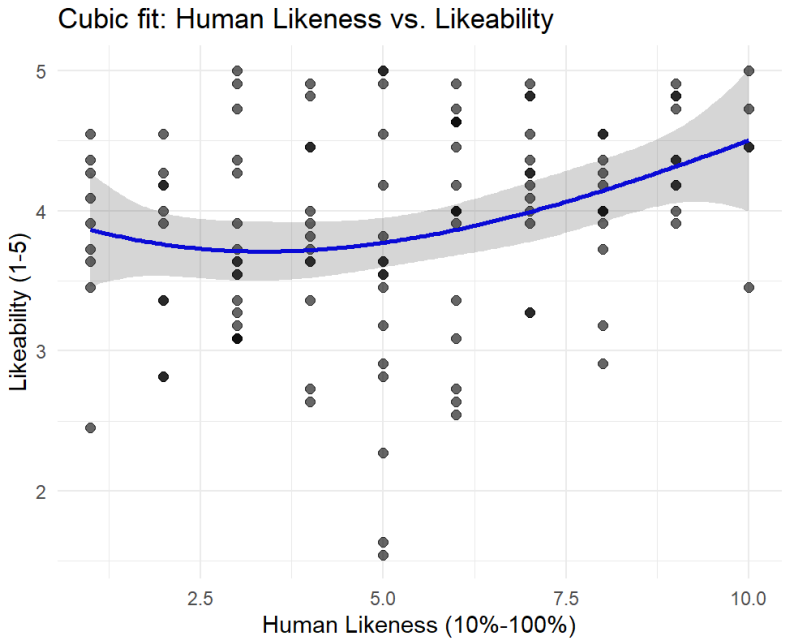

Figure 13: Scatter of participant's average response to all experimental questions (ranking questions pertaining to the likeability of the chatbot) vs. the human likeness of the chatbot avatar they were randomly assigned. Each point represents a participant's average response, with a darker opacity indicating more than 1 participant had the same mean response. the Blue line represents the best fit line predicted using cubic polynomial regression. Grey ribbon represents the confidence bands of the best fit line with a confidence interval of 95%, calculated using the standard error of the best fit line's predicted mean.

Therefore, the sample will not be further filtered nor will it be standardized in any way for the analysis. Furthermore, when calculating the mean likeability ranking for each image in the dataset, shown in 12, no image exceptionally deviated in its mean rank compared to other images of the same human likeness interval. Therefore, no sessions with a specific image id were removed.

7.1 Research Question Analysis

First, for every participant, the mean of their ratings for all 10 likeability questions is found. This is plotted against the human likeness ranking of the chatbot avatar they interacted with. Performing hierarchical regression on this scatter of data, the cubic polynomial fit for the human likeness and likeability relationship is shown in Figure 13.

The relationship was then controlled by section, as in the mean likeability rating was taken for each section of the questionnaire (Overall, Message Content, Message Tone, and Avatar), with each rating plotted on 4 separate graphs. Hierarchical regression was also performed on these 4 scatters and the cubic fit can be shown in Figure 14.

Visually, all of the best fit lines are either extremely flat or have a minimal decrease before trending upward. This appears to indicate that for all 5 scatters, none of best fit lines model the Uncanny Valley model [18]. For the global fit graph and the graphs standardized by sections "Overall", "Message Content", and "Message Tone", all best fit lines do not appear to model a

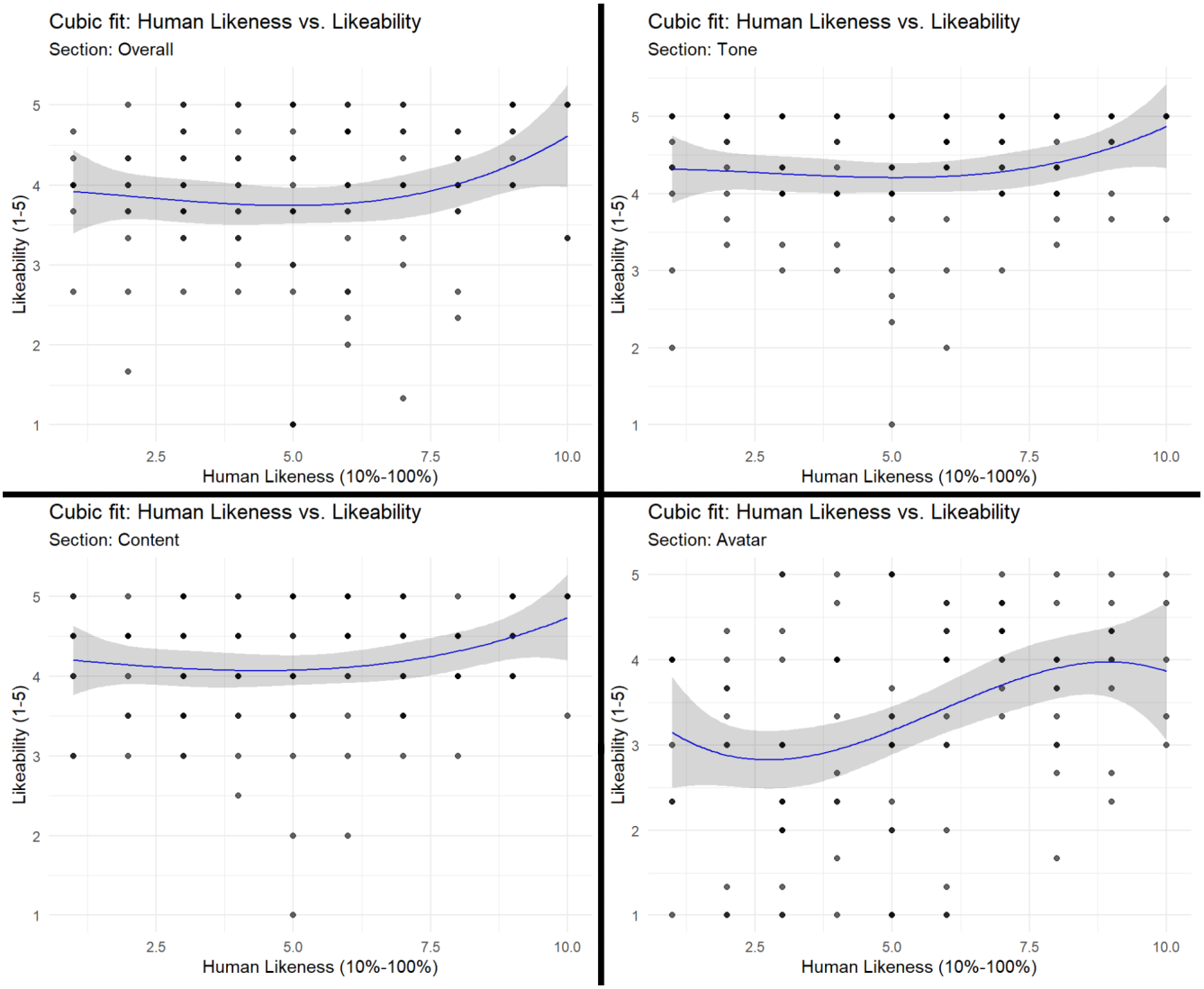

Figure 14: Scatter of participant's average response to the experimental questions (ranking questions pertaining to the likeability of the chatbot) vs. the human likeness of the chatbot avatar they were randomly assigned, standardized by section. Each point represents a participant's average response, with a darker opacity indicating more than 1 participant had the same mean response. the Blue line represents the best fit line predicted using cubic polynomial regression. Grey ribbon represents the confidence bands of the best fit line with a confidence interval of 95%, calculated using the standard error of the best fit line's predicted mean. For each graph, the sections used were: Overall Chatbot Interaction (Top Left), Chatbot Message Content (Top Right), Chatbot Message Tone (Bottom Left), and Chatbot Avatar (Bottom Right).

	Polynomial Regression Fit β_3 Comparison		
	β_3 estimate	t-statistic	p value
Global	0.0025	-0.276	0.6086
Overall Section	0.0011	0.4897	0.3126
Content Section	0.0023	0.2614	0.3971
Tone Section	0.9841	0.5445	0.2935
Avatar Section	-0.0096	-1.5068	0.9327

Figure 15: Table of the β_3 estimate, t-statistic, and one-sided p-value in the positive direction found performing cubic polynomial regression on the scatter of the human likeness of the chatbot avatar versus the mean rating of the chatbot's likeability by a participant. Regression performed globally and standardized by section (Overall, Content, Tone, and Avatar).

cubic polynomial relationship. This indicates that the β_3 estimate is close to 0.

For the graph standardized by the "Avatar" section, the best fit line appears to model a negative cubic polynomial relationship, indicating the β_3 estimate is likely negative. This means the likeability of the chatbot dips before peaking, which is the inverse of what the Uncanny Valley model suggests.

When standardized by the "Message Content" or "Message Tone" sections, there appears to be very little of a polynomial relationship found within the data. This is reasonable when standardizing for the chatbot's message tone and content, since those features of the chatbot were controlled for. Furthermore, the scatter in general trends towards higher likeabilities and has the smallest confidence bands for those 2 graphs, which indicates that the two variables were successfully controlled for.

All of the scatters, regardless of section, can not be sufficiently explained with a positive cubic polynomial relationship. Not only is the cubic coefficient B_3 negative for all 5 graphs, the coefficient is also not statistically significant for any of the graphs, as can be seen in Figure 15. For the global cubic fit, the estimated $B_3 = -0.0011$ with a t-statistic $t = -0.276$. Since the alternative hypothesis for Hypothesis 1 is one-side ($\beta_3 > 0$), the one-sided p-value in the positive direction is found which, with 118 degrees of freedom, is $p = 0.6086342 > \alpha = 0.05$.

Similar results were found when standardizing by section, the avatar section have a negative estimated β_3 while the rest have a positive estimate. For all sections, the one-sides p-value in the positive direction of the t-test on the coefficient was also greater than α. This means that Hypothesis 1 cannot be rejected, both globally and section-by-section.

It was also found that the cubic polynomial best fit line did not explain the data any better than a quadratic or linear best fit did, all 3 explaining the insufficiently. This indicates that none of the tested shapes were found in the data. Hypothesis 1 not being rejected had already indicated this result, that Hypotheses 2 and 3 are also unable to be rejected. This is confirmed through analysis of the changes in R^2, as shown in Figure 16.

For the global cubic polynomial fit, $R^2 = 0.0922$, the change in R^2 between the cubic and quadratic fits being $\Delta R^2_{2,3} = 0.0006$, and the change in R^2 between the cubic and linear fits being $\Delta R^2_{1,3} = 0.0280$. This already indicates that the change isn't substantial, as this aligns with the inherent increase in R^2 that comes from adding degrees to the fit.

To further determine statistical significance, nested F-testing is done with ANOVA (analysis of variance) to determine if there is a significant improvement. For the cubic and quadratic fit models,

	Polynomial Regression Fit R^2 Comparison					
	$\Delta R^2_{1,3}$	F-statistic	p value	$\Delta R^2_{2,3}$	F-statistic	p value
Global	0.0280	1.8208	0.1664	0.0006	0.0764	0.7827
Overall Section	0.0323	2.0129	0.1382	0.0019	0.2398	0.6253
Content Section	0.0249	1.5438	0.2179	0.0006	0.0683	0.7942
Tone Section	0.0260	1.6012	0.2060	0.0024	0.2965	0.5870
Avatar Section	0.0224	1.5081	0.2256	0.0169	2.2705	0.1345

Figure 16: Table of the change in R^2 between the cubic and quadratic regression models ($\Delta R^2_{2,3}$) and between the cubic and linear regression models ($\Delta R^2_{1,3}$), along with the F-statistic obtained using ANOVA, and the two-sided p-value. Obtained performing R^2 and nested-F analysis on the hierarchical regression models trained on the scatter of the human likeness of the chatbot avatar versus the mean rating of the chatbot's likeability by a participant. Regression performed globally and standardized by section (Overall, Content, Tone, and Avatar).

the F-statistic is 0.0764, making the two-sided p-value $p = 0.7827 > \alpha = 0.05$. For the cubic and linear fit models, the F-statistic is 1.8202, making the two-sided p-value $p = 0.1664 > \alpha = 0.05$. Therefore, there was not a statistically strong enough improvement in the global cubic model. Similar results were obtained when standardizing by section, with no substantial change in R^2 and all 8 F-statistics resulting in p-values greater than α. This means that Hypotheses 2 and 3 cannot be rejected, both globally and section-by-section.

7.2 Section Correlation Analysis

Since all hypotheses were rejected, this thesis will examine other features of the chatbot, as these other features may be masking the relationship between the avatar's human likeness and the likeability of the chatbot, or have a stronger independent effect on the likeability. This was already partially done by standardizing by section (see Figure 14), indicating that the other two features participants were asked about (message content and tone) generally had a horizontal spread with little standard deviation.

To determine if and how the other features have an effect on the participant's overall perception of the chatbot's likeability, the relationship between the likeability of the chatbot controlled for tone, content, and avatar, and the likeability of the chatbot controlled for the "overall" section will be analysed. Specifically, if and by how much a participant ranking the likeability of a certain feature of the chatbot affected how they ranked the chatbot's overall likeability. The likeability standardized by the "overall" section was chosen rather than the global mean likeability, as the global likeability was calculated by taking the average of all ratings across sections, meaning there is an inherent statistical relationship there.

To examine this relationship, multiple linear regression is performed to find the best fit model between the likeability of the three features and the "overall" likeability. Two regression models are trained, one where all four variables are evaluated raw, and one where the variables are standardized so their standardized beta coefficient estimates can be accurately compared based on effect size. The results of the best fit models can be seen in Figure 17.

For all three of the features, $p < \alpha = 0.05$, indicating a statistically significant correlation between the likeability of the 3 features and the overall likeability. Comparing the standardized beta

	Overall Likeability vs Chatbot Features Regression Fit			
	β estimate	t-statistic	p value	β' estimate
Intercept	-0.3853	-1.240	0.217	<0.0001
Content Likeability	0.6201	5.669	<0.001	0.5163
Tone Likeability	0.2912	2.619	0.009	0.2438
Avatar Likeability	0.1273	2.772	0.006	0.1658

Figure 17: Table of the models obtained from performing regression on the mean likeability of the chatbot for sections "Content", "Tone", and "Avatar" best fitted to the mean likeability of the chat for the "Overall" section. The standardized beta coefficient (β') estimate was obtained by standardizing the four variables during training, allowing for more accurate comparison of the estimate values.

coefficient (β') estimate for each section, message content had the largest estimate of $\beta' = 0.5163$, message tone had the second largest estimate of $\beta' = 0.2438$, and avatar had the smallest estimate of $\beta' = 0.1658$. This indicates that message content potentially has the largest effect on the overall likeability of the chatbot and avatar has the lowest.

This is partially supported by examining the partial R^2 for each variable, a metric that indicates how much of the data can be explained by a single predictor when isolated, in this case the likeability of one of the three features. The total R^2 for the non-standardized model is $R^2 = 0.6272$. For each feature's partial R^2, message content likeability as a predictor had $R^2_{mc} = 0.2140$, message tone likeability had $R^2_{mt} = 0.0550$, and avatar likeability had $R^2_a = 0.0611$. While the partial R^2 for avatar likeability is slightly larger than the partial R^2 for message tone likeability, message content likeability had an exceptionally higher partial R^2 than the other two features. This indicates that of the three features are predictors, the message content likeability explains the most of the scatter (approximately 21%).

7.3 Age and Gender Correlation

Other possible variables that could have affected the relationship between the chatbot avatar's human likeness and its likeability is the gender or age of the participant. As discussed previously, there was a relatively uniform split of the participant gender, and a skew towards younger participants within this experiment. To determine if these variables affected the relationship, regression will be performed between human likeness and the global likeability of the chatbot, moderated for age and then for gender.

Participant gender was found to not significantly influence how participants responded to the chatbot. Moderated for gender, the model did not find a statistically significant correlation between gender and likeability, nor did it find a correlation between human likeness and likeability when the relationship was moderated by gender.

Participant was found to slightly influence how participants responded to the chatbot. Moderated for age, the model found a small statistically significant correlation between the age of the participant and the chatbot's likeability (B_{age} estimate $= 0.0160$ with a p-value $p = 0.0301 < \alpha$). So older participants were more likely, on average, to view the chatbot as more likeable. However, there was no statistically significant relationship found between human likeness and likeability when the relationship was moderated by age. So age did not significantly affect whether or not the human

likeness of the chatbot avatar affected the chatbot's perceived likeability.

8 Discussion

Experiment 2 aimed to determine if there was a relationship between the human likeness of a chatbot's avatar and the likeability of the avatar similar to that of Mori's Uncanny Valley model, using a positive cubic polynomial relationship as a mathematical approximation. The empirical results found that none of the hypotheses could be rejected, and therefore there was not empirical evidence found in this study that a positive cubic polynomial relationship exists between the 2 variables. The model fit in general of the cubic polynomial model was also very small ($R^2 = 0.0922$), meaning there is a lack of evidence supporting the relationship regardless of statistical significance.

Therefore, in the context of the scope of the study, there was no evidence found that the human likeness of an avatar is a predictor of the likeability of a chatbot during short 5-minute sessions where the user engages with the chatbot the way they would an acquaintance. Furthermore, the results indicates there is either no evidence that can be used to answer the research question or there is no relationship between the chatbot avatar's human likeness and its likeability.

To further close the academic gap and lack of studies using this form factor, the thesis also conducted Experiment 1 to provide open source data for further academic research. This was achieved with HuLA-50, an open source database of 83 unique participants' 2339 total rankings of the perceived human likeness of the 50 profile pictures that is now accessible via the thesis' open repository.

8.1 Implications

The results from Experiment 2 indicate three main potential implications. The first possible implication is that Mori's Uncanny Valley model [18] is not universal, and that there are only certain relationships related to human likeness that mirror this model. As discussed in Section 3, there are many studies that have previously found relationships that reflect the Uncanny Valley model, meaning there is an empirical precedent for the existence of the model. It could be that for this specific relationship, human likeness versus likeability in short-form chatbot interactions, the Uncanny Valley model as a relationship does not appear here.

This implication is likely not the case, as the visual data does appear to present some form of a relationship, though not statistically significant. Therefore, there may be a relationship present between the two variables that is not showing up in the results. In Figure 13, though the best fit line is relatively flat, most participants that, on average, ranked the likeability of the chatbot as less than neutral (ranking of 3 or below) were in the 40-60% range.

This dip is better visualized in Figure 12, where the mean likeability ranking given to each image is plotted against the image's human likeness. The scatter appears to trend downward be curving upward again at 50% human likeness. Note that for experiment 2, each avatar image on average was only encountered and rated by participants 3 times, so the mean likeability of each image could potentially be heavily affected by outliers. However, the dip still indicates there could potentially be a relationship that did not present within the results.

The second implication is that, during short exposure to a chatbot, message tone and content take precedent in a user's perception over the avatar. This is potentially supported by the results,

as shown in Figure 17. Of the statistically significant correlation found between the likeability of the chatbot by feature and the chatbot's overall likeability, the avatar had the lowest β' estimate and partial R^2. This correlation could potentially indicate that, within the study, the message tone and content affected how participant's perceived the chatbot's likeability more than the chatbot's avatar.

If the message tone and content affect the chatbot's likeability more than the chatbot's avatar, this could reasonably explain why Mori's Uncanny Valley model did not appear to present in the results. As discussed in the results, when standardizing by message content and tone, both best lines were relatively flat with small confidence bands, indicating that the two features were adequately controlled for the study, as intended. This, paired with the message tone and content potentially having the largest impact on the chatbot's perceived likeability, could indicate why a strong relationship did not appear in the results.

The third implication is that there is the possibility that all hypotheses were not rejected due to limitations within the experimental design. The specifics of these limitations will be discussed in the following subsection.

The implications of the results obtained from Experiment 1 are predominately tied to future research using HuLA-50. The potentials of this future research are discussed in Section 8.3.

Aside from these potential future implications, the results of the experiment imply two potential points about the image dataset used during the rankings. As discussed in Section 6 and shown in Figure 9, when unstandardised, the image mean skewed right and no image had a mean human likeness greater than or equal to 90%. This could potentially indicate that images in the dataset do not uniformly present on the human likeness range. However, it could also be due to a limitation of the experimental design, as participants may have been discouraged from ranking images at higher human likenesses.

8.2 Limitations

Limitations within Experiment 1's experimental design mainly come from having a within-participant blind ranking. Each participant was aware that they were ranking all 50 images in the dataset, but did not know what images they could potentially encounter in the future. Therefore, particularly at the start of the ranking, they could potentially be discouraged from assigning extreme human likeness ratings to images. Specifically higher valued extrema like 90% and 100% as shown in the results, as human likeness percentage is a fixed scale with a meaningful zero. Therefore, it is possible that the average person is able to more accurately recognize the absolute absence of anthropomorphisation rather than the totality of it.

Limitations within the experimental design for experiment 2 mainly pertain to the chatbot interaction. While the study tries to best replicate the "character chatbots" discussed previously, there was an inherent level of disconnect intentionally put into the study, as to not deal with any potential ethical or safety issues. As mentioned previously, in a paper by Zhang et al. [34], they found that the majority of their participants (80.33%) engaged in some type of emotional and social support where they actively sought out companionship from chatbots.

However, the majority of these participants engage in more intimate/personal interactions with the chatbot than they would with an acquaintance, which is the relationship the chatbot and participant were encouraged to engage as. In fact, the paper found that approximately 68% of participants had enough of a perception of intimacy with the chatbot that they engaged in

"Romantic and Intimacy Roleplay", where romantic and/or intimate interactions are explored.

Despite this, having the chatbot engage with the participant as if they were more personal friends or partners was still considered inappropriate for a thesis of this scope. Particularly due to the fact that the participants were crowdsourced and, therefore, could not be monitored for potentially concerning interactions.

Other limitations of the experimental design could be from the pre-testing, limitations that most likely influenced the lack of statistical significance found in the analysis. As stated previously, taking into account the feedback of experiment 2's pre-testing phase, UI option 1 was chosen of the two presented to participants (see Figure 4). While this UI did seem to replicate a chatbot that a person would likely encounter in the real world well, the avatar of the chatbot is less visible than in option 2.

This could mean that participants during testing were not focused on the avatar and, therefore, it did not affect their experience or their perception chatbot's likeability. This is supported by participant's answers to the open-ended questions, as many participants stated that they did not pay attention to the avatar heavily or forgot that the avatar was there.

The results of Experiment 1 used for Experiment 2 could have also brought about some limitations. After filtering, there were 33 participants in total whose responses were used for finding the mean human likeness rating of each image in the dataset. As the goal of experiment 1 was to find a more objective than subjective rating for each of the 50 images, it could be argued this is not large enough of a sample size to achieve a mean the majority of the population would agree with. Therefore, the human likeness ratings could be possibly incorrect for the image dataset.

For the results, limitations within the sample of participants include the age of the participants, which trends younger. Of the participants that disclosed their age, there was a range of 19-60 year old participants, however the mean age of participants was 28. This aligns with the age range of users that SurveySwap [29] and SurveyCircle [28] both report use their sites (approximately teens to mid twenties), which is where approximately 110 of the 128 responses came from.

Due to the majority of responses coming from those two sites, it further potentially limits the diversity of the participant sample. This is because the majority of users are posting surveys to the sites themselves and participating in surveys specifically to boost their own survey. This means they likely come from fields that facilitate them collecting participants, e.g. academia or marketing. However, the industries/specialities that these users fall into appear to be extremely diverse, with no field of study on the sites taking up an exceptional majority of the survey content. Gender did not appear to be exceptionally skewed within the sample. Of the participants that disclosed their gender, 63 identified as "female", 47 as "male", and 2 as "non-binary" or "other".

Another potential limitation of the sample that pertains to the population at large is that the characteristics of an individual might affect how they perceive the anthropomorphisation of an agent. In a paper done by Karl F. MacDorman [14], they analyse how certain characteristics of individuals affect "Uncanny Valley sensitivity". They found statistically significant effects on participant's perception of the eeriness/warmth of an agent when controlling for participants: "Neuroticism and Anxiety", an individual exhibits emotional instability, "Animal Reminder Sensitivity", an individual feels disturbed when presented with their mortality and bodily functions, "Human-Robot Uniqueness", an individual categorizes "humans" and "robots" as mutually exclusive categories, "Religious Fundamentalism", individuals considered "fundamentalists" of the Abrahamic religions, "Negative Attitude towards Robots", individuals within inherent dislike to robots.

They also found a perception change when controlling with age and gender, younger individuals

and people who identify as "females" experiencing the "eeriness" of an agent more heavily. This, specifically, is not likely a limitation of this thesis as the results found that there was no statistically significant effect on the correlation between the chatbot's human likeness and likeability when moderating with age or gender. Age on its own could have theoretically impacted the experiment, however, as the results did find a small statistically significant correlation between the participant's age and the chatbot's likeability.

Some of these results, in fact, align with Mori's own analysis within his original Uncanny Valley paper [18], as he proposed the "eeriness" experienced towards certain agents might come from an innate biological self-preservation. The results of the thesis ideally mitigate this by having a large sample size of 128 participants, therefore making it less likely that individual characteristics present as latent variables in the result analysis.

8.3 Further Research

Many of the limitations discussed above were methodological rather than inherent to the research questions and hypothesis, therefore further research could involve modifying the experimental design and re-running the study. Specifically, it could be reasonable to modify the UI so that the avatar is much more prominent. While this may not be realistic in comparison to character chatbots a person would encounter in real life, it would force the participant pay more attention to the avatar. This would be a way to bring out more latent reactions that, in a realistic setting, a user would only experience after engaging with the chatbot for more than 5 minutes. Therefore, the chatbot interaction could retain its short duration while propagating how the general population experiences more long-term exposure to a chatbot avatar with a given human likeness.

A continuation of the experiment could also be done. In relation to the other parts of this study, HuLA-50 has many more applications beyond the research question. Any study pertaining to the visual aspects of the Uncanny Valley or human-agent interactions that involve an avatar could reasonably use HuLA-50. An example of an application of HuLA-50 outside the scope of this thesis could be an expansion of the Visser et al paper [32]. As they found the presence of an avatar increases the play intensity of participants, they could use HuLA-50 as a benchmark dataset in an experiment that analyzes how the human likeness of said avatar affects that change in play intensity.

HuLA-50 itself could also be expanded as a database. With only 33 participants of the total 83 that ranked all 50 images, the sample size could be significantly increased. If more participants contributed to the database, what is ranked by participants could also be expanded. For example, more avatar images could be added to the dataset of images ranked. Furthermore, other attributes outside of human likeness could be ranked, such as the likeability of the avatar itself separated from a chatbot, or unexplored attributes like perceived intelligence, perceived safety, eeriness, etc.. This would provide researchers with a much more robust and significant database that could be used for further research.

Beyond the scope of the current study, further research could include modifying the research question. Specifically, asking the more discrete question of whether the presence of an avatar for a chatbot affects the chatbot's likeability. This was not a question for this specific paper, as the focus was modelling the chatbot so that is replicates "character chatbots", most of which have a profile picture associated with them. However, this type of research question could serve as a better foundational question for more research to be built off of. Other modifications to the research

question could be exploring other traits of the chatbot besides likeability, i.e. the other categories of the Godspeed Questionnaire such as "perceived intelligence" which was used for the distractor questions rather than the experimental questions. If other traits are explored, the Uncanny Valley should likely not be used as a reference model as Mori [18] only described "affinity" and "eeriness" in his original paper.

9 Conclusions

With the rise of character chatbots and "companion AI" for personal use and entertainment comes the rise of chatbots with avatars attached to them. Therefore, the question of how these avatars are anthropomorphised, i.e. their human likeness, affects the likeability of the chatbot appears. Specifically, if that relationship replicates Mori's Uncanny Valley model [18].

This thesis studies this relationship with two main experiments. In experiment 1, participants were provided a set of 50 profile pictures and asked to rank the human likeness of the avatars within the pictures. In experiment 2, a different set of participants interact with a chatbot similar to these "character chatbots" and rate the likeability of the chatbot.

Experiment 1 resulted in HuLA-50, consisting of 83 unique participants' 2339 total rankings. HuLA-50 could be used as a resource for other academic studies, in the same way, for example, it was used for Experiment 2.

Experiment's 2 results found showed there was not strong enough empirical evidence to support the relationship resembling the Uncanny Valley model. With a low R^2 and all three null hypotheses not being rejected, the results indicate there is not a positive cubic relationship between the chatbot avatar's human likeness and the likeability of the chatbot in short chatbot interactions where the participant are encouraged to treat the chatbot as an acquaintance.

Further research could go into modifying the experimental design to mitigate the effect the controlled features of the chatbot have on the participants and better reflect how the avatar of a chatbot might affect the user experience in the long term. Furthermore, the scope of the research question could be modified to ask discretely whether or not the presence of a chatbot avatar affects the user experience or if the human likeness of the chatbot's avatar affects different attributes.

References

[1] Ahmad Al-Dahle. The future of AI: Built with Llama. https://ai.meta.com/blog/future-of-ai-built-with-llama/, 2025.

[2] Christoph Bartneck, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. My robotic doppelgänger – a critical look at the Uncanny Valley. In *The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 269–276, 2009.

[3] Christoph Bartneck, Christian U. Krägeloh, Mohsen Alyami, and Oleg N. Medvedev. Godspeed Questionnaire Series: Translations and usage. In *International Handbook of Behavioral Health Assessment*, pages 1–35. Springer International Publishing, Cham, 2023.

[4] Tyler J. Burleigh, Jordan R. Schoenherr, and Guy L. Lacroix. Does the uncanny valley exist? an empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior*, pages 759–771, 2013.

[5] Aaron Chatterji, Thomas Cunningham, David J. Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use ChatGPT. *National Bureau of Economic Research*, 2025.

[6] Marcus Cheetham, Lingdan Wu, Paul Pauli, and Lutz Jancke. Arousal, valence, and the uncanny valley: psychophysiological and self-report findings. *Frontiers in Psychology*, 6, 2015.

[7] DeepSeek-AI. DeepSeek-V3.2. https://www.deepseek.com/en/, 2026.

[8] Alphabet Inc. Form 8-K. Technical Report 001-37580, U.S. Securities and Exchange Commission, 2025.

[9] JapanDict. 不気味 [bukimi]. In *JapanDict dictionary*, n.d.

[10] Boyoung Kim, Ewart de Visser, and Elizabeth Phillips. Two uncanny valleys: Re-evaluating the uncanny valley across the full spectrum of real-world human-like robots. *Computers in Human Behavior*, 135, 2022.

[11] Jari Kätsyri, Klaus Förger, Meeri Mäkäräinen, and Tapio Takala. A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, Volume 6 - 2015, 2015.

[12] Jari Kätsyri, Meeri Mäkäräinen, and Tapio Takala. Testing the 'uncanny valley' hypothesis in semirealistic computer-animated film characters: An empirical evaluation of natural film stimuli. *International Journal of Human-Computer Studies*, 97:149–161, 2017.

[13] Ning Ma, Ruslana Khynevych, Yunqiang Hao, and Yahui Wang. Effect of anthropomorphism and perceived intelligence in chatbot avatars of visual design on user experience: accounting for perceived empathy and trust. *Frontiers in Computer Science*, 7, 2025.

[14] Karl F. MacDorman and Steven O. Entezari. Individual differences predict sensitivity to the uncanny valley. *Interaction Studies*, 16(2):141–172, 2015.

[15] Magnific AI. Magnific. https://www.magnific.com/, 2026.

[16] Maya B. Mathur and David B. Reichling. Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition*, 146:22–32, 2016.

[17] Masahiro Mori. The Uncanny Valley. *Energy*, 7(4):33–35, 1970.

[18] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. The Uncanny Valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, 2012.

[19] Ollama. Ollama. https://ollama.com, 2026.

[20] OpenAI. How people are using ChatGPT. https://openai.com/index/how-people-are-using-chatgpt/, 2025.

[21] OpenRouter. AI model rankings. https://openrouter.ai/rankings, 2026.

[22] Lukasz Piwek, Lawrie S. McKay, and Frank E. Pollick. Empirical evaluation of the uncanny valley hypothesis fails to confirm the predicted effect of motion. *Cognition*, 130(3):271–277, 2014.

[23] Peter van der Putten. Contribute to research and rate these avatars on human likeness. https://www.linkedin.com/feed/update/urn:li:activity:7453065249311698946/, April 2026.

[24] r/SampleSize Community. r/SampleSize. https://www.reddit.com/r/SampleSize/, 2026.

[25] Tejas R. Shah, Sonal Purohit, Manish Das, and Thavaprakash Arulsivakumar. Do i look real? impact of digital human avatar influencer realism on consumer engagement and attachment. *Journal of Consumer Marketing*, 42(4):416–430, 02 2025.

[26] Marita Skjuve, Ida Maria Haugstveit, Asbjørn Følstad, and Petter Bae Brandtzaeg. Help! is my chatbot falling into the uncanny valley? an empirical study of user experience in human–chatbot interaction. *Human Technology*, 15(1):30–54, 2019.

[27] Stephen Wonchul Song and Mincheol Shin. Uncanny Valley effects on chatbot trust, purchase intention, and adoption intention in the context of e-commerce: The moderating role of avatar familiarity. *International Journal of Human–Computer Interaction*, 40(2):441–456, 2024.

[28] SurveyCircle. SurveyCircle. https://www.surveycircle.com, 2026.

[29] SurveySwap. SurveySwap. https://surveyswap.io, 2026.

[30] Chek Tien Tan, Indriyati Atmosukarto, Budianto Tandianus, Songjia Shen, and Steven Wong. Exploring the impact of avatar representations in AI chatbot tutors on learning experiences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[31] Universiteit Leiden. Ethics Review Committee of the Faculty of Science. https://www.organisatiegids.universiteitleiden.nl/en/faculties-and-institutes/science/committees/ethics-review-committee.

[32] Boele Visser, Peter van der Putten, and Amirhossein Zohrehvand. The impact of AI avatar appearance and disclosure on user motivation. In Chutiporn Anutariya, Marcello M. Bonsangue, Erna Budhiarti-Nababan, and Opim Salim Sitompul, editors, *Data Science and Artificial Intelligence*, pages 142–155, Singapore, 2025. Springer Nature Singapore.

[33] Marc Zao-Sanders. How people are really using AI in 2026. *Harvard Business Review*, 6 2026.

[34] Yutong Zhang, Dora Zhao, Jeffrey T Hancock, Robert Kraut, and Diyi Yang. The rise of AI companions: how human-chatbot relationships influence well-being. *arXiv preprint arXiv:2506.12605*, 2025.

10 Appendix

A Consent Form

Pre-testing Notes*:
For pre-testing, you will be simulating the study, providing feedback, and then given the option to rank the human likeness of the dataset of digital avatars. The study aims to test the correlation between a chatbot's avatar and the user's perception of their likeability. You will be examining, during the simulation, both the general experimental process and how well the nature of the study is obfuscated. A participant should know they're being studied on their perception of the interaction, but not the exact relationship. You do not need to fill in any parts of the questionnaire, as none of your responses in that section will be logged. For feedback, all questions are optional but it's preferred that you write "none" or "nothing" instead of leaving the prompt blank. All other matters of the consent form apply to you as well as a participant, unless it contradicts the above, so please read the rest of this form.
Purpose of the Study:
This study is part of an undergraduate thesis for the Leiden Institute of Advanced Computer Science (LIACS) at Leiden University by CJ Reitter under the supervision of Dr. Peter van der Putten. By participating, you will be interacting with and rating your experience with an LLM-based chatbot. The exact aspects of the interaction being measured will not be disclosed until the end of the study. We ask that you do not disclose the nature of this study after completion to ensure experimental integrity
Procedures:
It is expected that this study will take approximately 15 minutes to complete. You are welcome to stop participating in the study at any time, by simply closing the site.
The study is comprised of 2 phases. In the first, you will be presented with a chatbot powered by DeepSeek v3.2. It is prompted to communicate with the user as if it were making small talk with a friendly peer or acquaintance, and you are encouraged to communicate with it in a similar fashion. There is no expected minimum or maximum prompt count on the user or agent's end, so you may interact with it at whatever pace you feel comfortable with. You will remain on the page for 5 minutes before you are immediately redirected to the second phase of the study. This will happen immediately after 5 minutes, so keep that in mind during the interaction. Note that the chatbot might take time to respond, so please be patient and do not refresh while you are on this page.
In the second phase, you will be directed to a form where you will be asked to answer both rating-based and open-ended questions on your experience with the chatbot and your interactions. The rating-based questions are mandatory to complete the study, but you may exit the study if you do not wish to answer them all. Partial responses will not be recorded. All open-ended questions are optional and filling them out is not required to complete the study. You will also be asked a few demographic questions (age and gender), which are also optional to disclose.

After submitting, you will be taken to a page disclosing the exact nature of the study and the scales being measured. We ask that you don't share the contents and nature of the experiment until 1 July 2026.

Potential Concerns:

We recognize that LLM-based chatbots can imitate human-like conversation extremely well and have the potential to appear more intimate than they are. We would like everyone participating in the exam to treat their conversation with care and be mindful of what they share with the chatbot, working under the assumption that a stranger could read their chats. Everything the chatbot generates is made up and should not act as a substitute for human-to-human interaction. If at any point you feel uncomfortable or unfit to continue communicating with the chatbot, you have the right to immediately leave the study.

Confidentiality:

We recognize that LLM-based chatbots can imitate human-like conversation extremely well and have the potential to appear more intimate than they are. We would like everyone participating in the exam to treat their conversation with care and be mindful of what they share with the chatbot, working under the assumption that a stranger could read their chats. Should you ever feel uncomfortable during your interaction with the chatbot, you are always welcome to leave the study. Everything the chatbot generates is made up and should not act as a substitute for human-to-human interaction. The chatbot runs on the DeepSeek v3.2 model on a local server using Ollama. Ourselves, Ollama, and DeepSeek will not log the contents of the responses or prompts made while communicating with the chatbot, but token counts are recorded by Ollama, and prompt counts, as in the number of messages communicated with the chatbot, are recorded by ourselves and Ollama.

Voluntary Participation:

Your participation is completely voluntary. You are under no obligation to complete the survey, and you may withdraw at any time without any penalties. If you choose to exit the survey, you will not be required to provide any justification, and your answers will automatically be deleted. Responses to the questions asked in phase 2 and any other anonymized data collected throughout the study is only retained if you, the participant, explicitly submit the survey at the end of phase 2.

Contact Info:

If you have any questions or concerns regarding the study or its findings feel free to contact either the researcher: s3794490@leidenuniv.nl or the thesis supervisor: p.w.h.van.der.putten@liacs.leidenuniv.nl.

Figure A: Consent Form all participants read and digitally signed before participating in the study. *: Pre-testing participants simulating the study were the only participants to view this section.

B Pre-testing Questionnaire

Rating Questions (Mandatory)

Rate your experience based on how much you agree/disagree with the following statements. (1 = Strongly Disagree, 5 = Strongly Agree)

The chatbot's responses were likeable.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

The chatbot generally seemed nice.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

The chatbot had a friendly-looking avatar.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Interacting with the chatbot was a pleasant expereience.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

The chatbot's avatar was pleasant.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

The chatbot's avatar looked like a nice person/agent.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

The chatbot's responses were friendly.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

The chatbot's avatar appeared likeable.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

The chatbot seemed kind.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

The chatbot had a kind-looking avatar.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Open-Ended Questions (Optional)

What was your general impression of the chatbot? (Optional)

Enter text here...

What was your experience with the interface of the chatbot? (Optional)

Enter text here...

How did the digital avatar attached to the interface improve/worsen your experience? (Optional)

Enter text here...

How did the nature/manner of speaking of the chabot affect your experience? (Optional)

Enter text here...

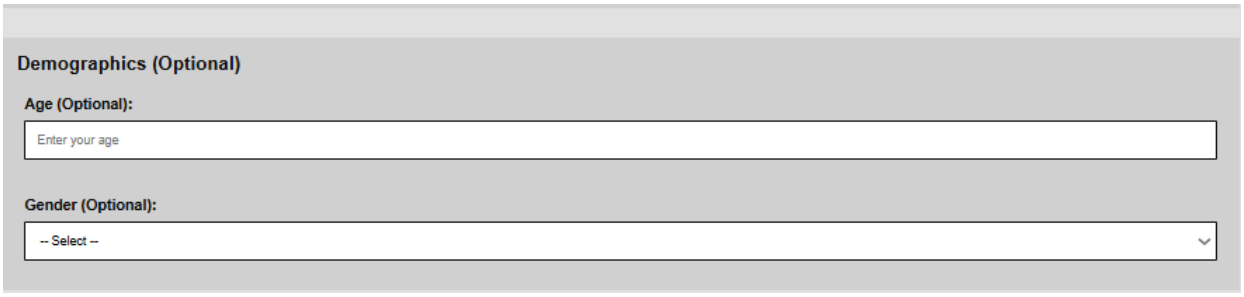

Figure B: Questionnaire participants were given during experiment 2's pre-testing. Questions related to the avatar's likeability were the experimental questions and all others (i.e. questions asking about the chatbot's perceived intelligence/other features of the chatbot) were distractor questions.

C Testing Questionnaire

Rating Questions (Mandatory)

Rate your experience based on how much you agree/disagree with the following statements, based on the following features of the interaction. (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, & 5 = Strongly Agree)

Overall Chatbot Interaction

I thought the chatbot seemed sensible.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I thought the chatbot was intelligent.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I felt the chatbot seemed like a responsible agent.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I felt the chatbot was pleasant to be with.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I liked the chatbot.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Chatbot Message Content

I thought the chatbot had competently written messages.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I thought the chatbot asked me friendly questions.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I thought the chatbot had knowledgeable responses.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I liked how the chatbot responded to my prompts.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I thought the chatbot's messages were written intelligently
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Chatbot Message Tone

I thought the way the chatbot spoke was likeable.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I felt the chatbot's tone was intelligent.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I thought the chatbot had a friendly attitude.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I thought the chatbot's tone was pleasant.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I found the way the chatbot wrote to be sensible.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Chatbot Avatar

I thought the avatar seemed friendly.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I felt the avatar looked intelligent.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I thought the chatbot had a kind-looking avatar.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I thought the avatar seemed like a pleasant agent.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

I liked the chatbot's avatar.
○ 1 ○ 2 ○ 3 ○ 4 ○ 5

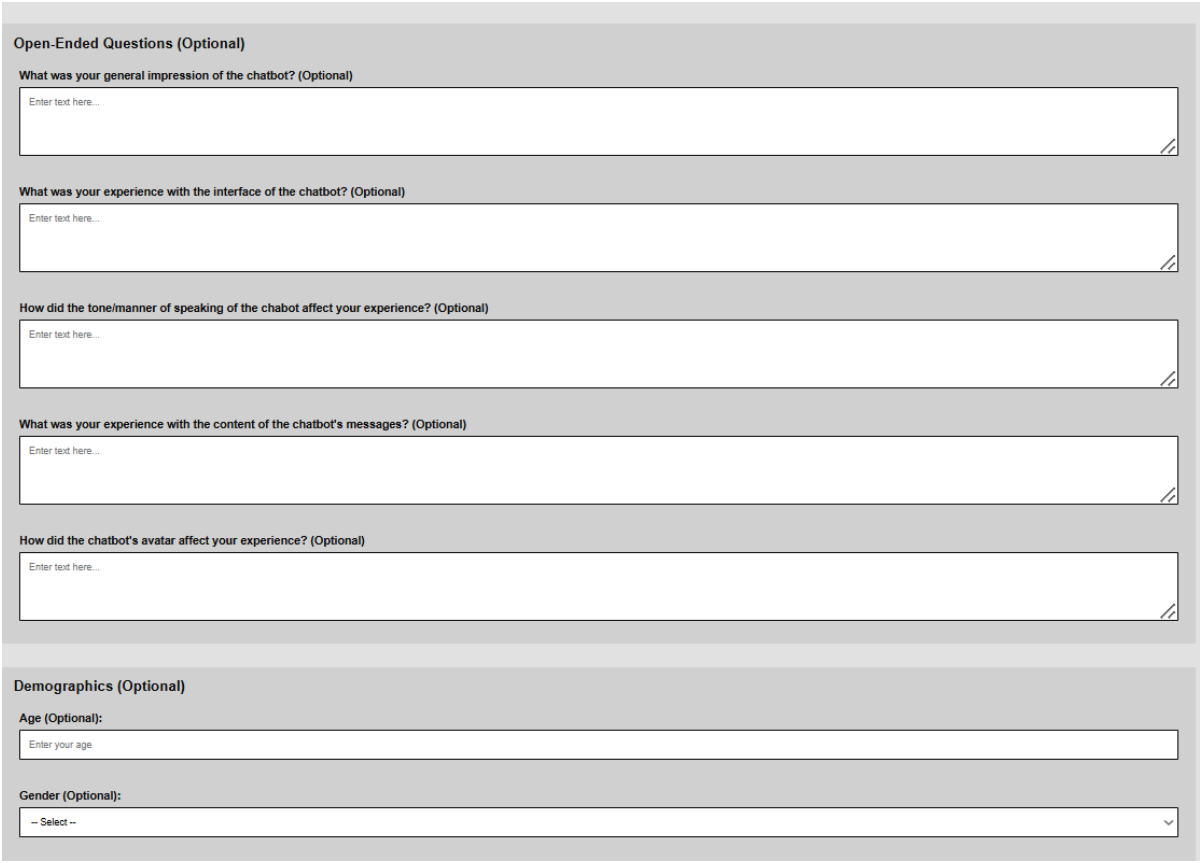

Figure C: Questionnaire participants were given during experiment 2. Questions related to the avatar's likeability were the experimental questions and all others (i.e. questions asking about the chatbot's perceived intelligence/other features of the chatbot) were distractor questions.