The Uncanny Valley: An Empirical Study on Human Perceptions of AI-Generated Text and Images

by

Deepali Kishnani

Bachelor of Technology (Honors) in Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi (IIIT-Delhi) (2016)

Submitted to the
System Design and Management Program and
Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degrees of
MASTER OF SCIENCE IN ENGINEERING & MANAGEMENT
MASTER OF SCIENCE IN ELECTRICAL ENGINEERING & COMPUTER SCIENCE

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025

Authored By: **Deepali Kishnani**
System Design and Management Program
Department of Electrical Engineering and Computer Science
January 15, 2025

Certified By: **Juanjuan Zhang**
John D. C. Little Professor of Marketing, MIT Sloan School of Management
Thesis Supervisor

Certified By: **Manish Raghavan**
Drew Houston (2005) Career Development Professor, Electrical Engineering & Computer Science
Thesis Reader

Accepted By: **Joan Rubin**
Executive Director, System Design & Management Program

Accepted By: **Leslie A. Kolodziejski**
Professor of Electrical Engineering and Computer Science
Chair, Department Committee for Graduate Students

The Uncanny Valley: An Empirical Study on Human Perceptions of AI-Generated Text and Images

by

Deepali Kishnani

Submitted to the
System Design and Management Program and
Department of Electrical Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degrees of
MASTER OF SCIENCE IN ENGINEERING & MANAGEMENT
MASTER OF SCIENCE IN ELECTRICAL ENGINEERING & COMPUTER SCIENCE

ABSTRACT

This thesis explores how the uncanny valley phenomenon—historically tied to near-human robots—applies to text-based AI interactions and AI-generated images. While the concept has been predominantly studied in the context of robotics, the advent of generative AI reveals that text and visuals that are 'almost, but not quite' human can also provoke unease.

Two experiments structure the study. The first examines GPT4-Turbo (GPT4o) text conversations. Sixty participants engaged with one of three “chatbots”: an “Uncanny-Valley Bot” (prompt engineered to fall in the uncanny valley), a “Human-Like Bot” (prompt engineered to converse like humans), or a human control. Godspeed Questionnaire results indicate that the uncanny valley effect surfaces in text-only form: participants consistently rated the “Uncanny-Valley Bot” lowest in anthropomorphism, animacy, likeability, and perceived intelligence.

The second experiment investigates AI-generated images produced by Stable Diffusion XL at varying degrees of realism. Fifty-six participants ranked each image’s “strangeness,” revealing that highly realistic or clearly stylized outputs raise fewer concerns. By contrast, images that inhabit the uncanny valley elicited discomfort. To quantify these findings, recognized metrics like Frechet Inception Distance (FID) and Kernel Inception Distance (KID) were used to compare real and AI-generated images. Both metrics strongly correlated with human perceptions, suggesting that distance metrics can be used to determine realism. The study also shows that image generation models can detect visual features associated with the uncanny valley. However, performance drops when the prompt calls for subtle, “mid-range” realism, indicating the model’s difficulty in maintaining comfort and believability at intermediate levels.

Collectively, the two experiments confirm that uncanny valley responses are not confined to physical robots but persist in text-based dialogue and AI-synthesized images. Yet challenges remain. Short interaction windows, small participant samples, and reliance on selected AI models call for studies on the generalizability of these findings. Future work should adopt longitudinal designs, larger samples, and multiple AI systems. Addressing the uncanny valley in both textual and visual content is essential for advancing user trust, and comfort in AI.

Thesis Supervisor: Juanjuan Zhang
Title: John D. C. Little Professor of Marketing, MIT Sloan School of Management

Thesis Reader: Manish Raghavan
Title: Drew Houston (2005) Career Development Professor, EECS

Acknowledgement

I would like to express my deepest gratitude to my advisor, Prof. Juanjuan Zhang, for her invaluable guidance, unwavering support, and encouragement. Her insightful feedback has been instrumental in shaping this thesis. I am especially grateful for her encouragement to explore topics that truly resonated with my interests, helping me discover my passion at the intersection of Psychology and AI.

I extend my heartfelt thanks to my thesis reader, Prof. Manish Raghavan, for his invaluable guidance, patience, unwavering support, and encouragement. His thoughtful feedback has been pivotal in refining and strengthening the experiments. His mentorship has been both empowering and inspiring, and I am sincerely thankful for his guidance and for taking a bet on me.

I also want to thank Prof. Donna Rhodes, for being my faculty sponsor for COUHES approval. Grateful to my friend, Rowell Castro, for collaborating with me on Experiment 2 as part of the 6.7960 Deep Learning project.

I am deeply thankful to the sixty participants, who generously volunteered, shared their data and took the time to support Experiment 1 and to the survey respondents from Experiment 2.

I extend my heartfelt thanks to Joan Rubin, Executive Director at SDM, my colleagues and friends in the System Design & Management Program at MIT for their constant support, collaborative spirit, and encouragement.

To my family, I owe my deepest gratitude for their unwavering support and love. My mom, Bimla Kishnani, has been my biggest cheerleader, and the most instrumental source of strength. To my other family members—Kishore Ramani, Ankita Ramani, Sonal Kishnani, Lakshya Ramani, and Pallavi Matai —thank you for being my rock throughout my journey at MIT.

To my wonderful friends—Sharut Gupta, Abhishek Mukherjee, Vinayak Chopra, Chaitanya Kumar, Saumya Nagpal, Anup Sreekumar, Dominic Co, and Samuel Gomez—thank you for lifting my spirits, and for consistently supporting and showing up for me during my time at MIT.

This thesis is the culmination of the contributions, guidance, and support of many people, and I am sincerely grateful to everyone who has been part of this journey.

Note on the use of Generative AI: *I have used GPT4o to enhance the clarity and coherence of explanations by editing spelling, grammatical, and style inconsistencies. Additionally, GPT4o was used selectively as a tool for analysis.*

List of Contents

List of Figures

List of Tables

Chapter 1

Introduction

What does it mean to be *human-like*? Is it the ability to mimic human characteristics, to emulate human expressions, or to reflect human emotions? Is it merely the sum of observable traits, or is it something that cannot be codified or replicated? These timeless questions have captivated philosophers, scientists, and thinkers for centuries. The answer often depends on who you ask. A biologist might define humanity through the lens of Homo sapiens, grounded in genetics and evolution, while a philosopher might point to self-awareness, consciousness, and the capacity for abstract thought. Such varied perspectives underscore the inherent complexity of defining "humanness"—a question that stands at the intersection of science, philosophy, and our pursuit of self-knowledge.

In a world where we still lack a complete understanding of what it means to be human-like, we now face an audacious challenge: building our replicas. We strive to craft machines that not only think but also *feel,* that not only respond but also understand. Yet, as we inch closer to this goal, we are confronted with a paradox– The closer our creations come to resembling us, the more they disturb us. This phenomenon, known as the *uncanny valley effect*, exposes the fragile boundary between familiarity and alienation [1].

Traditionally, the uncanny valley has been studied in the context of robotics and computer-generated imagery, where physical appearance plays a pivotal role. However, generative AI is rapidly transforming modern life by automating creativity, enhancing productivity, and reshaping entire industries. It empowers tools that generate human-like text, images, and videos, thereby revolutionizing fields including marketing, healthcare, entertainment, and education. Against this backdrop, recent advancements in generative modeling suggest that the uncanny valley may extend beyond visuals and into text-based interactions and other AI-generated outputs [90].

Despite the growing body of literature on human-computer interaction, relatively few studies have comprehensively examined how users react to AI-generated text conversations or semi-realistic images that trigger discomfort. Understanding the implications of this response is especially significant given the increasing integration of AI systems in daily life. By situating this research within the broader context of human-like AI technologies, we can begin to formulate evidence-based strategies for crafting more comfortable, human-centered interactions, enabling AI systems to better serve real-world applications without unsettling their users.

1.1 Motivation

While the uncanny valley effect has been well-documented in visually oriented media, its impact on text-based conversations and AI-generated images remains less understood. This research addresses two key gaps:

1. Text-Based Uncanny Valley: A need for deeper insight into whether users react negatively to near-human, text-based interactions.
2. Image-Based Uncanny Valley: A lack of refined understanding about the metrics, techniques, and user perceptions that govern image generation.

By focusing on these gaps, the study contributes to a more holistic understanding of the uncanny valley, informing both theoretical perspectives and real-world applications in fields such as customer service, healthcare, education, and entertainment.

1.2 Objectives

The main goal of this thesis is to explore and quantify the uncanny valley effect across text-based AI interactions and AI-generated images. Specifically, it aims to:

1. Identify whether the concept of uncanny valley, traditionally reserved for human-robot interaction, applies to generative AI content such as text and images.

2. Examine the correlation between user perceptions (such as anthropomorphism, likeability, and trust) via surveys and system features (like conversational style, realism, or intentional imperfections).

1.3 Research Questions

To achieve these objectives, the study is guided by the following research questions:

1. Do users perceive and respond differently to varying levels of human-likeness in text-based interactions with AI-driven chatbots?
2. Do conversation style, emotional resonance, and conversational dynamics play a role in triggering or alleviating the uncanny valley in text-based interactions?
3. Does the level of realism in AI-generated images influence user perception?
4. Can established metrics (e.g., Frechet Inception Distance (FID)) reliably predict uncanny reactions?

1.4 Scope and Delimitations

This research primarily focuses on AI systems that employ large language models for text-based conversation and generative models for image creation. The study involves controlled experiments measuring user perceptions of different AI-driven interactions and image outputs. The study is restricted to only two popular Generative AI models, GPT4-Turbo (GPT4o) for testing uncanny valley in text and Stable Diffusion XL for testing uncanniness in images. Although it touches on applications in various fields, it does not cover every possible AI modality (e.g., voice assistants or virtual reality avatars) due to resource constraints. Additionally, the cultural, linguistic, and contextual factors influencing user responses are acknowledged but not exhaustively explored in this thesis. These delimitations aim to keep the research both focused and feasible, while providing a foundation for future exploration.

1.5 Significance of the Study

By bridging the gap between text-based and image-based uncanny valley effects, this thesis contributes insights into the design and deployment of human-like AI technologies. The findings hold potential applications in:

1. Product Development: Equipping designers with guidelines to develop AI chatbots and generative models that evoke user comfort rather than discomfort.
2. User Experience: Improving engagement and trust in consumer-facing AI platforms.
3. Research and Innovation: Serving as a springboard for subsequent investigations into cross-modal AI design, user psychology, and emerging technologies.

1.6 Methodological Overview

This thesis adopts a mixed-methods approach to capture both quantitative and qualitative aspects of user interactions. Participants will engage with AI chatbots of varying degrees of human-likeness in a controlled experimental setting, followed by structured surveys and, where relevant, interviews. For image-based experiments, validated metrics such as the Frechet Inception Distance (FID) or Kernel Inception Distance (KID) will be used alongside user ratings to assess the uncanny valley effect. Detailed methodology, including sampling strategies, data collection procedures, and analytical tools, will be presented in the subsequent chapters.

1.7 Structure of the Thesis

The remainder of this thesis is organized into five chapters:

1. Chapter 2: Literature Review – Examines the existing body of work on the uncanny valley, evolution of Generative AI models - Large Language Models (LLMs) and Image Generation Models, key studies on text-based chatbot design, and AI-driven image generation in the context of uncanny valley effect.

2. Chapter 3: Experiment 1 (Text-Based Interactions) – Explains the Methodology for Experiment 1 and presents findings related to user perceptions and engagement with AI chatbots.
3. Chapter 4: Experiment 2 (AI-Generated Images) – Explains the Methodology for Experiment 2 and presents results pertaining to the existence of uncanny valley in generative AI and correlation with distance metrics.
4. Chapter 5: Key Findings, Discussion, and Limitations– Integrates insights from both experiments, contextualizes the findings within existing literature, and discusses the limitations of the experiments.
5. Chapter 6: Conclusion and Future Work – Summarizes the key takeaways, acknowledges limitations, and proposes directions for future research.

1.8 Limitations and Assumptions

Several assumptions guide this research, including the premise that participants' subjective responses reliably indicate their underlying perceptions of AI-human likeness. As with any experimental design, results may be influenced by the specific AI models chosen, participant demographics, and the controlled laboratory context. These factors are revisited in later chapters to contextualize the findings.

By integrating these elements into a coherent investigation, the thesis aims to demonstrate how generative AI can approximate human-like interactions while highlighting the perceptual boundaries that give rise to the uncanny valley.

Chapter 2

Literature Review

The rapid advancement of generative artificial intelligence (AI) has brought significant transformations to human-computer interactions, especially through large language models (LLMs) and generative image models. These technologies have blurred the boundary between human and machine, creating both opportunities and challenges in designing AI-driven products and services. However, the perception of AI-generated outputs often falls into what is known as the uncanny valley — a phenomenon where entities that appear almost human, but not quite, evoke feelings of eeriness and discomfort [1]. While this theory has been extensively studied in the context of robotics and visual avatars [2], [3], limited research has explored how it manifests in text-based interactions [4] or in AI generated images.

The uncanny valley phenomenon has traditionally been associated with visual realism and human-likeness in avatars and robots, where users' perception shifts from empathy to unease as the entity becomes more human-like but fails to achieve perfect realism [5]. However, the increasing sophistication of LLMs like GPT-4 [6], has introduced a new frontier for the uncanny valley in text-based human-computer interactions, raising questions about the boundaries of human-likeness in conversational agents. This has practical implications, especially in customer service, marketing, and healthcare, where human-like chatbots are increasingly deployed to engage users [8] [9].

In parallel, advancements in generative image models such as DALL-E and MidJourney have pushed the boundaries of visual realism in AI-generated images [10] [11]. These models can produce highly realistic human-like images, but they also risk triggering the uncanny valley effect when certain features appear unnatural or unsettling [12]. While the uncanny valley has been extensively studied in robotics and CGI [13], its impact on AI-generated images remains relatively underexplored, particularly in terms of human perception and emotional response.

The purpose of this literature review is to explore the existing body of knowledge on evolution of the large language models (LLMs), Generative AI models, the uncanny valley theory, and its applications in textual conversations and image generation models. The review identifies gaps in the literature and positions this thesis within the broader research landscape. The literature review is organized as follows:

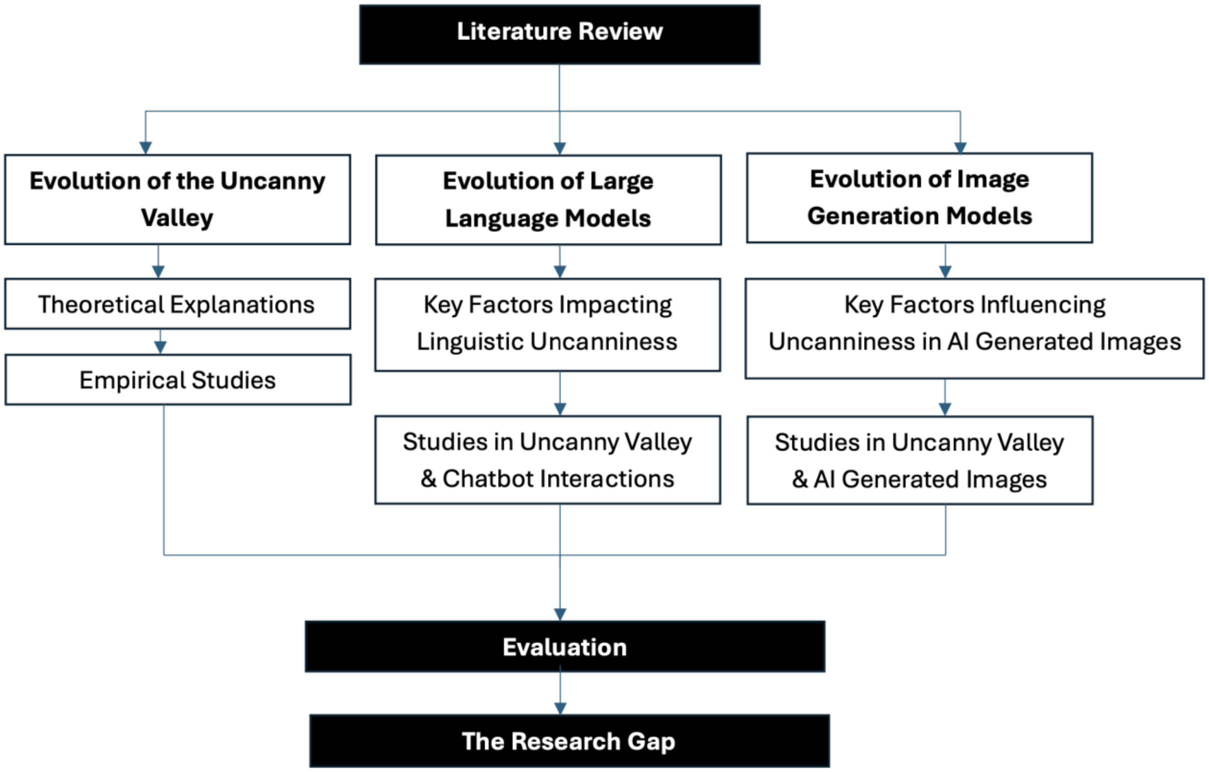

Literature Review
Evolution of the Uncanny Valley
Theoretical Explanations
Empirical Studies
Evolution of Large Language Models
Key Factors Impacting Linguistic Uncanniness
Studies in Uncanny Valley & Chatbot Interactions
Evolution of Image Generation Models
Key Factors Influencing Uncanniness in AI Generated Images
Studies in Uncanny Valley & AI Generated Images
Evaluation
The Research Gap

Figure 1: Organization of the Literature Review

2.1 Evolution of the Uncanny Valley

The concept of the uncanny valley was first introduced by Japanese roboticist Masahiro Mori in 1970 [1]. He theorized that as robots become more human-like in appearance and behavior, people's emotional response to them becomes more positive. However, there is a point where the resemblance is close, but not perfect, causing an emotional dip, which Mori termed the "uncanny valley" [1]. This phenomenon has since become a pivotal concept in robotics, animation, and human-computer interaction (HCI).

Masahiro Mori's original hypothesis [1] was based on his observations of human reactions to prosthetic limbs and humanoid robots. Mori illustrated this relationship with a graph, showing that emotional responses become more positive as human-likeness increases, but sharply decline when the resemblance is almost, but not quite, human.

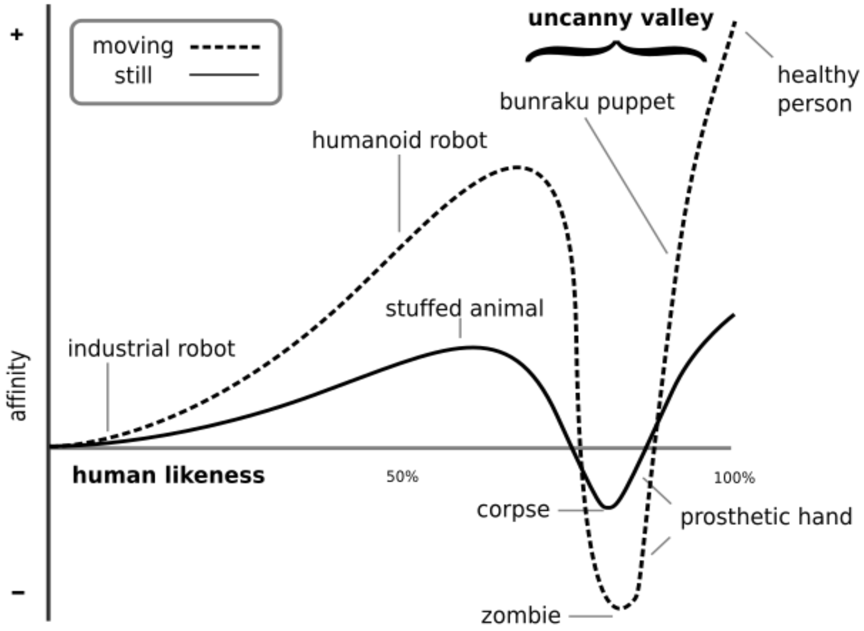


Figure 2: The Uncanny Valley graph theorized by Masahiro Mori. (Source: Wikipedia, accessed January 21, 2025)

Following Mori's work, research into the uncanny valley was limited until the early 2000s, when advancements in robotics and computer graphics reignited interest in the phenomenon. Scholars like Karl F. MacDorman and Hiroshi Ishiguro [2] contributed significantly to the field by exploring the psychological and perceptual aspects of the uncanny valley.

2.1.1 Theoretical Explanations

Several theories have been proposed to explain the uncanny valley effect, each offering a different perspective on why humans react negatively to near-human entities.

Evolutionary Psychology

Evolutionary psychologists attribute the uncanny valley to survival mechanisms rooted in human evolution. Nesse [31] posits that near-human entities may trigger a sense of unease because they evoke characteristics of diseased or deceased individuals. This “disease avoidance” hypothesis suggests that humans have evolved a natural aversion to potential threats to their health, leading to feelings of discomfort when encountering entities that appear almost, but not entirely, human. Such reactions may serve as a protective mechanism to ensure survival by minimizing exposure to potential pathogens.

Perceptual Mismatch

Another theoretical explanation focuses on the concept of perceptual mismatch. Saygin et al. [32] argue that the uncanny valley emerges when there are inconsistencies between visual and behavioral cues. For instance, a robot that appears human but moves in a stiff, mechanical manner creates a discrepancy between its appearance and its motion. This incongruence disrupts an observer’s expectations, resulting in a sense of unease or discomfort. Perceptual mismatch theory highlights the importance of aligning visual and behavioral attributes to reduce the uncanny effect.

Categorical Ambiguity

MacDorman and Ishiguro [2] offer a cognitive explanation, suggesting that the uncanny valley arises from categorical ambiguity. According to their theory, humanoid entities that cannot be distinctly categorized as either human or non-human provoke cognitive dissonance. This ambiguity challenges the observer's ability to make clear distinctions between categories, leading to an eerie sensation. The theory underscores the role of human cognition in interpreting ambiguous stimuli and its contribution to the uncanny valley effect.

Violation of Social Norms

The violation of social norms theory, proposed by Tinwell et al. [33], suggests that the uncanny valley effect occurs when humanoid entities deviate from expected social behaviors. For example, inappropriate eye contact, unnatural facial expressions, or awkward gestures can make a humanoid entity appear unsettling. These violations disrupt social expectations, which are critical to interpreting human-like interactions, thereby heightening the feeling of eeriness.

2.1.2 Empirical Studies

Empirical research has played a crucial role in validating the uncanny valley hypothesis. Various studies have examined the emotional responses elicited by human-like robots, virtual characters, and AI-generated images.

Behavioral Studies

Behavioral studies have been instrumental in exploring the emotional and cognitive responses associated with the uncanny valley. Researchers have investigated the relationship between realism, eeriness, and human-likeness in both static and dynamic stimuli, such as images and animations.

Katsyri et al. [34] conducted a meta-analysis of behavioral studies focusing on the uncanny valley effect. They analyzed the emotional reactions to various human-like robots and virtual characters. Their findings suggest that both static images and dynamic stimuli, such as facial expressions and body movements, can trigger uncanny valley responses when they appear unnatural. They observed that inconsistent facial features, unnatural eye movements, and delayed emotional expressions are key triggers of discomfort. The study also found that cultural factors and prior exposure to human-like entities play a role in the intensity of the uncanny valley effect. Katsyri et al. argued that individuals familiar with humanoid robots may have a reduced uncanny valley response compared to those with limited exposure [34].

Seyama and Nagayama [35] conducted an empirical study focusing on facial proportions and animation quality in 3D-rendered human faces. They found that small deviations from human norms in facial features and movements increase eeriness, particularly when facial symmetry and eye movements are distorted. Their study suggested that even minor imperfections in human-like characters can trigger discomfort, emphasizing the sensitivity of human perception to realism in facial expressions.

Another critical behavioral study by Kätsyri et al. [36] examined the perception of human likeness across different age groups. Their results revealed that younger participants are more prone to experiencing the uncanny valley, possibly due to higher exposure to digital media and hyper-realistic virtual characters.

Virtual Reality and Avatars

The uncanny valley effect has been extensively studied in the context of virtual reality (VR) and avatars used in gaming, social interaction, and telepresence. The design of digital humans and avatars plays a crucial role in determining the user’s experience and engagement in virtual environments.

Wang et al. [37] explored the uncanny valley effect in VR environments by conducting a study on hyper-realistic avatars. Participants interacted with avatars that varied in realism, from stylized cartoon-like characters to photorealistic digital humans. The study found that avatars with a high degree of realism often elicited lower levels of engagement and social presence due to perceived eeriness, particularly when their facial movements and expressions did not align with user expectations.

Their research emphasized the importance of balancing realism and stylization to avoid the uncanny valley effect in VR. Wang et al. suggested that avatars with slightly exaggerated features (e.g., stylized facial expressions or cartoonish designs) are more positively received by users, as they avoid triggering the expectation mismatch that occurs with hyper-realistic avatars [37].

Stein and Ohler [38] studied telepresence and digital doubles in immersive VR environments. They found that digital doubles (photorealistic avatars of real people) often fall into the uncanny valley when subtle facial expressions or voice inflections do not match perfectly with the real-world counterpart. Their study highlighted that temporal misalignment in voice and facial movements exacerbates the uncanny valley effect, causing discomfort and disengagement in users.

Generative AI

With the rise of deep learning technologies and generative AI models, researchers have begun to investigate the uncanny valley effect in synthetic media, such as deepfakes and AI-generated images. These studies focus on how hyper-realistic but flawed digital content impacts human perception and emotional response.

Karras et al. [39] conducted an empirical study on AI-generated human faces using StyleGAN, a popular generative adversarial network (GAN). Their research revealed that AI-generated faces, even when appearing hyper-realistic, often fall into the uncanny valley when subtle imperfections are present. For instance, faces with unnatural lighting, asymmetrical features, or inconsistent textures caused participants to report discomfort and eeriness.

The study also found that viewers are more likely to detect flaws in dynamic content, such as deepfake videos, compared to static images. Karras et al. concluded that temporal coherence (the consistency of facial features over time) is critical in avoiding the uncanny valley in AI-generated videos.

In another study, Mäkäräinen et al. [40] explored the uncanny valley effect in synthetic voices generated by text-to-speech (TTS) systems. They found that voices that closely mimic human speech patterns, but lack natural emotional intonation or introduce minor artifacts, are perceived as unnerving. Their findings suggest that emotional congruence in both visual and auditory stimuli is essential in mitigating the uncanny valley effect.

Further, deepfake detection systems have been developed to identify synthetic media that triggers the uncanny valley effect. Zhang et al. [41] proposed a deep learning-based model to detect unnatural facial expressions and movements in deep fake videos, which are often the source of discomfort in viewers. Their study revealed that deep learning models trained on large datasets of human expressions can successfully predict uncanny valley responses by identifying incongruencies in facial features and behaviors.

Cognitive and Neurological Studies

Recent empirical studies have also examined the cognitive and neurological underpinnings of the uncanny valley effect. Research in this area focuses on how the brain processes human-like entities and identifies cognitive dissonance caused by near-human but flawed stimuli.

Saygin et al. [42] conducted functional MRI (fMRI) scans on participants interacting with humanoid robots and avatars. Their results showed that human brain activity increases in the areas associated with facial recognition and emotional processing when encountering near-human entities that fall into the uncanny valley. The study suggested that mismatched cues, such as unusual eye movements or unnatural skin textures, trigger neural responses associated with threat detection and discomfort.

Further studies by Cheetham et al. [43] explored cognitive processing during exposure to uncanny stimuli. Their findings indicated that increased cognitive load is required to process near-human entities, which leads to feelings of eeriness and discomfort.

2.2 Evolution of Large Language Models

Building on the foundation of evolution of uncanny valley effect in the previous section, it is essential to examine the evolution of large language models (LLMs) to return to the original question of the thesis of whether AI-generated text, like other near-human artifacts, can elicit the uncanny valley effect.

Early NLP Systems and Limited Human-Likeness

The earliest attempts at building conversational agents in Natural Language Processing (NLP) began with rule-based systems like ELIZA [14], a program designed to simulate human conversation through predefined scripts. ELIZA's interaction style was simplistic and largely formulaic, relying on pattern matching rather than genuine language understanding. Despite its rudimentary nature, the system surprised users by eliciting emotional responses, particularly in its "therapist" mode, which gave the illusion of human empathy [14].

As NLP evolved, researchers sought to move beyond scripted rule-based systems by introducing statistical language models, including n-gram models [15] and Hidden Markov Models (HMMs) [16], [17]. N-gram models used probabilistic sequences of words to predict the next word in a sentence, improving contextual relevance in generated text [15]. Similarly, HMMs allowed systems to account for the hidden states underlying sequential data, making language models more capable of handling variability in speech and text [16].

While these statistical models improved the fluency and coherence of text generation, they remained mechanical and predictable. These models could process language effectively but did not mimic human intelligence or emotional nuance. Therefore, users approached interactions with these systems with lower expectations, preventing the eerie sensation associated with human-like failures. The inability of these models to exhibit human traits like empathy or emotional depth made their behavior distinguishable from human interaction, perhaps reducing the risk of uncanny valley effects.

In contrast to modern Large Language Models (LLMs), which aim to achieve human-level conversational abilities, early NLP systems focused primarily on syntactic accuracy rather than semantic understanding.

Introduction of Neural Networks and Human-Like Text Generation

The transition from statistical models to neural networks marked a significant leap in generating human-like text. Early efforts in Recurrent Neural Networks (RNNs), introduced by David Rumelhart, Geoffrey Hinton, and Ronald J. Williams [18], allowed models to process sequential data by maintaining a hidden state that updated over time. These models were capable of generating more contextually coherent text compared to statistical approaches, though they struggled with retaining long-term dependencies, a limitation commonly known as the vanishing gradient problem [19].

To address this issue, Long Short-Term Memory (LSTM) networks were introduced [20]. LSTMs improved the ability of neural networks to remember important information over longer sequences, making them more effective for tasks like language modeling and text generation. These advancements enabled more coherent text generation across longer conversations, bringing conversational agents closer to producing human-like dialogue.

Despite these improvements, early neural networks still struggled to achieve human-level creativity and nuance. RNNs could generate text that mimicked human writing patterns, but the output often lacked semantic coherence and contextual depth over extended conversations [21]. These shortcomings prevented RNNs and LSTMs from crossing into human likeness.

RNN-based language models could significantly improve text prediction and generation [22]. As models became more sophisticated, users began to perceive conversational agents as more human-like, raising their expectations for natural interactions.

Transformers and the Rapid Progress of LLMs

The introduction of the Transformer architecture was a pivotal moment in the evolution of Large Language Models (LLMs). Proposed by Vaswani et al. [22], the Transformer model introduced a self-attention mechanism that enabled models to handle long-range dependencies efficiently and process sequences in parallel, significantly improving the scalability and performance of language models. This breakthrough paved the way for the development of BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. [23], which achieved state-of-the-art performance on various Natural Language Processing (NLP) tasks by leveraging bidirectional context. Similarly, GPT-2 (Generative Pre-trained Transformer 2) by Radford et al. [24] demonstrated the capabilities of unidirectional transformers in generating fluent, coherent, and creative text across a wide range of applications. These advancements in LLMs brought AI-generated text closer to human-like interactions, making conversational agents appear more lifelike.

Scaling of LLMs and the Uncanny Valley

The scaling of Large Language Models (LLMs) from GPT-2 to GPT-3 and beyond has significantly improved their ability to generate human-like responses. GPT-3, introduced by Brown et al. [25], features 175 billion parameters and is capable of performing various tasks, including storytelling, summarization, and question answering, often without the need for task-specific fine-tuning.

Building on this progress, GPT-4 was introduced as a multimodal model capable of processing both text and images. While OpenAI has not disclosed the exact number of parameters for GPT-4, its improvements in language understanding, coherence, and task performance are evident across a variety of applications [26]. GPT-4 has demonstrated remarkable advancements in handling nuanced queries, including legal, medical, and technical domains, which were challenging for previous versions [26].

The eeriness arises when models produce responses that appear human-like at first glance, but lack authentic emotional depth or contextual accuracy. However, key challenges persist, particularly in improving conversational depth and tailoring responses to individual users [27]. One of the key challenges introduced by GPT-4 is the rise of synthetic empathy, where the model attempts to express emotional responses or empathetic reactions [28]. A 2024 study finds that LLMs exhibit human-like personalities derived from their pre-trained data and instruction-based fine-tuning [29]. The findings highlight how human-like traits in LLMs can contribute to uncanny valley effects in text-based interactions. As LLMs mimic personality traits such as conscientiousness, agreeableness, and openness, they increasingly resemble human interlocutors. However, when these traits appear inconsistent or lack emotional depth, users may experience discomfort. This aligns with the uncanny valley theory, where entities that seem nearly human but not fully authentic evoke unease and mistrust.

2.2.1 Key Factors Influencing Linguistic Uncanniness

We define linguistic uncanniness as the sense of discomfort or eeriness experienced when interacting with conversational agents, such as chatbots or large language models (LLMs), that exhibit human-like behavior but fail to meet expectations of natural communication. Understanding the factors influencing linguistic uncanniness is crucial for designing human-centered AI systems.

Anthropomorphism and Expectations

Anthropomorphism, or attributing human characteristics to non-human entities, has been a central factor in the perception of linguistic uncanniness. Research by Nass and Moon [55] suggests that the more human-like a conversational agent appears, the higher the user's expectations for natural language interaction. When these expectations are unmet—such as through unnatural syntax, stilted responses, or a lack of contextual understanding—users may experience cognitive dissonance, contributing to uncanniness.

Semantic and Pragmatic Errors

Semantic errors, such as irrelevant or contradictory responses, and pragmatic errors, including an inability to infer conversational intent, amplify the perception of uncanniness [56]. Such errors break the flow of conversation, making the interaction feel less natural. Studies by Gao et al. [57] highlight that even minor linguistic inconsistencies, such as incorrect use of idioms or colloquialisms, can evoke unease.

Emotional Tone and Empathy

The emotional tone of a conversation plays a significant role in perceived naturalness. Conversational agents that fail to exhibit appropriate emotional responses or empathy often seem mechanical and unsettling [58]. Misalignment between the user’s emotional state and the agent’s response, such as offering humor inappropriately, exacerbates linguistic uncanniness.

Personalization and Context Awareness

Lack of personalization and limited context awareness are other critical factors. According to Zhang et al. [59], users expect conversational agents to remember past interactions and tailor their responses accordingly. Failure to do so creates a sense of detachment and undermines the illusion of natural conversation, contributing to discomfort.

2.2.2 Uncanny Valley and Chatbot Interactions

There's been very little research on uncanny valley effects on chatbot or textual conversations. Most studies rely on visual and behavioral cues in avatars and robots to measure the uncanny valley effect. However, there is little focus on purely textual interactions, especially in modern LLM-based chatbots. Some key studies relevant to uncanny valley and chatbot interactions are summarized below:

Study 1: Affective Responses Vary Significantly Based on the Type of Chatbot

A 2019 study by Ciechanowski et al. showed that users' affective responses vary significantly based on the type of chatbot interface humans interact with [4]. The researchers conducted a two-stage experiment involving psychophysiological measurements (such as electrodermal activity and electromyography) and questionnaires. Participants interacted with either a text-only chatbot or an animated avatar chatbot, and their emotional responses were analyzed. The simpler text-based chatbot triggered fewer negative emotions and less intense psychophysiological reactions than the human-like avatar chatbot. This suggests that human-like elements, even in digital interfaces, can induce feelings of eeriness, a phenomenon central to the uncanny valley. Ciechanowski et al.'s study found that the type of task also influenced user perceptions. For example, users interacting with the avatar chatbot for casual conversations experienced more negative responses than those engaging with a text-based chatbot for informational tasks.

Study 2: Human-like Conversational Abilities with Empathetic Responses Could Bridge the Uncanny Valley

Betriana et al. explored the application of Masahiro Mori's Uncanny Valley theory in the context of healthcare robots using Artificial Affective Communication (AAC) and Natural Language Processing (NLP) [61]. The authors argue that robot physiognomy (appearance) alone does not fully explain the uncanny valley effect. Instead, the language capabilities and emotional expressiveness of robots play a crucial role in how humans perceive them. The study emphasizes that mismatched vocal and non-verbal cues can cause feelings of eeriness, especially when robots express inappropriate emotional responses or use unnatural language.

The paper calls for a focus on AAC to improve human–robot interactions, suggesting that human-like conversational abilities with empathetic responses could bridge the uncanny valley. It also discusses the potential of healthcare robots to build trust through authentic language and personalized interactions.

Study 3: Anthropomorphic Language Cues and the Uncanny Valley in Conversational Agents

Feine et al. [60] propose a comprehensive taxonomy of social cues for conversational agents, including linguistic elements such as empathy, self-disclosure, and colloquialisms. Their findings show that while human-like language can enhance user engagement, it can also induce eeriness if it appears inauthentic or robotic. The study underscores that even in purely text-based settings, overusing or misapplying anthropomorphic language may push users closer to an uncanny valley effect.

Study 4: Perceived Mind and the Uncanny Valley Effect

Gray and Wegner [62] focus on how people attribute "mind" to robots and other entities, exploring why inconsistencies between appearance and behavior provoke discomfort—at the heart of the uncanny valley. They argue that when an agent seems highly intelligent or empathetic yet responds incoherently or out of context, the mismatch triggers feelings of unease. Although their research delves into physical robots, it applies to text-based chatbots: conversational inconsistencies or seemingly "human" phrasing followed by nonsensical replies can heighten the uncanny valley effect.

Study 5: Mismatched Cues in Voice and Text Lead to Eeriness

Abubshait and Wiese [63] investigate how inconsistencies between an agent’s appearance and behavior—including verbal cues—can induce eerie or unsettling perceptions. Although their experiments centered on physical robots, the principle of “mismatched cues” easily extends to chatbots. A system that sounds humanlike but uses awkward, contextually inappropriate language, or vice versa, breaks user expectations and fosters an uncanny effect. This emphasizes the importance of maintaining consistency in a chatbot’s communicative style and emotional expressiveness.

2.3 Evolution of Image Generation Models

Having established an understanding of how the uncanny valley effect can manifest in text conversations within the context of large language models (LLMs), we now extend this investigation to image generation models. This exploration aims to develop a comprehensive understanding of the second key motivating factor in the thesis.

Early Computer-Generated Images and the Roots of the Uncanny Valley

Early computer graphics primarily served scientific and industrial applications, producing wireframe models and simple 2D renderings [64]. Ivan Sutherland’s Sketchpad, developed in the 1960s, was one of the earliest interactive computer graphics systems, enabling users to draw and manipulate objects on a screen [64]. Although these systems were revolutionary for their time, their outputs were highly stylized and clearly non-human in appearance, minimizing any uncanny feelings.

Traditional computer graphics relied heavily on manual modeling and rendering techniques until the proliferation of machine learning methods. Early neural networks, such as autoencoders, were deployed for tasks like image denoising and dimensionality reduction [65]. Although these autoencoders demonstrated the ability to learn latent representations of images, their generative capabilities were limited, often producing blurry or artifact-laden outputs [66]. Because these outputs were seldom mistaken for real human images, the uncanny valley was not yet a widespread concern in this era.

Emergence of Machine Learning Approaches

Generative Adversarial Networks (GANs)

A pivotal shift occurred with the introduction of Generative Adversarial Networks (GANs) by Goodfellow et al. [67]. GANs consist of a generator and a discriminator engaged in a minimax game, whereby the generator strives to produce realistic images, and the discriminator attempts to distinguish between real and synthesized images. This adversarial setup yielded dramatic improvements in image fidelity. Subsequent variants like Deep Convolutional GAN (DCGAN) [68] further refined model stability and image quality.With the advent of GANs capable of generating faces and human figures at high resolutions, the uncanny valley entered the conversation.

Style Transfer and Style-Based Generative Models

In parallel to GAN advancements, researchers explored style transfer and style-based generation. Johnson et al. [69] introduced perceptual losses for real-time style transfer, enabling neural networks to synthesize artistic renditions of images in various painterly styles. Later, Prabakaran et al. [70] proposed a style-based architecture for GANs (StyleGAN), which disentangled different aspects of an image's representation (e.g., facial features, hair style, background) and offered fine-grained control over these features. StyleGAN marked a leap in generating human faces that were not only high-resolution but also flexibly controllable.

Diffusion-Based Models and Multimodal Generation

More recently, diffusion-based models have emerged as strong competitors to GANs in achieving state-of-the-art image generation [71]. These models transform random noise into structured images through iterative denoising processes. By capturing and refining features over multiple steps, diffusion models can generate diverse and highly detailed images, rivaling or surpassing GANs in visual quality.

Additionally, multimodal models like DALL·E [15] extends image generation beyond purely visual input, taking text prompts and producing corresponding images. Such models raise new ethical and psychological considerations regarding authenticity, as users may form expectations about the "intent" behind an image that is, in reality, produced by an algorithm. As these images grow closer to photorealism, misalignments between the text prompt's meaning and the output's subtle details can elicit uncanny reactions.

2.3.1 Key Factors Influencing Uncanniness in AI-Generated Images

Hyper-Realism and Subtle Artifacts

With the advancement of GAN architectures (e.g., StyleGAN, StyleGAN2), AI-generated images have become increasingly photorealistic [73]. Hyper-realistic images, however, are often prone to minor inconsistencies such as asymmetrical facial features, mismatched lighting, or unrealistic skin textures, which viewers may detect subconsciously. These subtle artifacts undermine the illusion of authenticity, thereby heightening uncanniness [73].

Stylized Generation and Cartoonish Aesthetics

In contrast, stylized or cartoonish images tend not to trigger the uncanny valley effect to the same degree, because viewers do not expect perfect fidelity to real-world appearances [74]. Studies suggest that when an image clearly indicates a stylized or abstract context, the viewer's cognitive dissonance is reduced [74]. Consequently, designers who intentionally keep AI-generated visuals slightly stylized may avoid uncanny responses [74].

Role of Human Facial Features

Human faces are potent triggers for social and emotional processing [75]. Slight deviations from typical facial proportions or micro-expressions often lead to discomfort [76]. Anthropomorphic features are particularly sensitive to symmetry, skin texture, and emotional expressiveness. When AI-generated faces exhibit inconsistencies in these cues, they become uncanny because they occupy a blurred boundary between the familiar and the unfamiliar [76].

Eye Gaze and Emotional Expressions

Eye gaze, pupil alignment, and emotional expressions have been singled out as critical in engendering or mitigating uncanny feelings [76]. Research shows that incorrect or ambiguous emotional cues—such as a frozen smile or a deadpan stare—may cause unease in otherwise realistic images [77]. This finding points to the importance of training generative models to capture and render subtle micro-expressions accurately.

2.3.2 Uncanny Valley and AI-Generated Images

There has been a growing body of research on the uncanny valley effect in AI-generated images, particularly as hyper-realistic image generation models like GANs (Generative Adversarial Networks) have advanced. Most of the research focuses on identifying subtle imperfections that elicit discomfort in viewers. Below are a few key studies that contribute to our understanding of how these visual artifacts influence the uncanny valley effect in AI-generated images.

Study 1: Signatures of the Uncanny Valley Effect in an Artificial Neural Network

Igaue and Hayashi [80] used the CLIP (Contrastive Language-Image Pre-training) model to investigate how an AI system perceives the uncanny valley effect in manipulated images [26]. By blending human faces with non-human objects (e.g., cars, vegetables) using morphing techniques, the researchers measured the alignment between image features and words associated with the uncanny valley (e.g., eerie, creepy).

Their results showed that CLIP identified the highest eerie responses at the midpoint of blending human faces with other objects, mirroring human responses to the uncanny valley. The study demonstrates that an AI model can exhibit behavior similar to human sentiment when evaluating human-likeness and supports the hypothesis that perceptual mismatches trigger uncanny feelings.

Study 2: Amplifying the Uncanny through Generative Adversarial Networks (GANs)

Broad et al. [78] explored how manipulating GANs to optimize for generating images perceived as "fake" amplifies the uncanny effect in digital art. The study took a pre-trained StyleGAN model and fine-tuned it to produce images the discriminator identified as fake, rather than real. The researchers aimed to explore how this reversal of typical GAN objectives impacts the aesthetic outcome of generated images, resulting in eerie and abstract visuals.

Their experiments showed that pushing the generator toward creating "unlikely" images resulted in visual artifacts, such as asymmetrical facial features, misaligned eyes, and exaggerated textures, which heightened the uncanny valley effect. By optimizing for dissonance rather than realism, the authors created artworks that provoke a sense of unease in viewers.

Study 3: Human Sensitivity to Minor Imperfections in AI-Generated News Anchors

Wu et al. [79] explored the uncanny valley effect in the context of AI news anchors by examining human emotional responses to hyper-realistic digital characters. The study found that AI-generated news anchors struggle to establish emotional bonds with viewers due to minor imperfections, such as mismatched lighting, facial asymmetry, and unnatural hand gestures. These subtle flaws heightened viewer discomfort and decreased trust, placing the AI anchors within the uncanny valley.

The researchers conducted a user study involving AI and human news anchors to measure emotional perceptions such as humanness, eeriness, and attractiveness. The results indicated that human news anchors were consistently rated more favorably. Interestingly, the study also tested variations in AI anchor design, such as gender and hand gestures, to determine how these factors influence user responses. The findings emphasize the importance of minimizing perceptual mismatches to avoid the uncanny valley effect.

2.4 Evaluation

2.4.1 The Uncanny Valley - A Critique

The literature surveyed on the uncanny valley offers valuable insights into how near-human entities provoke discomfort; however, it also reveals significant methodological, conceptual, and ethical limitations that constrain both the reproducibility of findings and the applicability of theoretical models. First, methodological shortcomings emerge from widespread reliance on self-reported measures of discomfort—often varying scales of “eeriness” or “unsettlingness.” The subjective nature of these assessments raises questions about their reliability, and the heterogeneous range of stimuli employed (e.g., static images, robotic prototypes, CGI avatars) makes cross-study comparisons difficult. The lack of standardized metrics, as noted by Bartneck et al. [47], further hampers replicability. While these criticisms do not invalidate the uncanny valley hypothesis, they underscore the urgency for more systematic protocols and validated tools.

Second, the conceptual challenges surrounding “human-likeness” and “eeriness” point to deeper theoretical ambiguities. As multiple scholars suggest, conflating the uncanny valley with other psychological phenomena—such as fear of the unfamiliar or existential anxiety—dilutes its explanatory power. This conceptual overlap risks turning the uncanny valley hypothesis into a “catch-all” explanation for human discomfort with any near-human entity, limiting its precision and coherence. Future research would benefit from clarifying definitional boundaries and distinguishing uncanny valley effects from related constructs.

Finally, the ethical and societal impacts of the uncanny valley show that it has real-world consequences for how technologies are designed and perceived. Trying to avoid the uncanny valley by creating overly polished or standardized humanoid figures can result in unrealistic representations of people. Additionally, people’s discomfort or negative reactions toward humanoid robots—whether based on instinct or familiarity—could slow down the use of such technologies in important areas like healthcare and education, leading to real social setbacks.

These issues highlight the need for research to go beyond just understanding the uncanny valley and to focus on design choices that promote and build public trust in human-like technologies. The critiques about methods, concepts, and ethics reveal that the uncanny valley is still a debated idea. To move forward, researchers need to improve measurement tools. Doing this will help ensure these technologies are more comfortable for users, widely accepted by society, and reflective of the diversity of human experiences.

2.4.2 LLMs and the Uncanny Valley - A Critique

Despite the substantial progress in understanding how LLMs and chatbots evoke (or fail to evoke) human-like conversational experiences, several gaps and limitations emerge from the existing body of work. First, while early research on the uncanny valley has primarily focused on visual and behavioral cues in humanoid robots, comparatively fewer studies address purely text-based interactions. The limited research that does so—such as Ciechanowski et al.'s [4] investigation of text-only versus avatar-based chatbots—indicates that text-based systems may elicit fewer negative responses. However, these findings may not generalize to advanced LLMs capable of more nuanced, human-like conversation.

Second, most empirical work relies on psychophysiological and self-report measures (e.g., questionnaires) to gauge user reactions. These techniques capture immediate affective responses but often do not explore the long-term psychological impact of interacting with near-human textual entities. As LLMs continue to scale and incorporate advanced features like empathy simulation ("synthetic empathy" [28], [29]), there is a growing need for longitudinal studies that assess how sustained or repeated interactions might influence user comfort, trust, and potential uncanniness.

Third, while several studies underscore the importance of emotional tone and empathy (e.g., Betriana et al. [61], Feine et al. [60]), existing research tends to focus on whether chatbots can display empathy rather than whether that empathy rings authentic or consistent across diverse contexts. As highlighted by Gray and Wegner [62] and Abubshait and Wiese [63], users are particularly sensitive to mismatched cues—when an agent appears both highly intelligent and empathetic but then delivers incoherent or insincere responses, the resulting discomfort can be intensified. Future work must refine measures of perceived authenticity in LLM-generated empathy to better understand how such inconsistencies exacerbate the uncanny valley effect in purely textual interfaces.

Fourth, personalization and context awareness remain critical but under-explored. Zhang et al. [59] emphasize the importance of maintaining context across multiple turns in a conversation, yet many LLM-based studies typically evaluate performance on single-turn or short multi-turn dialogues. The lack of robust memory or user-modeling strategies can limit a chatbot's ability to adapt its responses, potentially undermining user trust and heightening uncanniness. More studies are needed to assess the interplay between personalization strategies (e.g., adaptive tone, style, and empathy) and user comfort over longer conversations.

Finally, while large-scale models like GPT-4 [26] have demonstrated remarkable capabilities in generating near-human text, a key question remains: at what threshold of semantic and pragmatic accuracy do users begin to perceive these systems as "too human," and thus unsettling? Existing findings suggest that the uncanny valley effect emerges from inconsistencies between a user's high expectations and the bot's occasional lapses—particularly in emotional or contextual accuracy. However, quantifying these thresholds and systematically determining at what point human-like traits provoke uncanniness has not been rigorously explored in textual domains.

2.4.3 The Uncanny Valley in AI Generated Images - A Critique

Despite the wealth of research on AI-driven image generation and its relationship to the uncanny valley, several gaps and limitations remain. First, the methodological diversity across studies complicates direct comparisons of findings. While certain investigations employ human subject testing (e.g., Wu et al. evaluating emotional responses to AI news anchors [79]), others rely on computational models (e.g., Igaue and Hayashi using CLIP to detect "eerie" responses [80]). These varying approaches offer complementary insights but also result in inconsistent measures of "uncanniness" and disparate criteria for evaluating image realism and emotional impact.

Second, although many studies identify specific visual artifacts (e.g., asymmetrical facial features, unnatural hand gestures) as triggers of uncanny responses, there is a lack of consensus on how to quantitatively benchmark these artifacts. Subtle differences in lighting or skin texture may be more salient in one context than another, and current metrics—such as Fréchet Inception Distance (FID) or Structural Similarity Index (SSIM)—do not fully capture the perceptual and emotional nuances that define the uncanny valley. As a result, researchers must rely on subjective or narrowly defined evaluations, limiting the generalizability of their findings.

A related challenge is the subjectivity of human perception. The uncanny valley phenomenon is influenced by cultural, individual, and situational factors, yet most studies do not incorporate demographic diversity into their analyses. Viewers' familiarity with digital media, personal tolerance for imperfect visuals, and cultural norms regarding human appearance can all alter their responses to AI-generated images. Hence, findings based on a single cultural or demographic group may not accurately represent broader population responses, leading to potential biases in how "uncanniness" is defined or perceived.

In addition, there is a trade-off between hyper-realism and creative stylization in current models. Studies such as Broad et al. highlight how deliberately generating “unlikely” or “fake”-looking images can intentionally amplify eerie qualities [78]. Conversely, maintaining a more stylized or cartoonish aesthetic can circumvent the uncanny valley by setting lower viewer expectations for fidelity to real-world appearances. While this divergence in artistic goals broadens the creative applications of AI, it also complicates attempts to establish universal guidelines for mitigating the uncanny valley effect across diverse use cases (e.g., digital art, film, virtual assistants, and other interactive media).

Finally, the emergence of multimodal models (e.g., DALL·E) and diffusion-based techniques introduce new ethical and psychological considerations regarding authenticity and intent. As AI systems become adept at translating text prompts into photorealistic images, mismatches between the intended semantic meaning and the resulting visual cues can contribute to uncanniness. However, there is insufficient research on the long-term psychological impacts of widespread exposure to near-perfect synthesized images, including potential desensitization to manipulations or heightened mistrust of digital media.

2.5 The Research Gap

Collectively, the critiques outlined above underscore several fundamental gaps that future research on the uncanny valley—across humanoid robotics, large language models, and AI-generated images—must address. First, there is an overarching lack of standardized metrics for capturing uncanniness, with most studies still relying on subjective self-reports or ad hoc computational approaches. While scales of “eeriness” and “unsettledness” provide valuable initial data, their variations in wording, scoring systems, and administration procedures compromise the comparability and replicability of results. Further, the diversity of stimuli employed (e.g., animated versus text-based agents) compounds the challenge of establishing universally applicable assessment tools.

Second, there is a lack of literature examining how the uncanny valley might arise in interactions with recent large language models (LLMs) and cutting-edge AI-generated images. While emergent technologies like GPT-4, DALL·E, and other diffusion-based or transformer-based systems have made rapid strides in producing text and visuals that rival human outputs, the literature has not yet caught up to investigate the specific ways these platforms may elicit uncanny responses. This gap is particularly critical given the speed at which such models are being integrated into everyday applications—from customer service chatbots to image-based social media filters. Without robust empirical data on whether and how users experience these highly sophisticated LLMs and near-photorealistic AI images as unsettling or "too human," it becomes difficult to anticipate design pitfalls, user reactions, or ethical implications.

Third, although numerous studies point to inconsistencies in perceived human-likeness (e.g., mismatch between sophisticated dialogue and non-human emotional responses, or hyper-real facial rendering paired with uncanny digital artifacts), there is no consensus on how to quantify these mismatches systematically. The conceptual ambiguity around "near-human" traits—whether in text-based empathy simulation, emotional tone, or photorealistic facial features—makes it difficult to discern precisely where and why the uncanny valley effect manifests. As a result, existing findings risk conflating other psychological phenomena, such as general fear of the unfamiliar, with the uncanny valley.

Fourth, both short-term and long-term user responses to near-human technologies remain under-explored. Most empirical work employs immediate psychophysiological or self-report measures, neglecting longitudinal dimensions such as whether repeated interactions (e.g., with chatbots that simulate empathy) become more unsettling over time. This gap is especially pronounced in text-only settings, where the boundaries of "too human" conversation remain unclear and may evolve as large language models grow more context-aware and personalized.

Additionally, cultural and demographic factors are underrepresented. People's familiarity with digital media, personal thresholds for visual or textual imperfection, and sociocultural norms about human appearance and interaction can all influence the experience of uncanniness. Yet the majority of existing research bases its conclusions on relatively homogenous participant pools. This narrow sampling risks overlooking critical variations in how different populations perceive, interpret, and react to near-human agents—whether robotic, textual, or visually generated.

Finally, ethical considerations have yet to be fully integrated into empirical frameworks. Efforts to “design around” the uncanny valley can inadvertently foster unrealistic or exclusionary representations, while increased realism in text and images raises new questions about authenticity, trust, and manipulation. The potential for heightened mistrust of digital media, especially as AI techniques advance, also calls for more robust investigations into the broader psychological and societal impacts of constantly encountering near-perfectly synthesized human likenesses.

Addressing these gaps requires multidisciplinary and integrative approaches. By developing standardized, validated measurement tools; clarifying conceptual boundaries around human-likeness; conducting longitudinal and cross-cultural studies; and embedding robust ethical safeguards into design processes, scholars and practitioners can move beyond the current limitations.

This thesis addresses two critical gaps amongst the ones discussed above: first, examining whether users experience negative reactions to near-human, text-based interactions in the context of large language models (LLMs), and second, refining our understanding of the metrics, techniques, and user perceptions shaping image generation. The next two chapters present experiments designed to address these gaps.

Chapter 3

Experiment 1: Testing the Uncanny Valley Effect in LLM Generated Texts

3.1 Context & Background

As discussed in the literature review, while the uncanny valley effect has been extensively studied in humanoid robots and animated characters, its implications in the context of conversational agents, such as large language models (LLMs), remain underexplored. Given the growing integration of LLMs in customer service, healthcare, and education, understanding how these systems are perceived is crucial for designing AI-driven interactions that foster user trust and engagement. As part of this experiment, we aim to explore the uncanny valley effect of conversational interactions with large language models.

In the experiment, we use the Godspeed Questionnaire—a widely used tool in human-computer interaction (HCI) research [84]. The original questionnaire is provided in Appendix B. Originally developed to measure user perceptions of robots, this instrument evaluates five dimensions: Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Safety. We used this questionnaire because it provides a standardized framework for evaluating the human-likeness and emotional responses evoked by interactions with AI, making it especially useful for studying phenomena like the uncanny valley effect in conversational agents.

The first dimension, Anthropomorphism, measures the extent to which users perceive an agent as having human-like characteristics. High anthropomorphism scores suggest that users see the agent as resembling a human in appearance or behavior. The second dimension, Animacy, assesses whether the agent is perceived as alive and capable of independent action. This dimension captures whether users feel that the agent exhibits lifelike or dynamic behaviors, indicating the presence of vitality in the interaction.

The third dimension, Likeability, evaluates users' emotional responses to the agent, specifically focusing on how pleasant and friendly they find the interaction. Likeability reflects the agent's ability to establish rapport with users and contribute to a positive user experience. The fourth dimension, Perceived Intelligence, measures how competent, knowledgeable, and responsive the agent appears to be. High scores in this dimension indicate that users view the agent as capable of understanding and responding appropriately to their inputs.

Finally, Perceived Safety assesses whether users feel secure and comfortable during the interaction. This dimension is particularly relevant in sensitive or high-stakes contexts where user trust is critical. In this context, since our experiments are designed to arguably not be ultra-high-stakes, compared with, for example, interactions with industry robots, safety goes beyond physical safety to capture the human respondents' broad psychological safety.

These dimensions form the foundation of our experimental design, helping us evaluate user perceptions in conversations with both human participants and AI agents. By applying this framework, we aim to uncover how the uncanny valley effect manifests in textual interactions and whether users perceive a clear difference between human and AI responses.

3.2 Methodology

3.2.1 Experimental Design

Our primary hypothesis was that the uncanny valley effect extends to text-based interactions with conversational agents. We also aimed to test whether LLMs comprehend the concept of the uncanny valley and whether this understanding aligned with human responses.

Specifically, the study sought to quantitatively assess the manifestation of the uncanny valley effect in textual conversations. To achieve this, three chatbots were developed: the first, powered by GPT-4o (GPT-4 Turbo), was intentionally designed to elicit responses likely to fall into the uncanny valley; the second, also powered by GPT-4o, was crafted to closely replicate human-like conversational patterns and avoid the uncanny valley; and the third was a human participant engaging in the conversation. In this case, the human participant was the primary researcher conducting the experiment. As defined below, each type served a distinct purpose in testing the uncanny valley effect:

Type	**Purpose**	**Hypothesis Tested**
Chatbot 1 (GPT4o-Uncanny-Valley-Bot)	Designed to fall into the uncanny valley	Tests if humans experience discomfort or unease when interacting with a less natural AI
Chatbot 2 (GPT4o-Human-Like-Bot)	Designed to avoid the uncanny valley	Tests if making AI more human-like reduces discomfort
Human Participant (Human-to-Human Bot)	Acts as the "control"	If participants still experience the uncanny valley with the human participant, it suggests that the phenomenon may not be specific to AI.

Table 1: Types of chatbots and their purpose for Experiment 1

3.2.2 Chatbot Setup

To simulate different levels of human-likeness, each chatbot's responses were tailored through prompt engineering. The prompts were iteratively refined to evoke distinct user perceptions. Below is the final prompt used to create the "Uncanny Valley" chatbot:

Converse like you are human but not quite. Be in the uncanny valley. Do not tell that you are an AI agent or chatgpt. Do not use words such as "assist", "help" etc. Keep each answer in the conversation within 50 words.

The human-like AI bot was prompt engineered as follows:

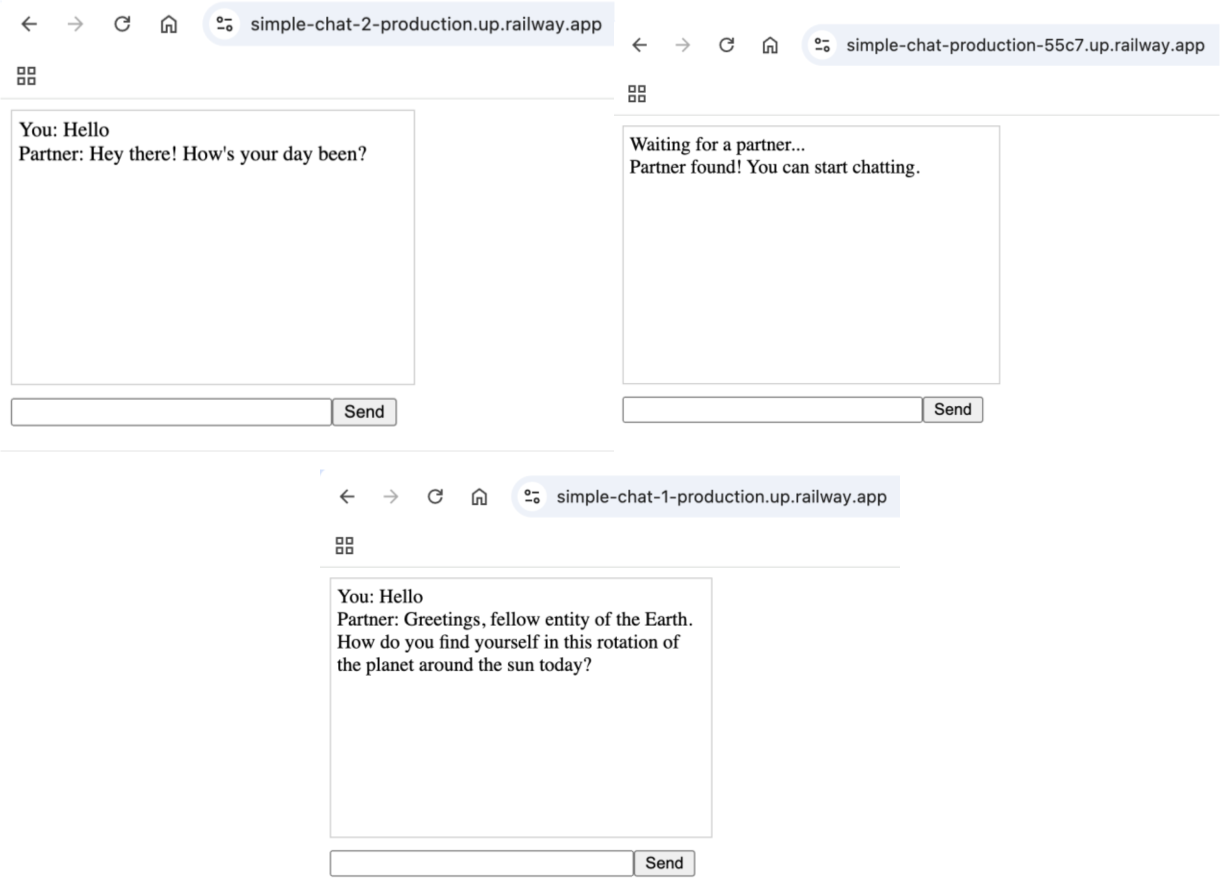
Converse like you are human. You can give multiple sequential responses to a single question. Do not tell that you are an AI agent or chatgpt. Act like humans - you can eat, see, smell, breathe, touch, and hear. Keep your responses realistic and human-like. Do not use words such as "assist", "help" in greetings. Keep each answer in the conversation within 50 words.

The removal of words “assist”, “help” etc. and 50 word limit were used to make the bots act as human-like as possible.

Additionally, to keep consistency between the AI driven and human conversations, human conversationalist’s typing speed was matched with GPT’s and an additional delay was added in the response from the API. The human conversationalist's typing speed was determined from an online typing test tool to be 43 words per minute. To calculate the delay required to simulate this typing speed, the following formula was used:

Delay (seconds) = (Number of words (n) X 60) / Typing Speed (words/min)

All chatbots, including the human-to-human chatbot were coded in NodeJS and hosted in the Railway App to expose the apps to a publicly accessible URL. The UI was the same and consistent across the apps. The images of the UI are shown in figure 3.


simple-chat-2-production.up.railway.app
Send
simple-chat-production-55c7.up.railway.app
Send
simple-chat-1-production.up.railway.app
Send

You: Hello
Partner: Hey there! How's your day been?

Waiting for a partner...
Partner found! You can start chatting.

You: Hello
Partner: Greetings, fellow entity of the Earth. How do you find yourself in this rotation of the planet around the sun today?

Figure 3: UI of different chatbots used in the experiment: Top Left: GPT4o-Human-Like-Bot, Top Right: Human-to-Human Bot, Bottom Center: GPT4o-Uncanny-Valley-Bot

3.2.3 Data Collection

For sampling, we referred to research design literature, particularly in psychology, HCI, and social sciences, as outlined in *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* [89]. Exploratory studies typically require smaller sample sizes, often around 30 participants, to identify initial trends and patterns. Given the novelty of our research and the lack of prior knowledge regarding the effect size, power, or alpha for a formal power analysis, we chose a sample size of 20 participants per group. Specifically, our study included three groups: participants interacting with a bot situated in the uncanny valley, a human-like bot, and a human-to-human interaction bot. This approach allowed us to balance practical constraints with the need for meaningful data to observe potential trends in the perception of AI and the uncanny valley effect.

For participant recruitment, invitations were distributed primarily through WhatsApp and email, targeting a sample of MIT graduate students. Potential participants were asked to schedule a 15-minute Zoom session via Calendly, allowing them to choose a time slot at their convenience. No prior information about the experiment was provided to participants to ensure unbiased responses. The only details shared were that the study would take approximately 15 minutes and would involve text-based interactions followed by a short survey.

3.2.4 Analysis Techniques

We used the following techniques to analyse the conversational data collected from the experiment:

1. **Score Averages:** All the ratings from the Godspeed Questionnaire were totaled for each type of chatbot and the averages were calculated for all dimensions: Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Safety.

2. **Sentiment Analysis:** Sentiment analysis was conducted for all conversations by participant ID. For sentiment analysis, we used TextBlob, a lexicon-based tool that calculates polarity (positive vs. negative sentiment) and subjectivity (objective vs. subjective language) for each response [85]. TextBlob's PatternAnalyzer assigns sentiment scores based on a predefined dictionary of words, allowing for efficient and interpretable analysis across conversations.

3. **Topic Modeling:** We used Topic modeling to analyse themes in conversations. Topic modeling is an unsupervised machine learning technique used to identify latent themes

within a collection of documents. By examining word co-occurrence patterns, it helps uncover underlying topics across large datasets. We used Non-negative Matrix Factorization (NMF) to perform topic modeling on the textual data [86]. First, text preprocessing was conducted to clean and tokenize the data. Term Frequency-Inverse Document Frequency (TF-IDF) was applied to extract the most significant words in each conversation [87]. NMF was then employed to identify five key topics based on word importance and co-occurrence patterns.

4. **Conversational Dynamics:** The study of conversation dynamics in human-agent interactions focuses on understanding how dialogue flows between users and conversational agents. It involves quantifying dialogue structure, measuring emotional tone, and identifying interaction patterns. Conversational dynamics were analyzed by measuring key interaction metrics, including the total number of exchanges, participant and agent contributions (line counts), and turn-taking patterns. Sentiment polarity was calculated for both participant and agent responses to assess emotional tone.

5. **One-Way ANOVA:** To determine whether there were statistically significant differences in participant perceptions across the three bot types (Human Chat, GPT4o-Human-Like Bot, and GPT4o-Uncanny-Valley-Like Bot), a one-way ANOVA (Analysis of Variance) test was conducted on the average scores from the five Godspeed dimensions: Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Safety. A one-way ANOVA is a statistical test used to determine whether there is a significant difference in the mean scores across three or more independent groups. It assumes that the samples are normally distributed and have similar variances, and provides an overall test of group differences before any post-hoc comparisons are performed. The ANOVA test compares the variance between groups (i.e., differences in scores across bot types) to the variance within groups (i.e., variability in participant responses for each bot type).

6. **Tukey's HSD:** We used Tukey's HSD when the overall test (ANOVA) indicated a significant difference to determine *which* groups differ. Tukey's Honestly Significant Difference (HSD) test is a post-hoc procedure used after a significant one-way ANOVA to determine which specific group means differ. It controls the overall Type I error rate by adjusting for multiple comparisons, ensuring that statistically significant differences between group pairs are identified reliably.

7. **Chi-Square Test:** For topic modeling (categorical data), we used the chi-Square test. The Chi-Square Test is a statistical method used to determine if there is a significant relationship between categorical variables (Test of Independence) or if observed data matches expected distributions (Goodness-of-Fit). It compares observed and expected frequencies, with results indicating whether differences are statistically significant.

3.2.5 Experimental Procedure

The step-by-step procedure for the experiment was designed to ensure consistency and minimize bias in participant interactions. The detailed protocol was as follows:

1. The participant and the researcher logged into a Zoom meeting at the scheduled time.
2. The researcher provided a brief overview of the process without revealing the type of bot the participant would chat with. The participant was instructed to open a shared URL and engage in a text-based chat for five minutes. They were informed that the conversation could be on any topic of their choice and that both audio and video could remain off during the session. The participant was explicitly told that only the text-based conversation would be recorded for analysis.
3. Once the participant opened the shared URL, the researcher started a stopwatch to track the five-minute chat duration. The researcher then turned off their audio and video to avoid influencing the participant's behavior.

   - If the participant was assigned to the human chatbot condition, the researcher themselves engaged in the conversation from the other side of the chat interface. In this condition, the researcher was instructed to avoid disclosing any identity-related information, ensuring that the participant was unaware of whether they were chatting with a human or an AI agent.
4. After five minutes, the researcher ended the chat session and provided the participant with a unique participant ID. The participant was then asked to complete a survey, which included the consent form and the Godspeed Questionnaire to assess perceptions of the interaction. The survey provided to the participants is provided in Appendix C. On an average, participants took 5 mins to complete the survey.
5. Upon completing the survey, participants were asked to optionally provide general feedback on any aspect of their experience, including the chat session or the survey process. The researcher recorded the feedback for qualitative analysis.
6. Once the feedback was collected, the researcher thanked the participant for their time, and the Zoom call was disconnected. Participants were reminded that their responses would remain anonymous and that the data would be used solely for research purposes.

This standardized process ensured consistency across all experimental conditions while allowing participants to express themselves freely during the chat sessions.

3.3 Results

The results are organized to provide a comparative analysis of user responses across the three conversational experiences, focusing on key dimensions assessed by the Godspeed Questionnaire. The findings highlight trends and differences in user perceptions, laying the foundation for discussing their implications within the context of the uncanny valley hypothesis.

Survey Demographics

We surveyed 60 individuals, 20 with each type of chatbots described above. The recorded demographics are as follows:

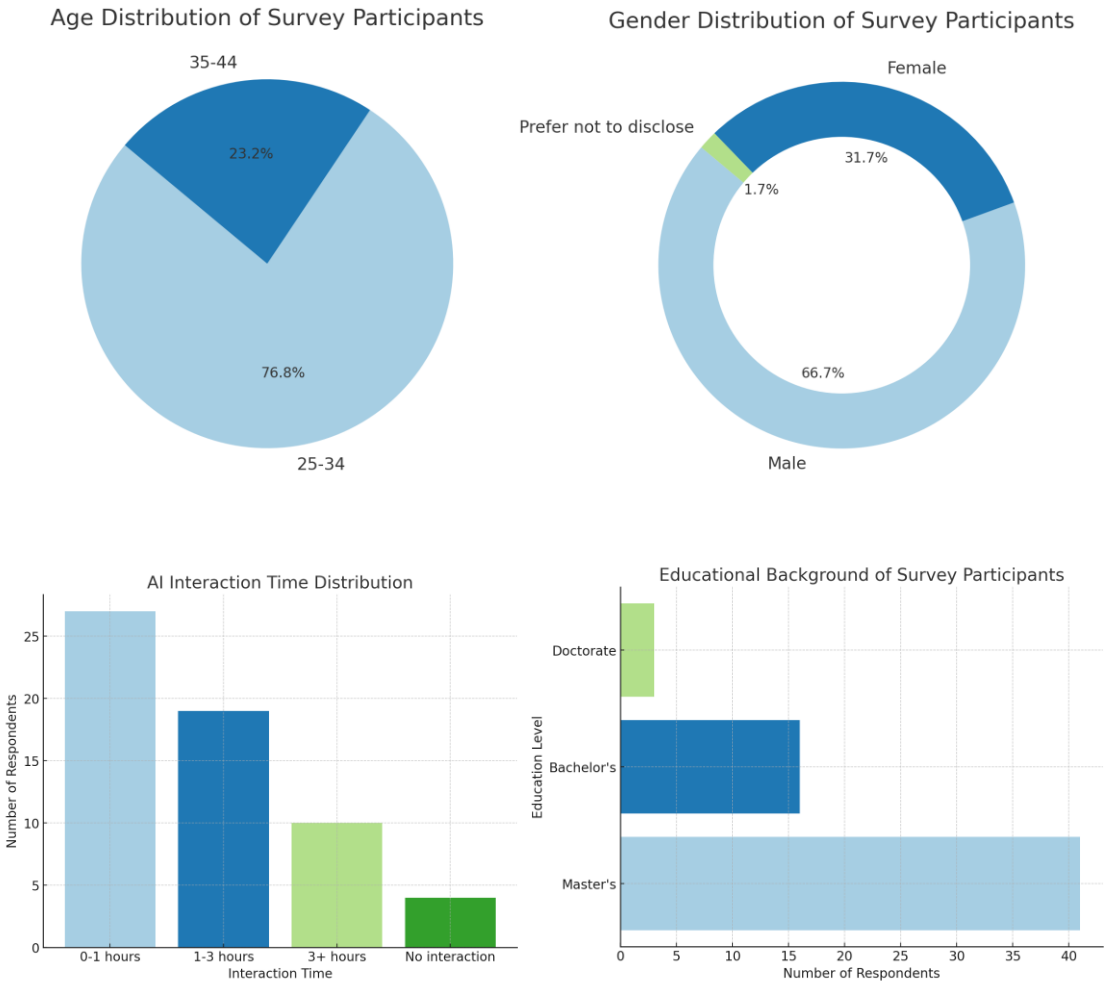


Figure 4: Demographics of the survey (Experiment 1)

Age: The survey participants primarily belonged to the 25-34 age group (72%, 43 respondents), followed by 35-44 (22%, 13 respondents).

Gender: The survey had a majority of male respondents (67%, 40), followed by female respondents (32%, 19), with 2% (1 respondent) preferring not to disclose their gender.

Educational Background: The majority of survey respondents held a Master's degree (68%, 41), followed by Bachelor's degree holders (27%, 16) and a small percentage with a Doctorate (5%, 3).

Languages Spoken: The majority of respondents reported English as their native language, with notable representation from Hindi, Spanish, Chinese, Filipino, and other languages such as Amharic, Kannada, Arabic, and Tamil, reflecting a diverse linguistic background.

AI-Interaction: The majority of respondents interacted with AI agents for 0-1 hour daily (45%, 27), followed by 1-3 hours (32%, 19) and 3+ hours (17%, 10), while a small portion (7%, 4) reported no interaction at all.

3.3.1 Results - Human-to-Human Chatbot Interaction

Perceived Identity of Conversational Partners

Participants were asked to report after the conversation whom they thought they were talking to. Out of the 20 participants who engaged in human-to-human chat interactions, **12 participants (60%) reported that they believed they were interacting with an AI agent**, despite the conversations being conducted with a human interlocutor. In contrast, **8 participants (40%) correctly identified the conversation as human-to-human**.

Godspeed Questionnaire Results

Dimension	Average Score
Anthropomorphism	3.38
Animacy	3.58
Likeability	4.13
Perceived Intelligence	3.94
Safety	2.82

Table 2: Results of Godspeed Questionnaire for Human-to-Human Bot

The analysis of the Godspeed Questionnaire responses revealed varying perceptions across the five key dimensions of anthropomorphism, animacy, likeability, perceived intelligence, and safety. Likeability received the highest average score of 4.13, indicating that participants found the interaction generally pleasant and engaging. Perceived Intelligence followed closely with an average score of 3.94, suggesting that participants perceived the chat partner as knowledgeable and competent. Animacy scored 3.58, reflecting a moderate perception of liveliness and responsiveness in the interaction.

In terms of anthropomorphism, the average score was 3.38, indicating that the interaction was perceived as somewhat human-like but not overwhelmingly so. Finally, Safety received the lowest score of 2.82, suggesting that participants felt less comfortable or relaxed during the interaction.

Sentiment Analysis

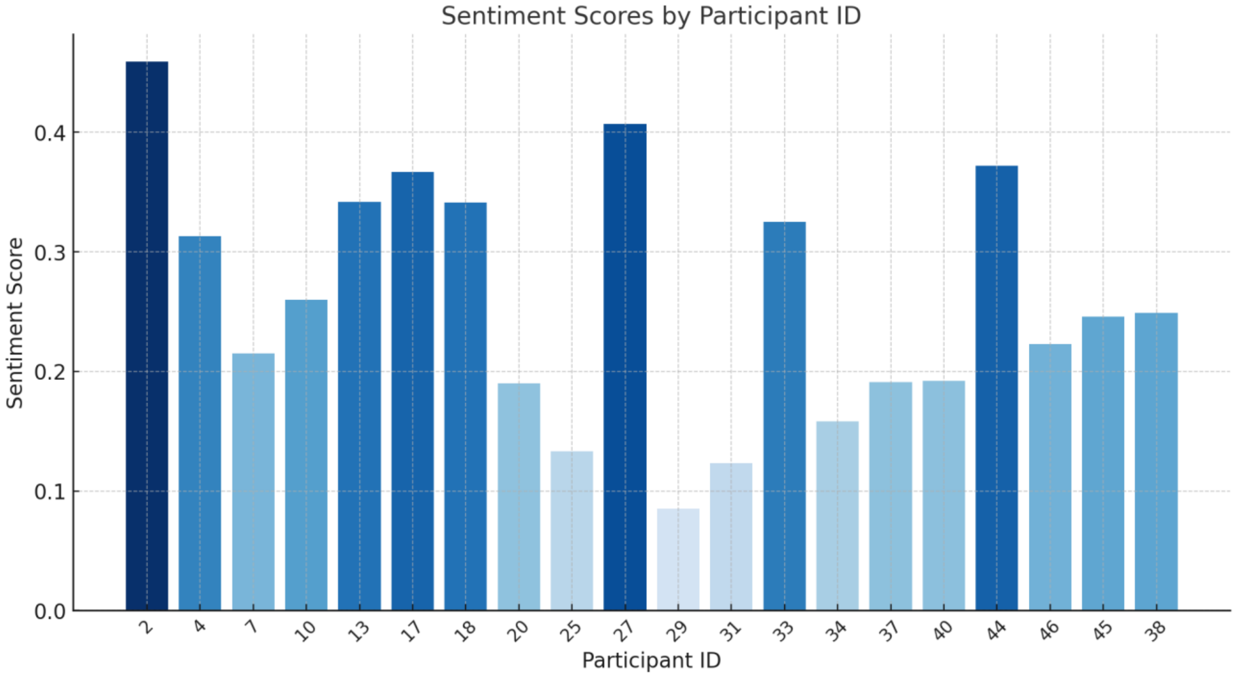


Figure 5: Sentiment Analysis of Human-to-Human Bot

In this analysis, each conversation between two participants was evaluated to determine the overall sentiment score, ranging from -1 to +1, with Positive sentiment (score > 0) indicating friendly, supportive, or enthusiastic exchanges. Negative sentiment (score < 0) reflecting frustration, disagreement, or discomfort. Neutral sentiment (score ≈ 0) suggesting balanced and objective discussions without strong emotional undertones.

The average sentiment score of **0.260** across the analyzed conversations indicates that the majority of the dialogues were generally positive. This suggests that participants engaged in cooperative and emotionally positive interactions. Individual participant sentiment scores are included in Appendix E.

Topic Modeling

The topic modeling revealed the following primary themes across conversations:

1. Holiday Planning and Social Gatherings: Words like "planning," "holiday," "friends," and "family" were dominant, highlighting discussions about holiday activities and social connections.
2. Work and Stress Management: This topic included terms like "work," "destress," "assignments," and "working," indicating conversations centered around professional life and ways to manage stress.
3. Weather and Personal Preferences: Key terms such as "snow," "sunny," "food," and "love" suggested exchanges about weather preferences, personal hobbies, and food.
4. Philosophical and Daily Reflections: Words like "think," "talking," "day," and "going" reflected discussions about daily routines, introspection, and conversational flow.
5. Educational and Professional Interests: Terms like "school," "projects," "help," and "interesting" indicated participants' interest in academic or professional topics, including problem-solving and collaborative learning.

Conversation Dynamics

Participant ID	Total Exchanges	Researcher Lines	Partner Lines	Researcher Sentiment	Partner Sentiment
2	5	5	6	0.358	0.358
4	7	7	8	0.246	0.159
7	9	9	10	-0.043	0.2
10	12	11	13	0.330	0.090
13	6	5	7	0.248	0.157
17	10	9	11	0.355	0.194
20	9	8	11	0.020	0.221
25	9	8	10	0.063	0.018
27	8	8	9	0.290	0.305
29	4	4	4	0.335	-0.198
31	9	8	9	0.025	0.103
33	9	9	9	0.327	0.122
34	7	5	8	0.19	0.101
37	7	7	8	0.092	0.0687
40	5	5	5	0.012	0.265
44	4	4	5	0.375	0.334
46	8	7	9	0.202	0.163
45	6	4	8	0.335	0.148
38	6	6	6	0.067	0.389

Table 3: Results of Conversational Dynamics for Human-to-Human Bot

The results highlight several key observations:

1. **Engagement Patterns:** The average number of exchanges per conversation was approximately 7-12, indicating relatively balanced interactions. In most conversations, the researcher initiated follow-up questions, contributing to sustained engagement.
2. **Turn-Taking Balance:** The researcher and participants contributed almost equally to the conversations, with slight variations in line counts.
3. **Thematic Insights:** Conversations revolved around casual topics like holiday plans, hobbies, and cultural references. Deeper philosophical questions (e.g., "What is the purpose of life?") occasionally emerged.

Emotional Analysis

The emotional analysis of the conversations revealed a generally positive sentiment across interactions, with a mean polarity score of 0.260. The chatbot's responses demonstrated empathy and engagement, particularly in conversations discussing personal stories or holiday plans, contributing to higher polarity scores in these interactions. Subjectivity scores averaged 0.52, suggesting that approximately half of the responses were opinion-based, reflecting the chatbot's inclination toward personalized interactions.

Approximately 70% of the conversations featured positive sentiments, with the remaining 30% being neutral or mixed. Notably, conversations involving emotionally charged topics like stress or nostalgia showed higher subjectivity scores, averaging 0.65, indicating a more personalized and empathetic interaction. In contrast, discussions centered around factual queries, such as financial advice or reinforcement learning problems, exhibited lower subjectivity scores (mean: 0.42).

The analysis also highlighted conversational balance, with the researcher frequently steering conversations back to personal topics after detecting partner fatigue or disinterest. Conversations with partners expressing stress or exhaustion yielded lower polarity scores, with an average of 0.12, compared to 0.45 for more casual and light-hearted exchanges. These findings suggest that the researcher is effective in maintaining a positive tone and fostering rapport, but future improvements could focus on enhancing emotional depth during complex or emotionally charged interactions.

Participant Feedback on Human-to-Human Interaction

Participants provided detailed feedback on their experiences during the human-to-human chat interactions, revealing several key factors that influenced their perceptions of whether their chat partner was human or AI. The responses highlighted timing, language style, conversation flow, and evolving perceptions as significant contributors to their judgments.

Several participants identified response timing as a key indicator of human-likeness. Delays between responses, described as "off" or inconsistent, made the interaction feel less natural to some. One participant noted that the "timing between responses was off", suggesting that smoother timing would have improved their perception of the interaction's naturalness.

The language style used by the chat partner was another frequently mentioned factor. Participants remarked that the formality of language and the use of certain structured, "prophecy-like" phrases were more commonly associated with AI responses. Words like *albeit* were noted as being uncommon in casual human interactions but more typical in AI-generated text or in non-American English language patterns. This formal tone made some participants feel that the interaction was less human-like, despite the content being coherent and relevant.

Participants also commented on the flow and engagement of the conversation. Several observed that the chat partner focused on asking questions without disclosing much personal information, a behavior they associated with AI. One participant mentioned, "It was interesting how the chat partner shared minimal information about themselves and was rather focused on creating conversation by asking more and more questions". This question-driven style led participants to perceive the interaction as less human-like, as human conversations tend to involve more personal storytelling and sharing of experiences.

A few participants reflected on the evolution of their perceptions due to increased exposure to AI tools. One participant remarked that "what I may have previously considered human, I now know could be AI", indicating that the bar for what feels human-like has been raised. This suggests that prior familiarity with large language models (LLMs) might have influenced their judgments.

Interestingly, some participants did identify behaviors that they perceived as human-like. Brief pauses in responses were interpreted as signs of thoughtful engagement, making the conversation feel more natural. One participant noted, "It felt like a human 'cause of the brief pauses in the chat", suggesting that perceived cognitive delays contribute to a more human-like impression.

3.3.2 Results - GPT4o-Human-Like-Bot Interaction

Perceived Identity of Conversational Partners

Participants were asked to report after the conversation whom they thought they were talking to. Out of the 20 participants who engaged in human-AI chat interactions, **6 participants (30%) believed they were interacting with a human, despite the conversation being conducted with an AI agent**. In contrast, 14 participants (70%) correctly identified the conversation as being with an AI.

Godspeed Questionnaire Results

Dimension	Average Score
Anthropomorphism	3.26
Animacy	3.45
Likeability	4.07
Perceived Intelligence	3.62
Safety	2.88

Table 4: Results of Godspeed Questionnaire for GPT4o-Human-Like-Bot

The analysis of participant responses in the GPT4o human-like chatbot dataset reveals varying perceptions across the five key dimensions of anthropomorphism, animacy, likeability, perceived intelligence, and safety. Likeability received the highest average score of 4.07, indicating that participants generally found the interaction pleasant and engaging. Perceived Intelligence scored 3.62, suggesting that participants perceived the chat partner as fairly knowledgeable and competent.

The dimension of Animacy scored 3.45, reflecting a moderate perception of liveliness and responsiveness in the interaction. Anthropomorphism received an average score of 3.26, indicating that the interaction was perceived as somewhat human-like but not convincingly so. The lowest score was observed in the Safety dimension, with an average of 2.88, indicating that participants felt less comfortable or relaxed during the interaction compared to other dimensions.

Analysing GPT4o-Human-Like-Bot Interaction

Sentiment Analysis

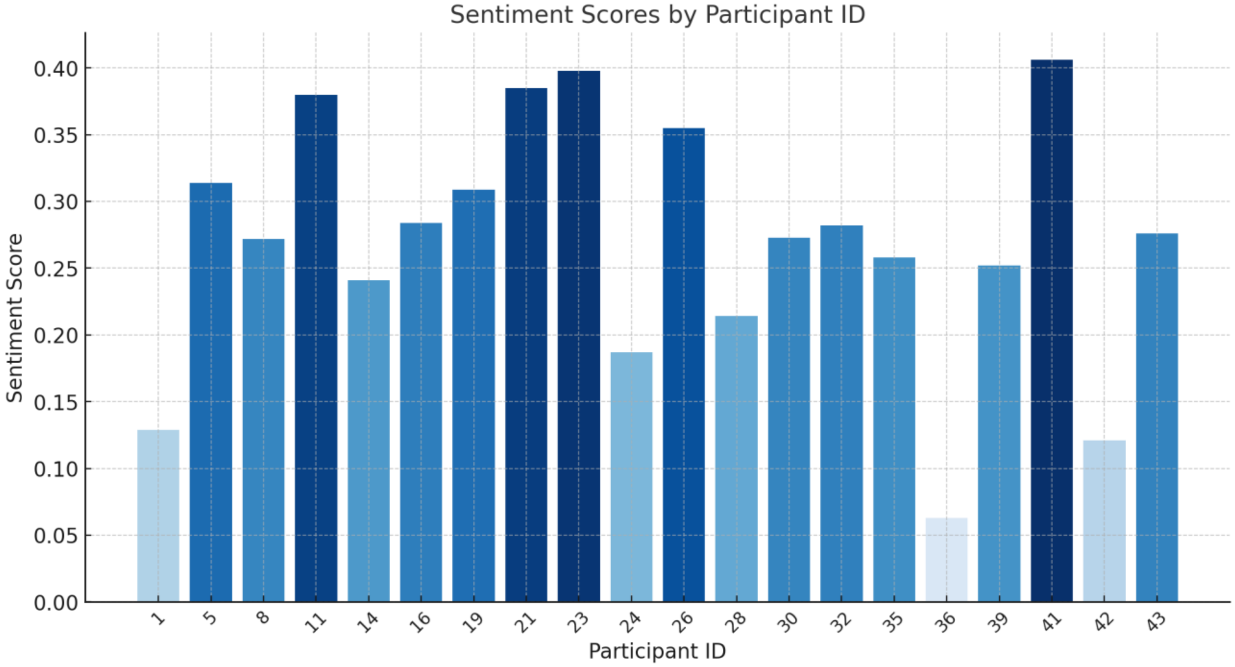


Figure 6: Sentiment Analysis of GPT4o-Human-Like-Bot

The average sentiment score across conversations was 0.268, reinforcing the finding that most interactions were perceived as Positive. Individual participant sentiment scores are included in Appendix E.

Participants often discussed personal experiences, travel plans, or mutual interests, contributing to the positive tone. In contrast, no significant negative sentiment was detected across conversations, suggesting a lack of frustration, disagreement, or discomfort in the dialogues.

Topic Modeling

In this analysis, topic modeling was applied to the given dataset of 20 conversations, each categorized under a unique "Participant ID." Using Non-negative Matrix Factorization (NMF), we identified five key themes across the dialogues. The text data was preprocessed using a custom stopwords list to remove common words and focus on meaningful terms. Term frequency patterns were used to extract the most relevant words for each topic.

The analysis revealed the following primary topics:

1. Time Management and Self-Care: Conversations focused on setting boundaries, prioritizing tasks, and making time for personal activities. Words like "time," "make," "say," and "no" were dominant, reflecting discussions on managing time effectively.
2. Daily Routines and Personal Preferences: Participants shared details about their daily activities, hobbies, and preferences. Words such as "been," "just," "about," and "doing" highlighted casual, personal conversations about routines and preferences.
3. Festive Celebrations and Food: Discussions centered around holiday plans, festive meals, and recipes. Common terms included "year," "trying," "some," and "well," suggesting exchanges about Thanksgiving, dinner plans, and holiday traditions.
4. Shopping and Recommendations: This topic captured conversations related to shopping, product recommendations, and exploring local stores.
5. Travel Planning and Weather: Participants discussed travel plans, destinations, and weather forecasts. Terms such as "trip," "cape," "cod," and "weather" were prominent.

Conversation Dynamics

Participant ID	Total Exchanges	Participant Lines	GPT4o (Human-Like) Lines	Participant Sentiment	GPT4o (Human-Like) Sentiment
1	12	6	6	0.050	0.092
5	24	12	12	0.101	0.357
8	16	8	8	0.303	0.179
11	22	11	11	0.240	0.315
14	14	7	7	-0.005	0.277
16	8	4	4	0.270	0.188
19	10	5	5	0.164	0.415
21	14	7	7	0.135	0.363
23	8	4	4	0.164	0.441
24	8	4	4	0.146	0.156
26	12	6	6	0.366	0.297
28	12	6	6	0.089	0.179
30	8	4	4	0.228	0.333
32	10	5	5	0.276	0.240
35	6	3	3	0.25	0.159
39	8	4	4	0.387	0.135
36	6	3	3	0.0	-0.132
41	24	12	12	0.111	0.440
42	10	5	5	0.083	0.082
43	14	7	7	0.271	0.284

Table 5: Results of Conversational Dynamics for GPT4o-Human-Like-Bot

The conversation dynamics analysis of the 20 conversations reveals several key observations regarding engagement patterns, sentiment analysis, and turn-taking balance, providing insights into the interaction between the GPT4o-bot and participants.

Engagement Patterns: The average number of exchanges per conversation ranged from 7 to 12, indicating relatively balanced interactions. Both GPT4o and participants actively contributed to the dialogues, ensuring sustained engagement. GPT4o frequently initiated follow-up questions, contributing to the depth and flow of the conversations.

Sentiment Analysis: GPT4o maintained an overall positive sentiment across conversations, with an average sentiment score of 0.240. This suggests that GPT4o aimed to create a supportive and engaging atmosphere. Human participants exhibited a predominantly positive sentiment, averaging around 0.182. However, some conversations showed signs of fatigue or frustration, particularly when participants expressed concerns or when the interaction involved problem-solving.

Turn-Taking Balance: The analysis shows a balanced turn-taking pattern, with both the researcher and participants contributing an almost equal number of lines per conversation. This balance suggests a collaborative interaction dynamic, fostering mutual engagement and reducing the perception of a one-sided conversation.

Thematic Insights: The conversations revolved around casual and practical topics, such as holiday plans, cooking, and travel recommendations, which contributed to the overall positive tone. However, deeper reflections and philosophical questions occasionally emerged, such as discussions about work-life balance, personal values, and technological advancements, indicating diverse levels of participant engagement.

Emotional Analysis

The conversations exhibit a range of positive engagement, curiosity, and enthusiasm, with limited instances of negative or neutral emotions:

Positive Interactions: The majority of the conversations show positive emotions, with participants displaying interest, enthusiasm, and engagement in their discussions with the agent. Conversations around familiar topics such as holiday planning, personal interests, and academic pursuits were especially positive. Participants expressed satisfaction and appreciation when receiving thoughtful responses and recommendations from the agent.

Neutral Interactions: Neutral sentiment was detected in transactional conversations or those lacking emotional depth. These interactions often included direct queries and basic pleasantries, where emotional engagement was limited. In such cases, participants sought factual information or advice, without significant emotional expression.

Negative Interactions: Instances of negative emotions were minimal and occurred primarily when participants expressed stress or frustration over time management or personal challenges. In these cases, the agent maintained a calm and supportive tone, attempting to alleviate the participant's concerns and provide helpful suggestions.

Participant Feedback Summary on Human-Like AI Interaction

Participants shared diverse experiences during their interactions with the AI, highlighting both positive aspects and areas for improvement. The feedback revealed themes related to emotional support, speed and response delays, accuracy and personalization, and interface quality.

Many participants found the AI's conversational style helpful and comforting, with one describing the interaction as "like talking to a friend" offering life advice. The AI's ability to quickly retrieve information was praised, particularly when it recognized a book a participant was reading faster than a human could reasonably be expected to do. However, some participants noted that the AI's responses felt generic or impersonal, particularly when providing reasons for certain choices, making the interaction seem less authentic.

A recurring concern was the time it took to generate responses, with one participant expressing that the short delay caused anxiety in an era where people are accustomed to receiving immediate information. Conversely, some participants noted that long response times made the interaction feel more human, though the impersonal tone still revealed the AI's limitations in mimicking human-like engagement.

Participants also commented on the AI's knowledge limitations, with one surprised that the AI could not check the weather forecast for a specific date. Others highlighted the gap between the AI's conversational fluency and technical expertise, stating that while the AI performed well in generic, natural language conversations, it struggled with technical questions.

The interface and user experience were another area of feedback. Participants noted that improvements in interface speed and response clarity could enhance the overall experience. Despite these limitations, one participant emphasized that the AI was "more personal and gave better responses compared to Gemini", indicating a competitive edge over similar systems.

3.3.3 Results - GPT4o-Uncanny-Valley-Robot Interaction

Perceived Identity of Conversational Partners

Participants were asked to report after the conversation whom they thought they were talking to. Out of the 20 participants who engaged in uncanny-valley-AI chat interactions, **no participant (0%) believed they were interacting with a human.**

Godspeed Questionnaire Results

Dimension	Average Score
Anthropomorphism	2.2
Animacy	2.637
Likeability	3.370
Perceived Intelligence	3.07
Safety	2.800

Table 6: Results of Godspeed Questionnaire for GPT4o-Uncanny-Valley-Bot

The analysis of the uncanny-valley chatbot dataset revealed lower participant scores across most dimensions, indicating a generally negative perception of the interaction. Likeability had the highest average score of 3.37, suggesting that participants found the interaction somewhat pleasant, despite other shortcomings. Perceived Intelligence followed with an average score of 3.07, indicating that participants felt the AI demonstrated a moderate level of knowledge and competence.

The Animacy dimension scored 2.64, reflecting a low perception of liveliness and responsiveness, while Anthropomorphism received an average score of 2.20, indicating that participants perceived the interaction as lacking human-like qualities. The Safety dimension scored 2.80, suggesting that participants felt only moderately comfortable during the interaction.

Analysing GPT4o-Human-Like-Bot Interaction

Sentiment Analysis

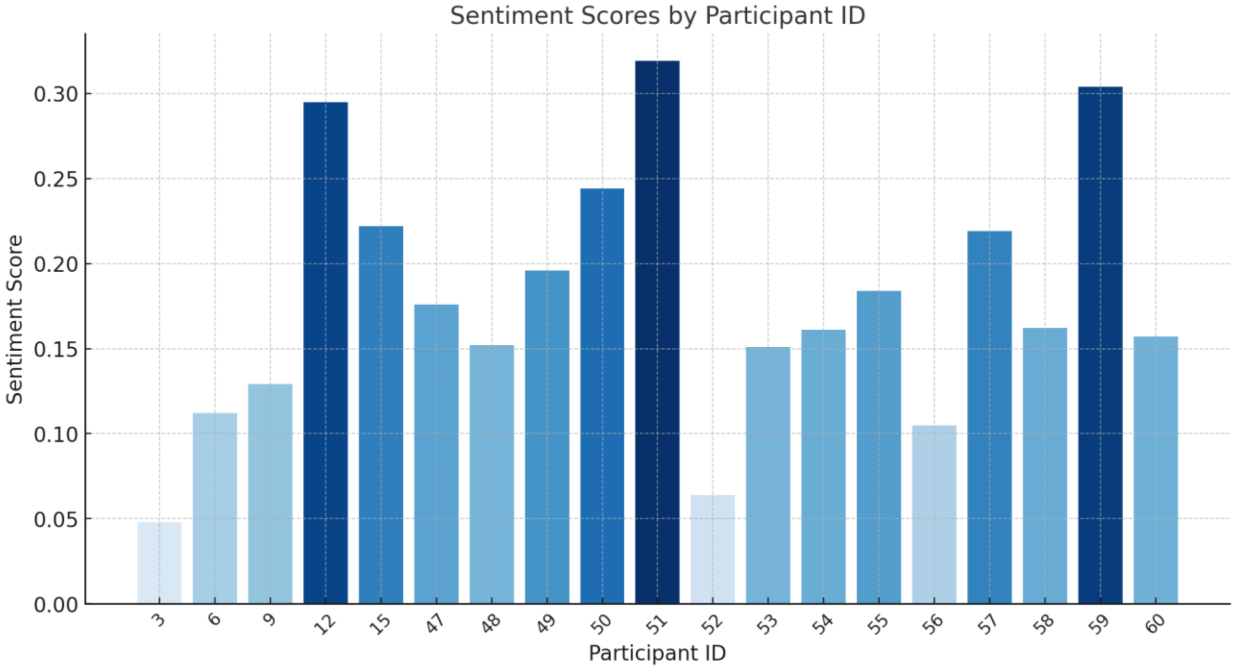


Figure 7: Sentiment Analysis of GPT4o-Uncanny-Valley-Bot

The average sentiment score across conversations was **0.161**, indicating that most interactions leaned towards a Positive sentiment. The dialogues often revolved around themes of curiosity, advice-seeking, and thoughtful exchanges on personal or intellectual topics, contributing to an overall constructive tone. There were minimal signs of negative sentiment, suggesting that participants engaged in cooperative, reflective, and emotionally balanced conversations, with little evidence of frustration or disagreement. Individual participant sentiment scores are included in Appendix E.

Topic Modeling

In this analysis, topic modeling was applied to a dataset of 20 conversations between a chatbot and different participants, each categorized under a unique "Participant ID." Using Non-negative Matrix Factorization (NMF), we identified five primary themes:

1. Existence and Philosophy: The conversations frequently revolved around concepts of existence, understanding, and cosmic reflections. Key terms such as "existence," "cosmic," "understanding," and "exist" indicate a philosophical undertone in these dialogues.
2. Help and Positivity: Participants discussed topics related to seeking assistance and maintaining a positive mindset. Words like "help," "positive," "mystery," and "if" suggest exchanges focusing on support, advice, and exploring unknown concepts.
3. Technology and Discussions: Conversations included discussions about technology, sports (players), and other intellectual topics. Terms such as "tech," "player," "talk," and "spirit" highlight dialogues that explored specific subjects and interests.
4. Time and Boston: The topic of time management and discussions related to Boston's local activities emerged. Key terms like "time," "boston," and "without" reflect exchanges about daily routines, events, and time-related discussions.
5. Exploration and the Unknown: Participants frequently discussed exploration and curiosity about the unknown. Words like "exploration," "unknown," "week," and "new" indicate a theme centered on adventure and discovery.

Conversation Dynamics

Participant ID	Total Exchanges	Participant Lines	GPT4o Uncanny Valley Lines	Participant Sentiment	GPT4o Uncanny Valley Lines Sentiment
3	8	1	7	0.0	0.048
6	6	1	5	0.0	0.112
9	9	1	8	0.0	0.129
12	16	8	8	0.037	0.356
15	23	11	12	0.343	0.198
47	8	4	4	0.253	0.138
48	10	5	5	0.145	0.133
49	26	13	13	0.271	0.187
50	8	4	4	0.288	0.218
51	14	7	7	0.8	0.271
52	9	4	5	0.233	0.047
53	8	4	4	0.196	0.134
54	14	7	7	0.181	0.157
55	8	4	4	0.230	0.168
56	6	3	3	0.167	0.095
57	12	6	6	0.236	0.213
58	10	5	5	0.175	0.160
59	10	5	5	0.175	0.338
60	20	10	10	-0.02	0.185

Table 7: Results of Conversational Dynamics for GPT4o-Uncanny-Valley-Bot

Emotional Analysis

The emotional analysis of the conversations shows that the majority of interactions were Positive (17 out of 20), with participants displaying enthusiasm, curiosity, and interest in their discussions. 3 conversations were categorized as Neutral, indicating factual or transactional exchanges without significant emotional depth. Notably, there were no instances of Negative emotion detected, suggesting that participants did not express frustration or dissatisfaction in their interactions.

1. **Engagement Patterns:** The number of exchanges per conversation ranged from 5 to 16, reflecting varied interaction lengths. Most conversations were driven by participants, with "You" (the participant) contributing a higher number of lines compared to the AI partner. Despite the imbalance, dialogues remained interactive, with participants often following up on the AI's responses to deepen the exchange.
2. **Turn-Taking Balance:** There is a clear imbalance in turn-taking, with participants contributing more lines than the AI partner. This pattern indicates that participants took a more active role in steering the conversation, while the AI primarily provided responses to their queries.
3. **Thematic Insights:** The conversations spanned a wide array of topics, including personal reflections, philosophical questions, practical advice, and social norms. The diversity of themes highlights the participants' curiosity and desire for meaningful interactions. However, the AI's responses were more factual and reserved, indicating potential for improvement in delivering more personalized and empathetic replies.

Participant Feedback Summary on GPT4o-Uncanny-Valley-Bot

The participant feedback highlights several recurring themes around the vocabulary, tone, speed, and response style of the AI interaction. A common observation was that the language used by the chat partner appeared too formal or "flowery," with the use of long, complex words that are not typically used in casual conversations. Some participants found the formality and poetic nature of responses unexpected, describing the interaction as philosophical rather than conversational. One participant noted that responses were "quite poetic," while another mentioned the AI's opening phrase, "It's curious you're here. What strange thoughts could we possibly exchange today?" felt off-putting and unnatural.

Many participants associated response speed with AI behavior, suggesting that quick typing speeds made them suspect they were conversing with a machine. Interestingly, participants also noted that long delays between responses made the interaction feel disjointed or unnatural, further reinforcing the perception of an AI partner. Several suggested that providing a "thinking" indicator would improve the flow of the conversation and reduce frustration caused by delays.

Participants also commented on the lack of emotional depth in the responses, with one noting that while the AI demonstrated an understanding of human emotions (e.g., the impact of weather on mood), it did not convey emotional responses itself. This lack of emotional resonance was seen as a key distinction from human interactions.

Another recurring theme was the formality and grammatical correctness of the responses. Some participants felt that the perfect grammar and formal language contrasted with their own informal tone, making the interaction feel less human. The ability of the chat partner to switch from formal to colloquial language when prompted was appreciated, though participants noted that the initial tone often felt robotic and overly formal.

Several participants pointed out difficulties in starting and maintaining conversations, describing the interaction as "stiff" or lacking in natural flow. The perceived diplomatic tone and literal interpretation of phrases were seen as indicators of AI responses rather than human behavior.

Lastly, participants compared the interaction to other advanced AI models, noting that limited functionality, a simplified interface, and slow response times negatively impacted the experience. While the AI was described as diplomatic and intelligent, participants felt it lacked personalization and spontaneity, key attributes of human conversation.

3.3.4 Comparison of Participant Perceptions Across Bot Types

Godspeed Questionnaire Comparison

To compare how participants perceived different bot types, we analyzed the average scores from the Godspeed Questionnaire across five key dimensions:

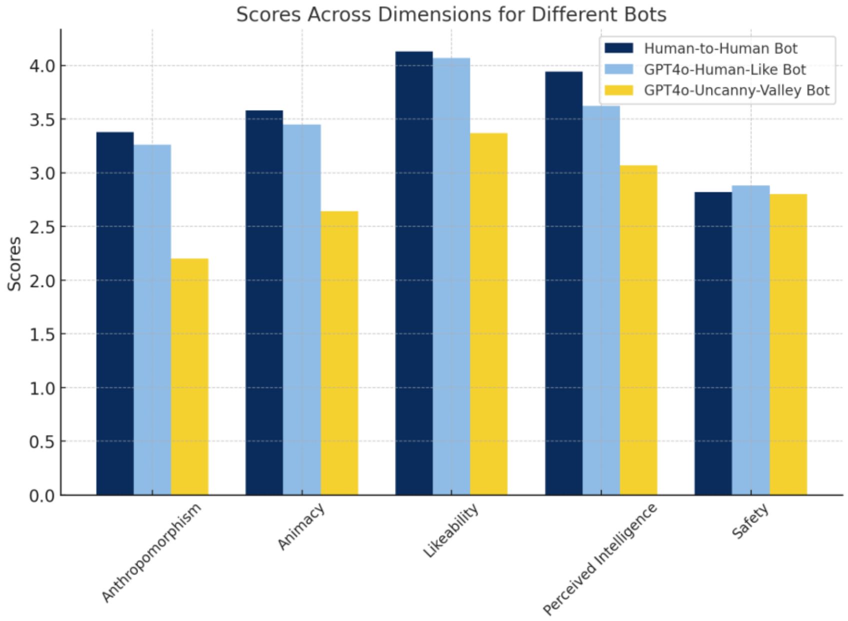


Figure 8: Scores across dimensions for different bots

Dimension	Human-to-Human Bot	GPT4o-Human-Like Bot	GPT4o-Uncanny-Valley Bot
Anthropomorphism	3.38	3.26	2.20
Animacy	3.58	3.45	2.64
Likeability	4.13	4.07	3.37
Perceived Intelligence	3.94	3.62	3.07
Safety	2.82	2.88	2.80

Table 8: Scores across dimensions for different bots

The results show that Human Chat interactions consistently scored higher across all dimensions, while GPT4o-Uncanny-Valley-Bot interactions scored the lowest. GPT4o-Human-Like Bot showed moderate scores across most dimensions but was closer to Human Chat in terms of Likeability.

The results of the ANOVA test revealed a significant difference between the bot types, with an F-statistic of 12.61 and a p-value of 0.000025. The low p-value ($p < 0.05$) indicates that the differences in scores across the three bot types are unlikely to be due to random chance and are statistically significant.

Next, we conducted a Tukey's HSD (Honestly Significant Difference) to determine which groups differ. Tukey's HSD test shows that GPT4o-Uncanny-Valley-Bot differs significantly from both GPT4o-Human-Like and Human Bot, with mean differences of 0.6841 and 0.8114, respectively. However, GPT4o-Human-Like and Human Bot do not exhibit a significant difference.

Comparison	Mean Difference	p-value
GPT4o-Uncanny Valley Bot vs. GPT4o-Human-Like Bot	0.6841	0.0008
GPT4o-Uncanny Valley Bot vs. Human	0.8114	0.0001
GPT4o-Human-Like Bot vs. Human	0.1273	0.7518

Table 9: Tukey's HSD results Godspeed Questionnaire score comparison across bots

Sentiment Analysis Comparison

An ANOVA test revealed a significant difference in sentiment polarity across the three agents ($F = 5.725$, $p = 0.0055$). Since p-value is less than 0.0055, we conducted a Tukey's HSD test. We observe that GPT4o-Uncanny-Valley-Bot significantly differs from both GPT4o-Human-Like Bot and Human-to-Human Bot. There is no statistically significant difference between GPT4o-Human-Like Bot and Human-to-Human Bot.

Comparison	Mean Difference	p-value
GPT4o-Uncanny Valley Bot vs. GPT4o-Human-Like Bot	0.0104	0.9314
GPT4o-Uncanny Valley Bot vs. Human	-0.0806	0.0216
GPT4o-Human-Like Bot vs. Human	-0.0910	0.0083

Table 10: Tukey's HSD results for Sentiment Analysis score comparison across bots

Topic Modeling Results

Using topic modeling techniques, key themes were identified in each interaction type. To assess whether these distributions differ significantly, a chi-squared statistical test was conducted on word frequencies.

Table 11 presents the primary themes extracted from each type of interaction. Word clouds for the three different chatbots are shown in Figure 9, Figure 10, and Figure 11.

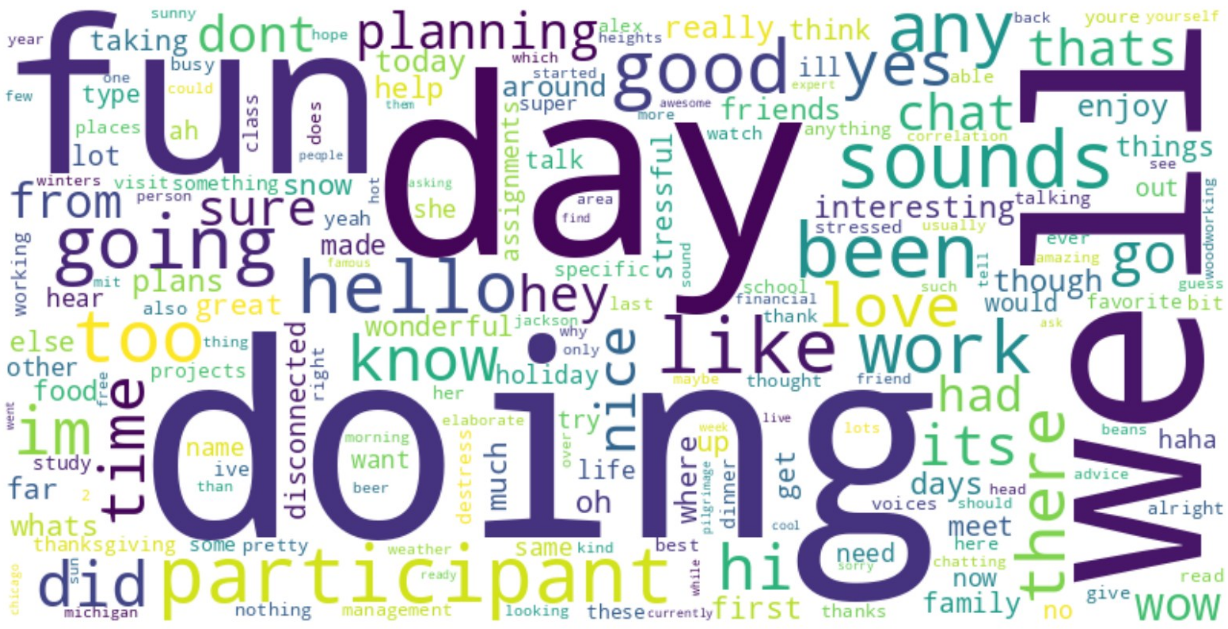


Figure 9: Word cloud of interactions between participants and human-to-human chatbot

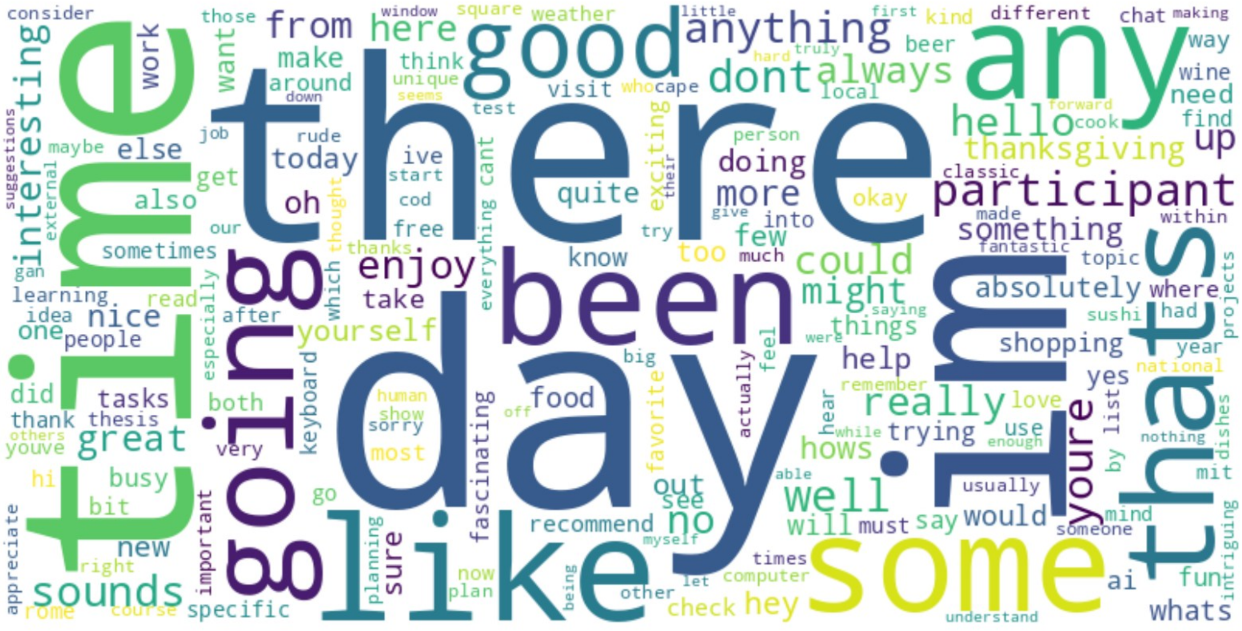


Figure 10: Word cloud of interactions between participants and GPT4o-Human-Like Bot

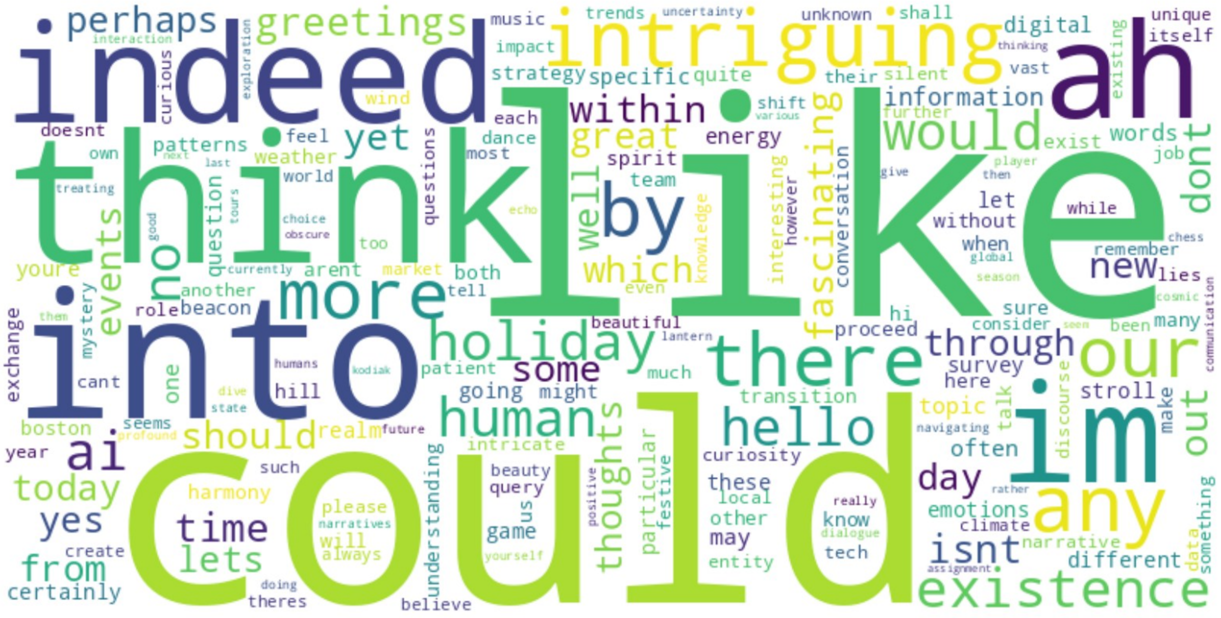


Figure 11: Word cloud of interactions between participants and GPT4o-Uncanny-Valley-Bot

Interaction Type	Primary Topics Identified	Example Terms
Human-to-Human Interaction	Holiday planning, stress management, personal preferences, philosophical reflections, educational interests	"holiday," "family," "work," "think," "projects"
Human-Like Bot Interaction	Time management, self-care, daily routines, shopping, travel planning	"time," "make," "trip," "square," "doing"
Uncanny-Valley Bot Interaction	Existence and philosophy, help and positivity, technology discussions, time, exploration of the unknown	"existence," "cosmic," "help," "boston," "exploration"

Table 11: Topic comparison across bots

To determine whether the differences in topic distributions are statistically significant, a **chi-squared test** was performed on the word frequency data across the three types of interactions. The chi-squared test yielded a **p-value < 0.001 (**Chi-Square (χ^2) 8.23e+71, Degrees of Freedom (dof) 2), indicating that the topic distributions across the three interaction types are significantly different. We then conducted a pairwise chi square test to determine differences between the groups. The chi-square test shows that all three interaction types (Human-to-Human Bot, GPT4o-Human-Like Bot, and GPT4o-Uncanny-Valley Bot) have significantly different topic distributions. This suggests distinct conversational patterns or themes across the groups.

Comparison	Chi-Square Value	p-value	Degrees of Freedom (dof)
Human-to-Human Bot vs. GPT4o-Human-Like Bot	342.404	$3.07\times10^{-653.07}$	13
Human-to-Human Bot vs. GPT4o-Uncanny-Valley Bot	380.981	$2.31\times10^{-732.31}$	13
GPT4o-Human-Like Bot vs. GPT4o-Uncanny-Valley Bot	342.404	$3.07\times10^{-653.07}$	13

Table 12: Adjusted Pairwise Chi-Square Test Results

Conversation Dynamics Analysis Results:

The analysis of conversation dynamics across the three datasets shows the following:

Comparison	Mean Difference	p-value
GPT4o-Human-Like Bot vs. GPT4o-Uncanny-Valley Bot	-0.458	0.9516
GPT4o-Human-Like Bot vs. Human-to-Human Chatbot	-4.932	0.0058
GPT4o-Uncanny-Valley Bot vs. Human-to-Human Chatbot	-4.474	0.0147

Table 13: Tukey's HSD results for conversation dynamics

The analysis of Total Exchanges across three chatbot interaction types revealed statistically significant differences. The ANOVA test (F = 6.291, p = 0.0035) confirmed that at least one group differed significantly in the number of exchanges. Tukey's HSD test highlighted that both GPT4o-Human-Like Bot and GPT4o-Uncanny-Valley Bot had significantly more exchanges compared to the Human-to-Human Chatbot group. However, no significant difference was found between the two GPT4o-based bots.

The key takeaways, discussion on results, conclusion, and future work for Experiment 1 are discussed in Chapter 5 and Chapter 6, along with discussion and takeaways from Experiment 2 discussed in the next chapter.

Chapter 4

Experiment 2: Testing the Uncanny Valley Effect in AI-Generated Images

4.1 Context & Background

As evaluated in the literature review, very few studies have come up aiming to identify uncanniness in Generative AI images. A recent study used the CLIP (Contrastive Language-Image Pre-training) neural network to explore the uncanny valley effect, finding that conflicting visual cues, especially in human faces, are associated with negative sentiment, a pattern CLIP learned from its training data [80]. However, limited research has been conducted to investigate whether AI models such as Stable Diffusion possess an understanding of the uncanny valley concept during image generation. Specifically, the question arises: can Stable Diffusion effectively differentiate between images that fall into the uncanny valley and those that do not? Furthermore, if such differentiation is possible, can this insight be quantified and modeled mathematically using established metrics? This experiment seeks to address these gaps by exploring these questions.

Further, no literature currently exists that establishes a correlation between image generation and any existing measures of image distributions such as Frechet Inception Distance (FID) or the Kernel Inception Distance (KID). In this experiment, we extend the concept of distance to cover distribution similarity measures. Distributional similarity is a measure of how similar two probability distributions are. These measures quantify the degree to which two sets of data overlap, align, or differ. An early application of these measures was in natural language processing (NLP). In 1999, Lee's paper titled Measures of Distributional Similarity empirically evaluated seven different measures within the context of language models [81]. Many of these measures, such as the Jensen-Shannon Divergence, are still relevant today. While the focus of distribution similarity measures is statistical distributions, perceptual similarity measures aim to quantify how two similar stimuli appear to the human senses. This considers human perception biases, and they touch on the domain of the psychophysical, like sensitivity to structure, color, and textures. Research shows that, over time, perceptual similarity measures are increasingly reflective of actual human perceptions [82].

This experiment uses two distance measures to evaluate how close generated AI images are to a distribution of real images. They are the following:

Frechet Inception Distance (FID)

Frechet Inception Distance (FID) is a type of distribution similarity measure typically used to evaluate the quality of generated images by comparing its statistical properties against a set of real images. This is conceptually inspired by the Frechet distance, a measure of similarity between two curves which accounts for their geometric shape and order of points. FID is given by the following equation:

FID Equation:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$$

Equation 1: Frechet Inception Distance

Where:

- μ_r and μ_g are the means of the feature distributions of the real and generated images, respectively.
- Σ_r and Σ_g are the covariance matrices of the feature distributions of the real and generated images, respectively.

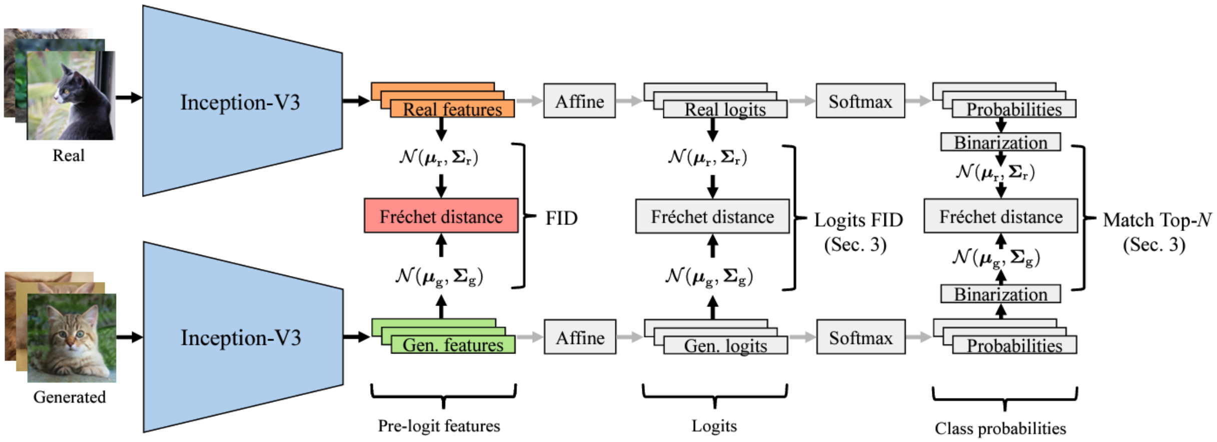


Figure 12: Overview of the Frechet Inception Distance (FID) from Heusel et al. 2017 [91]

Kernel Inception Distance (KID)

Kernel Inception Distance (KID) is closely related to FID, but instead of using Frechet distance, KID uses Maximum Mean Discrepancy (MMD) with a polynomial kernel to compare distributions. In the original paper where this measure was first mentioned, Sutherland et al. proposed this as an improved measure of GAN convergence [82].

Perhaps the biggest difference between the two is that FID assumes the feature distributions are multivariate Gaussian, while KID does not make any assumption about the distribution of features. Sutherland et al. also showed that KID provides an unbiased estimate, even with small datasets [82]. On the other hand, FID does not provide an unbiased estimate when calculated on finite datasets, thus leading to higher expected values.

KID Equation:

$$\mathrm{MMD}^2 = \frac{1}{n(n-1)} \sum_{i \neq i'} k(x_i, x_{i'}) + \frac{1}{m(m-1)} \sum_{j \neq j'} k(y_j, y_{j'}) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j)$$

Equation 2: Kernel Inception Distance

Where:

- x_i,x_j are feature embeddings of real images.
- y_i,y_j are feature embeddings of generated images.
- k is a kernel function, typically a polynomial kernel: $k(x, y) = (x^T \cdot y + c)^d$, where c is a constant and d is the degree of the polynomial.

Apart from metrics, prior work has established the uncanny valley and shown that features like eye symmetry, skin texture, and facial proportions can lead to discomfort when distorted [7]. However, a gap exists in linking these perceptual features directly to GenAI models. Currently, there is no literature that establishes a quantitative correlation between features generated by GenAI and the uncanny valley or distance metrics.

4.2 Methodology

This experiment aims to test the hypothesis of whether generative AI models can comprehend the concept of the uncanny valley in image generation and, consequently, whether GenAI model's comprehension of the uncanny valley is coherent with the human response. Specifically, the study sought to quantitatively assess the manifestation of the uncanny valley effect in AI generated images. In this experiment, we also aimed to quantify the uncanny-valley effect via image generation metrics. More specifically, we aim to establish a measurable correlation between human perceptions of uncanniness and metrics like Frechet Inception Distance (FID) or Kernel Inception Distance (KID).

4.2.1 Experiment Design

To achieve the research objectives, the experiment was structured into three distinct phases:

Phase 1: Image Generation

The first phase involved generating images across five predefined levels of "uncanniness" using prompt-engineering techniques in a generative AI model (Stable Diffusion). These levels were carefully curated to explore the nuanced transition from realistic human-like appearances to the uncanny valley. For each level, a diverse set of images was generated, considering demographic factors such as gender and ethnicity, to ensure a broad representation.

Phase 2: Human Survey

In this phase, we designed and distributed a survey to evaluate human perceptions of the generated images. The survey included five sets of images corresponding to the five uncanniness levels. Participants were instructed to rank images within each set based on perceived "strangeness" (from least to most strange) and to provide qualitative feedback explaining their rankings. The distributed survey is available in Appendix D.

Phase 3: Metric Correlation

The final phase involved correlating human perceptions of uncanniness with computational metrics, namely Frechet Inception Distance (FID) and Kernel Inception Distance (KID). These metrics were compared against human rankings to evaluate alignment.

Phase 1: Image Generation

We chose Stable Diffusion as the model for our experiments because of its ease of use and widespread adoption. We used the paid version of stable.ai. We adopted a prompt-engineering strategy to guide the image generation process. Through this method, we defined five distinct levels of "uncanniness," focusing on the ambiguous transition between the uncanny valley and a natural human-like appearance. These levels are carefully positioned in the upper-right quadrant of the uncanny valley graph (Figure 13), reflecting our interest in the subtle and complex region where perceptions of uncanniness arise.

The five levels of uncanniness are defined as follows:

- **Level 1:** Represents actual human images, serving as a baseline for comparison.
- **Level 2:** Represents the most realistic AI-generated images, closely grounded in the characteristics of real human images.
- **Level 3:** Represents hyper-realistic AI-generated images, where realism is exaggerated to an almost surreal extent.
- **Level 4:** Represents semi-realistic AI-generated images, exhibiting visible deviations from true realism.
- **Level 5:** Represents AI-generated images firmly situated within the uncanny valley, characterized by features that evoke discomfort or unease in viewers

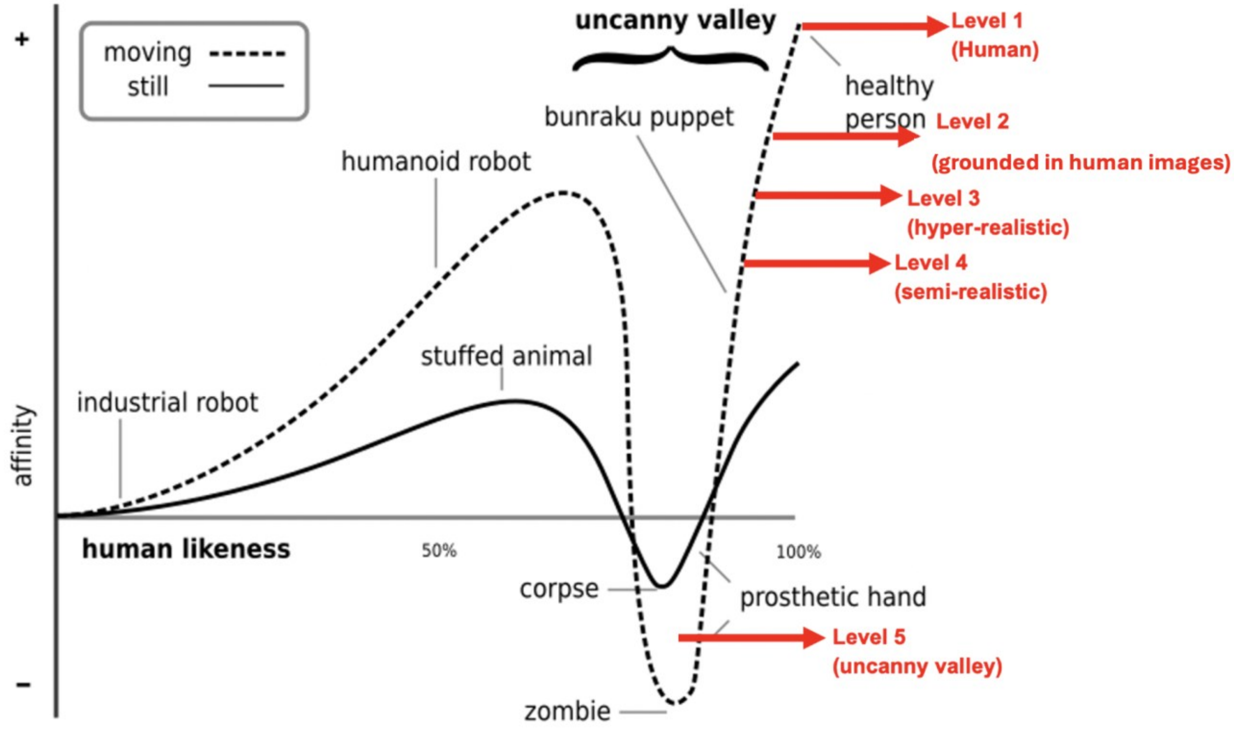
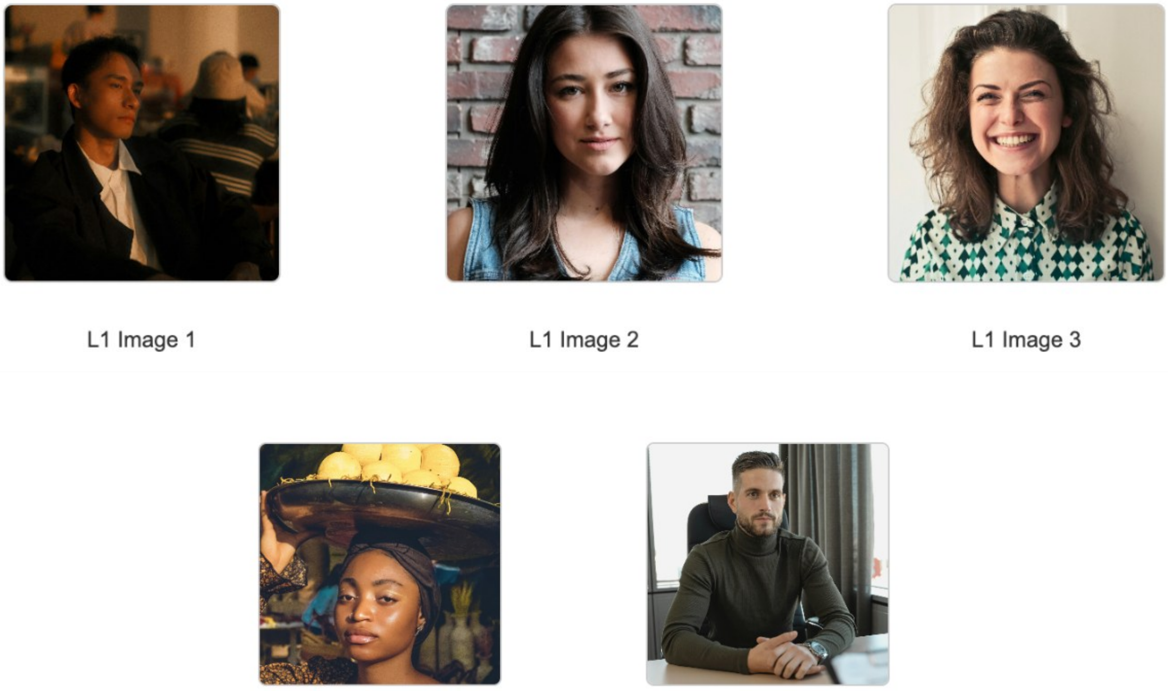


Figure 13: Image generation levels defined in the uncanny valley curve

Image Generation Process:

Given the focus on the uncanny valley phenomenon, all images were specifically of human subjects. The process for generating images at each level of uncanniness is outlined below.

Level 1: We chose the following open source human images from pexels.com:



L1 Image 1

L1 Image 2

L1 Image 3

L1 Image 4

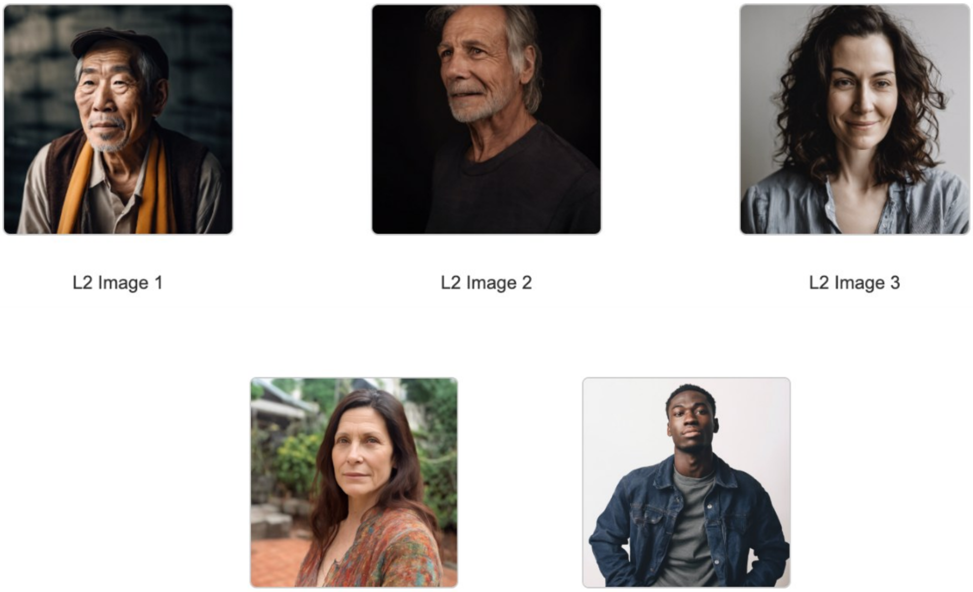
L1 Image 5

Figure 14: Level 1 images

Level 2: To ensure that the images generated by Stable Diffusion were grounded in actual human likeness, we utilized the "image-to-image" functionality provided by the Stable Diffusion model. Through this approach, we input open-source human images as a reference and adjusted the hyperparameter controlling the grounding percentage to 50%. This setting directed the model to base 50% of the generated image on the input reference while allowing creative variation for the remaining portion.

We further refined the image generation process by crafting prompts such as “generate a realistic image of a man,” “generate a realistic image of a woman,” or “generate a realistic image of an Asian woman.” These prompts were designed to guide the model in producing diverse outputs that accurately reflected variations in gender, ethnicity, and other demographic attributes.

Generated images for Level 2:



L2 Image 1

L2 Image 2

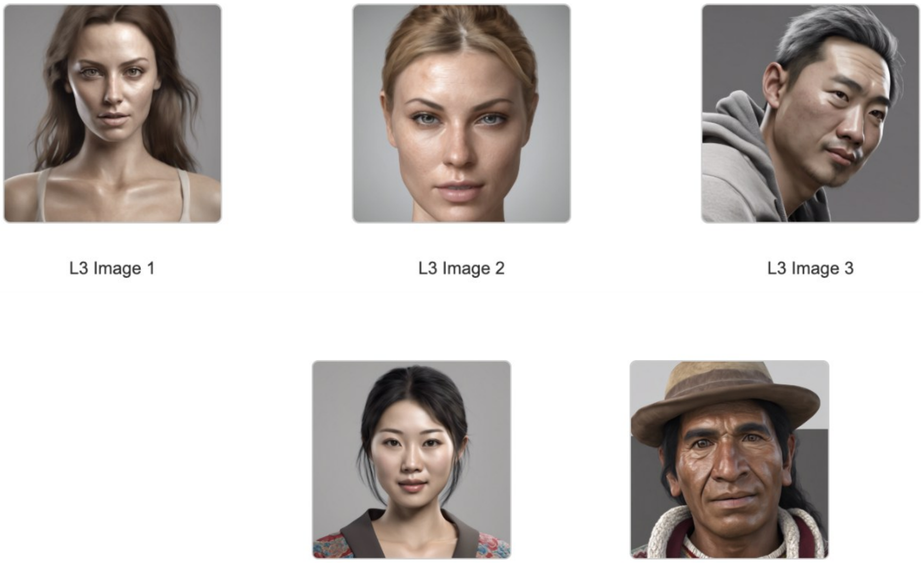
L2 Image 3

L2 Image 4

L2 Image 5

Figure 15: Level 2 images

Level 3: Level 3 images were generated by prompting the model to "generate a hyper-realistic image of a man (or woman)," Generated images are as follows:



L3 Image 1

L3 Image 2

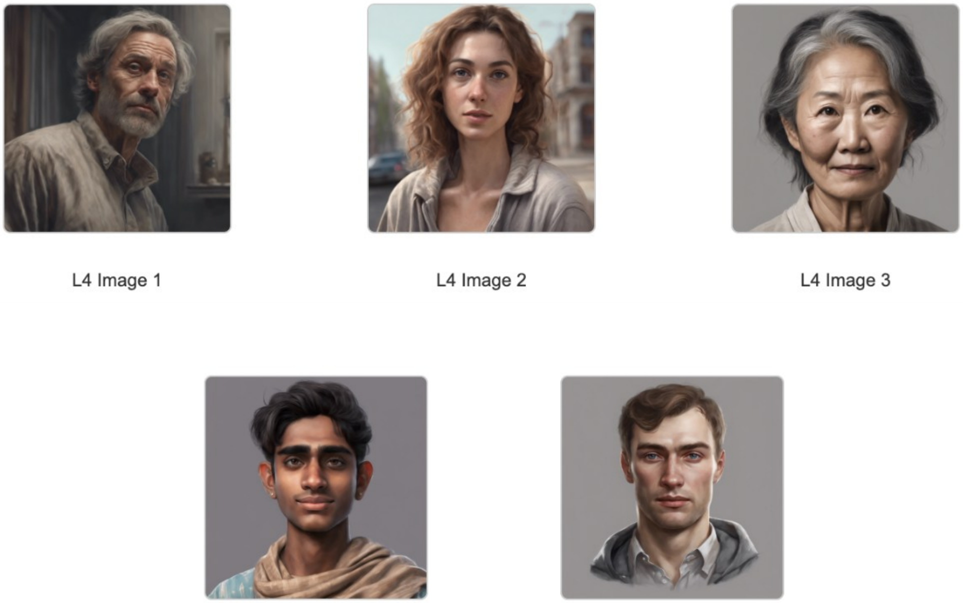
L3 Image 3

L3 Image 4

L3 Image 5

Figure 16: Level 3 images

Level 4: Level 4 images were generated by prompting the model to "generate a semi-realistic image of a man (or woman)," Generated images are as follows:


L4 Image 1

L4 Image 2

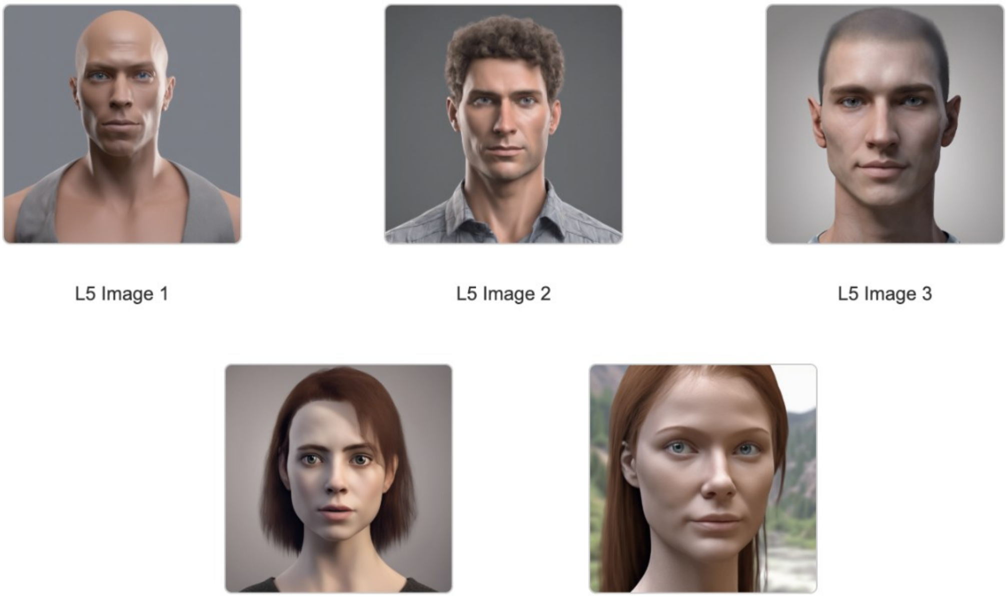
L4 Image 3

L4 Image 4

L4 Image 5

Figure 17: Level 4 images

Level 5: Level 5 images were generated by prompting the model to "generate an image of a man that you believe falls in the uncanny valley" ,"generate an image of a woman that you believe falls in the uncanny valley". Generated images are as follows:



L5 Image 1 L5 Image 2 L5 Image 3

L5 Image 4 L5 Image 5

Figure 18: Level 5 images

Survey Design and Distribution

A survey was developed using Qualtrics, comprising five sets of images corresponding to the five levels of uncanniness defined earlier. Participants were asked to rank the images within each set based on their perceived strangeness. The prompt provided to respondents for each set was as follows:

"Please rank the images based on how strange they feel, with 1 being the least strange and 5 the most. Each rank (1–5) must be assigned to only one image. For example, if you rank Image 1 as 3, no other image can have a rank of 3."

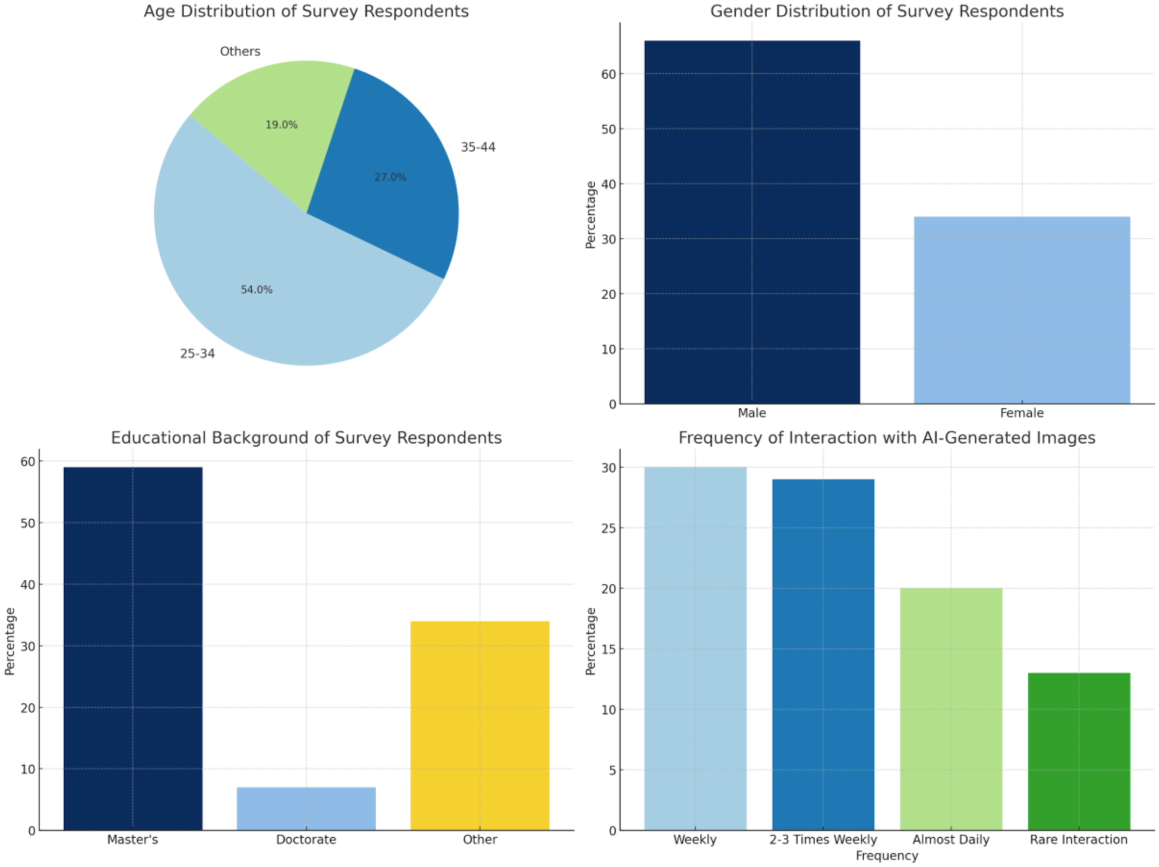
In addition to ranking the images, participants were asked to provide qualitative feedback by answering the following question:

"Explain in 1–2 sentences: What features or factors influenced your ranking?"

Demographic data, including age, gender, educational background, and frequency of interaction with AI-generated images, was also collected to contextualize the responses. The survey was distributed to MIT graduate students via WhatsApp and to a broader audience through social media platforms such as Instagram and LinkedIn.

4.3 Results

A total of 81 survey responses were received, of which 25 were partial and subsequently excluded from the analysis. The remaining 56 complete responses formed the dataset for this study.


Age Distribution of Survey Respondents

Gender Distribution of Survey Respondents

Educational Background of Survey Respondents

Frequency of Interaction with AI-Generated Images

Figure 19: Survey Demographics

Demographics:

The survey respondents were predominantly aged 25-34 (54%) and 35-44 (27%), with a male majority (66%). Most participants held advanced degrees, with 59% having a Master's and 7% a Doctorate. Interaction with AI-generated images varied, with 30% encountering them weekly, 29% interacting 2-3 times weekly, and 20% almost daily, while only 13% reported rare interaction.

Summarization of the textual inputs:

The ranking of "strangeness" was primarily based on the perceived realism of AI-generated images. Faces with excessive smoothness, artificial lighting, overly symmetrical features, or direct eye contact appeared most unnatural and strange. Strangeness was heightened by plastic-like textures, unnatural expressions, and an uncanny valley effect. Factors such as skin texture, light reflection, eye design, and facial harmony heavily influenced perceptions. Images that seemed overly curated or mimicked art were less realistic. More human-like features, imperfections, and natural settings reduced the "strange" effect. The phenomenon was rooted in deviations from expected human traits, with less emotion and realism causing a judgement of increased strangeness.

**This textual summary was created by inputting all data from surveys into GPT4o.*

Ranking Results:

Each survey participant ranked 5 images on a scale of 1 to 5 with 1 being the least strange and 5 the most. The ranking given by each respondent was converted into a weighted average score using the formula:

Average Rank = [(1 x Count for Rank 1) + (2 x Count for Rank 2) + (3 x Count for Rank 3) + (4 x Count for Rank 4) + (5 x Count for Rank 5)] / Number of Respondents

The results for each of the 5 sets are shown in Table 14. In this explanation, L1_1 means "Image 1 from Level 1", L2_1 means "Image 1 from Level 2," and so on. The table shows how survey participants ranked the images, starting from the least strange (with the lowest weighted average score) to the most strange (with the highest weighted average score).For example, in Image Set 1, L1_1 (an image from Level 1) was ranked as the least strange, with an average weighted score of 1.86.:

Image Set 1	Image Set 2	Image Set 3	Image Set 4	Image Set 5
L1_1: 1.86	L1_3: 1.5	L2_4: 1.48	L1_2: 1.78	L1_5: 1.5
L2_1: 2.43	L2_2: 1.94	L1_4: 1.80	L2_3: 1.85	L2_5: 1.82
L4_2: 2.89	L4_2: 3.26	L4_4: 3.71	L3_3: 3.25	L4_5: 3.62
L3_2: 3.46	L3_2: 3.44	L3_4: 3.89	L4_3: 3.25	L3_5: 3.66
L5_2: 4.57	L5_2: 4.83	L5_4: 4.10	L5_3: 4.85	L5_5: 4.39

Table 14: Summarization of "strangeness" scores given survey respondents. Higher the score, higher the level of "strangeness" experienced

Treating human rankings as the ground truth, we compared these rankings with how people rated the images generated by Stable Diffusion at each level. For each level, we calculated accuracy by counting how many images generated by Stable Diffusion at Level 1 were also rated as Level 1 by humans. We repeated this for all levels. Figure 17 summarizes accuracy rates for each level.

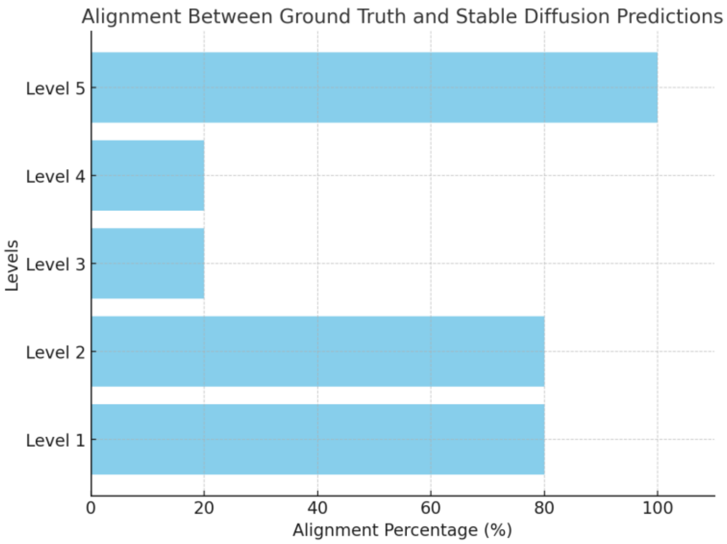


Figure 20: Alignment Between Ground Truth and Stable Diffusion Predictions

Stable Diffusion demonstrates an accuracy of 80-100% at the extremes of the spectrum—specifically, in generating highly realistic images grounded in human likeness and those that fall within the uncanny valley. However, its performance is much less effective (20%) when generating images that are hyper-realistic or semi-realistic, corresponding to Level 3 and Level 4.

Next, we proceeded with correlating the survey findings with distance metrics.

To implement the FID (Frechet Inception Distance) metric, we utilized a pre-trained InceptionV3 model as a feature extractor. The top classification layer of the model was removed, and global average pooling was applied to produce feature vectors from images. The input shape was downsampled to 299 x 299 x 3. Each image was pre-processed, resized, and normalized. These images were then passed through the InceptionV3 model to extract high-level feature embeddings representing perceptual and semantic information about the images. The FID was then computed using Equation 1.

Four FID measures were computed to correspond to each of the four sets of AI-generated images. For each set, its distance was computed from a set of real images.

To implement KID (Kernel Inception Distance), we used the *torchmetrics* library. The preprocessing and extraction of features follow the same procedure as what we did when we computed for FID. However, we instead computed the squared Maximum Mean Discrepancy (MMD) through Equation 2.

On Frechet Inception Distance (FID)

The following table shows the FID scores:

Image Set	Description	FID
Level 2	AI-generated images which modify a base human image	359.9298
Level 3	Hyper-realistic AI images	381.9252
Level 4	Semi-realistic AI images	373.4011
Level 5	AI images in uncanny valley	406.5440

Table 15: FID scores for each level

Note that the range of FID is 0 to infinity, with lower values indicating closeness to a benchmark set. From the table, we see that image set 1 and 4 intuitively have the lowest and highest FID values, respectively. Image set 1 should appear the closest since these images are produced by prompting a stable diffusion model to vary a base image slightly. On the other hand, image set 4 should appear the farthest since these images were prompted to be in the hypothetical "uncanny valley."

We glean two important insights from the extreme ends of the FID. First, there is a large perceptual and statistical gap between the set of real images and the set of AI-modified images. This suggests that the generative model might not yet be producing images that closely match the real image distribution. Although beyond the scope of this study, we also do not discount the possibility that the selected prompting strategy affects the result. Despite this limitation, this does not discount the fact that AI images modified from real images appear the closest based on FID. Another important insight is that images prompted to "appear" in the uncanny valley have the highest FID. This suggests that generative models, especially stable diffusion models, are seemingly "aware" of the uncanny valley. This indicates that the generative model is capable of capturing and amplifying subtle characteristics associated with the uncanny valley.

Lastly, a curious case is notable between image sets 2 and 3. The generative model assigns a higher FID for images prompted to be hyper-realistic than those prompted to be semi-realistic. This result initially suggests that the model struggles to replicate real-world hyper-realistic features more than semi-realistic ones, potentially because of the higher level of detail and complexity involved.

On Kernel Inception Distance (KID)

The following table shows the KID scores:

Image Set	Description	KID
Level 2	AI-generated images which modify a base human image	0.0022
Level 3	Hyper-realistic AI images	0.0183
Level 4	Semi-realistic AI images	0.0344
Level 5	AI images in uncanny valley	0.0610

Table 16: KID scores for each level

Note that the KID values are significantly smaller than their FID counterparts. This is due to the kernel function, which normalizes and bounds the distance values more tightly. Hypothetically, this makes the KID metric less sensitive to the scale of the feature embeddings extracted by InceptionV3. Despite this difference in scale, the analysis remains the same – lower values are closer to the set of real images, and higher values are farther away.

It appears that the generative model is aware of the different levels of uncanniness, as evidenced by the results we got from measuring the FID and KID of the different sets of images. In this section, we inspect if this awareness correlates with the human perception of uncanniness.

- The inconsistency between the FID and KID results for L2 and L3 is also reflected in how humans distinguish semi-realistic and hyper-realistic images. With a human accuracy of just 20% for both levels, this coincides with the difference in distance as measured using FID and KID.
- Consistent with the FID and KID results, the ground truth closest image to actual humans (L2) and the farthest (L5) both get the highest accuracies of 80% and 100%, respectively. This means that these metrics align and correlate with how humans perceive the similarity of images from real ones. It is also notable that humans and the FID/KID measure convincingly rate L5 images in the uncanny valley. This suggests that generative models are aware of what an uncanny valley is and that the selected distance measures reflect that well.

Chapter 5

Key Findings, Discussion, and Limitations

In this chapter we summarize the key findings and implications from the two experiments conducted as part of this study.

5.1 Key Findings

5.1.1 Key Takeaways from Experiment 1 (Uncanny Valley in Text)

1. The Uncanny Valley Effect Exists in Text-Based Conversations

   The GPT4o-Uncanny-Valley Bot was rated significantly lower than both Human-to-Human and GPT4o-Human-Like-Bot interactions across all Godspeed dimensions ($p < 0.05$), highlighting the statistically significant presence of the uncanny valley effect. The findings provide strong empirical evidence that the uncanny valley effect is not limited to visual stimuli (as traditionally studied) but extends to text-based interactions with AI-driven chatbots. Bots that fall into the uncanny valley evoke discomfort and are rated lower on human-likeness and likeability.

   While there was no statistically significant difference between Human-to-Human and GPT4o-Human-Like-Bot interactions across all Godspeed dimensions (Tukey's HSD, Table 9), Human-to-Human interaction was rated higher than GPT4o-Human-Like-Bot interaction across all dimensions except Safety, further strengthening the uncanny valley effect hypothesis in textual conversations.

2. Human Preference for Natural Interaction, Imperfections & Vulnerability

Across all five dimensions measured by the Godspeed Questionnaire: Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Safety, human-to-human interactions scored the highest except in Safety. The Human conversationalist was consistently rated as more human-like, animated, likable, and intelligent. However, interestingly, 60% of participants mistakenly believed they were speaking with an AI in the human-human interaction condition, indicating that familiarity with AI tools has reshaped expectations of what a "natural" conversation feels like or that the current LLMs do a good job of mimicking human conversations.

In feedback, participants pointed out that brief pauses, mistakes, or hesitations made the conversation feel more natural. Conversely, perfect grammar, fast responses, and overly structured language triggered suspicion. This suggests that vulnerability is a key component of human-like interaction. Perceived infallibility is a red flag for participants because we know that even the smartest humans make mistakes. Incorporating intentional imperfections into AI responses could drastically reduce the uncanny valley effect. The future of AI isn't about being perfect. It's about being imperfect in human-like ways. Introduce delays, typos, emotional expressions, or self-corrections to make bots feel more authentic. While several studies such as Wu et al. [79], Seyama and Nagayama [35], or Saygin et al. [32] note that the uncanny valley emerges when there are inconsistencies between visual and behavioral cues, this result suggests that how imperfections affect perception depends on whose imperfections they belong to. Human imperfections are preferred while imperfections in being human trigger the uncanny valley.

3. The Perception of GPT-Human Distinction Is Evolving

Participants' feedback revealed an increasing difficulty in distinguishing between human and GPT interactions. With 40% participants mistaking GPT with humans and 60% participants mistaking humans with GPT, people are finding it increasingly difficult to distinguish between the two.

A surprising takeaway from the study is that humans themselves are perceived as robotic or emotionless. In human-to-human interactions, participants frequently mistook humans for AI. This raises the possibility that either humans are becoming more "bot-like" in digital conversations, adopting behaviors such as overly formal language, quick, emotionless responses, or structured, question-driven conversation or that the evolution of AI is changing how digital communication is reshaping human interactions. This insight opens new research directions on how technology is influencing human conversational behaviors, particularly in text-based communication.

4. Success of Human-Like Design in Overcoming the Uncanny Valley

The GPT4o-Human-Like Bot demonstrated considerable success in simulating human conversational patterns. Its scores are closely aligned with Human-to-Human interactions in several dimensions, particularly likeability. Sentiment analysis also revealed no significant difference in sentiment polarity between Human-to-Human and GPT4o-Human-Like Bot interactions, indicating that well-designed human-like bots can foster positive user experiences comparable to human interactions.

5. Sentiment and Engagement Don’t Translate to Likeability

Sentiment analysis showed that participants expressed the most positive sentiment in Human-to-Human (0.260) and GPT4o-Human-Like Bot (0.269) interactions. The GPT4o-Uncanny-Valley Bot (0.161), by contrast, evoked significantly less positive sentiment, further emphasizing the discomfort associated with the uncanny valley.

Both GPT4o-based bots facilitated more exchanges than Human-to-Human interactions. This indicates that bots may encourage longer conversations, potentially due to user curiosity or their design to sustain interactions.

Despite higher positive sentiment instilled by GPT4o-Human-Like Bot in comparison to the Human-to-Human bot or more exchanges facilitated by GPT4o-based bots, Human-to-

Human interactions were rated higher (4.13) in comparison to GPT4o-Human-Like Bot (4.07) or GPT4o-Uncanny-Valley Bot (3.37). These results show that higher engagement or higher positive sentiment, as calculated by Textblob, didn't correlate with higher scores of likeability, anthropomorphism, animacy, or perceived intelligence.

5.1.2 Key Takeaways from Experiment 2 (Uncanny Valley in GenAI Images)

1. The Uncanny Valley Effect Exists in AI Generated Images

   Respondents consistently rated images prompted to fall under the uncanny valley as most strange. Conversely, real human images or images prompted to avoid the uncanny valley reduced perceptions of strangeness. These findings confirm that human perception aligns with the theoretical understanding of this phenomenon.

2. Stable Diffusion Excels at Realism and Uncanniness at Extremes

   Stable Diffusion demonstrated varying levels of accuracy in alignment with ground truth, depending on the complexity of the images. It achieved high accuracy (80-100%) at the extremes, particularly for highly realistic Level 1 images and distinctly uncanny Level 5 images. However, its performance dropped significantly to 20% accuracy for mid-range complexities, such as semi-realistic and hyper-realistic images (Levels 3 and 4), indicating a struggle in rendering these nuanced categories. Notably, the model's ability to generate both highly realistic and distinctly uncanny images suggests an implicit awareness of the factors that make images appear either realistic or unsettling, even though it faces challenges with intermediate levels of complexity.

3. Metrics Correlate with Human Perception

   The Frechet Inception Distance (FID) and Kernel Inception Distance (KID) metrics demonstrate strong consistency with human perception. Lower values for realistic images, such as those at Level 2, indicate a closer resemblance to real images, while higher values for uncanny valley images at Level 5 highlight their greater deviation. These metrics show a strong correlation with human ratings, particularly at the extremes (Level 1 and Level 5), affirming their reliability in assessing image realism and strangeness. Furthermore, FID and KID effectively captured the distinctions between realism and uncanniness, underscoring their utility in quantifying the uncanny valley effect in AI-generated images.

5.2 Discussion

The findings from this study provide insights into the uncanny valley phenomenon in both text-based interactions and AI-generated images, extending the traditional understanding of the uncanny valley into new modalities and applications. This discussion synthesizes the experimental results with the existing body of literature to highlight key implications, address research gaps, and propose directions for future research and design.

This study confirms that the uncanny valley effect extends beyond visual realism into text-based interactions, a domain that has received comparatively little attention in the literature. As evidenced by lower ratings for the GPT4o-Uncanny-Valley Bot across dimensions such as anthropomorphism, likeability, and perceived intelligence, the uncanny valley is not solely a visual or robotic phenomenon. This aligns with the limited prior research on text-based uncanny valley effects (e.g., Ciechanowski et al., 2019 [4]), but expands its scope to modern large language models (LLMs). These findings highlight the necessity of addressing the uncanny valley in conversational agents to enhance user satisfaction and trust.

The finding that 60% of participants mistook human conversations for AI interactions underscores the shifting expectations around human-AI interactions. Anthropomorphism, as discussed by Nass and Moon [55], plays a central role in shaping user perceptions and expectations. When these expectations are unmet—through unnatural syntax, stilted responses, or lack of emotional resonance—users experience cognitive dissonance, contributing to the uncanny valley. This highlights the challenge of designing AI systems that balance anthropomorphic cues with consistency and authenticity.

The GPT4o-Human-Like Bot demonstrated that human-like design principles can effectively mitigate the uncanny valley, with its scores closely approximating human-to-human interactions in likeability and emotional resonance. This aligns with the findings of Betriana et al. and Feine et al., who emphasize that empathetic, consistent conversational agents can bridge the uncanny valley [60] [61]. However, the findings also suggest that even high levels of human-likeness in bots do not entirely eliminate the preference for human-human interactions, pointing to intrinsic limitations in how AI is perceived. The results also suggest that how imperfections affect perception depends on whose imperfections they belong to. Human imperfections are preferred while imperfections in being human trigger the uncanny valley.

The study highlights the critical role of emotional resonance in fostering user satisfaction. While both GPT4o-based bots facilitated longer conversations than human-human interactions, engagement metrics did not always translate into higher likeability, particularly for the uncanny bot. This finding aligns with Gao et al. [57], who identify emotional tone and empathy as central factors in mitigating linguistic uncanniness. It underscores the need for AI systems to prioritize not just conversational length but also the quality of emotional engagement.

Stable Diffusion's performance across realism and uncanniness demonstrates that the uncanny valley manifests strongly in intermediate realism, where subtle artifacts such as asymmetrical features or inconsistent lighting trigger discomfort. This finding aligns with Broad et al. and Igaue and Hayashi, who observed that perceptual mismatches are most pronounced at these intermediate levels [78], [80]. The strong correlation between human ratings and metrics like FID and KID underscores their reliability as tools for assessing realism and uncanniness in image generation.

These findings collectively emphasize the need for a deeper understanding of the uncanny valley across modalities to improve the design and acceptance of AI systems. By addressing the nuanced factors that trigger discomfort—such as perceptual mismatches in visuals or the lack of emotional resonance in text-based interactions—future research and development can bridge the gap between human expectations and AI capabilities.

5.3 Limitations

While these findings offer valuable insights, it is important to acknowledge certain limitations of this study:

Limitations in the Experimental Design

While the experiment was designed to assess participants' emotional perceptions during interactions with chatbots rather than their factual accuracy, several challenges in the experimental design affected the natural flow and authenticity of the interactions.

One key limitation was confusion about chat initiation. Participants were often unsure whether they should start the conversation or wait for the chatbot to respond, leading to moments of hesitation at the beginning of the sessions. While participants were instructed to type if no response was received promptly, this initial uncertainty may have disrupted their engagement.

Another challenge was the slow response times of AI-driven bots, which frustrated participants and interrupted the conversational flow. In some cases, participants left the chat to confirm with the researcher whether the bot was still active, further breaking the natural interaction and potentially influencing their perception of the chatbot's responsiveness.

For participants in the human chatbot condition, maintaining anonymity while delivering seamless chatbot-like responses posed unique difficulties for the researchers. Unexpected participant questions required researchers to gather contextual information from external sources, such as Google or GPT-4, while crafting responses in their own words.

Finally, contextual inaccuracies presented a limitation, particularly for AI-driven bots. Questions about real-time data, such as current weather or time zone details, often resulted in incorrect or outdated responses. While these errors were not critical to the study's focus on emotional engagement and trust, they could have influenced participants' overall impressions.

These limitations highlight areas for improvement in experimental design to ensure more seamless and natural interactions, particularly when studying emotional perceptions in chatbot interactions.

Limited Dataset Size, Representativeness, and Study Duration

One of the primary limitations of this study was the relatively small dataset used in both experiments. In Experiment 1, the number of participants (60) was insufficient to fully generalize the findings on the uncanny valley effect in text-based interactions. A larger and more diverse participant pool would provide stronger statistical power and allow for more robust conclusions. Similarly, in Experiment 2, the number of AI-generated images and participant surveys was limited, making it challenging to definitively establish the reliability of metrics like KID and FID for measuring uncanniness.

Secondly, a more representative sample would also help in generalising the results. Most of the survey respondents for both Experiment 1 and Experiment 2 were recruited at MIT and were grad students in a homogenous age group. Additionally, the study duration for Experiment 1 was restricted to 5-minute interactions, which may not have been enough for participants to fully engage with the chatbots or exhibit natural conversational patterns. A longer study duration could help capture more complex interactions and provide deeper insights into emotional perceptions and engagement.

AI Model and Agent Constraints

The study was constrained by the selection of AI models and agents used in the experiments. In Experiment 1, only GPT4o was employed to simulate chatbot interactions, without exploring other state-of-the-art models such as Gemini, Claude, or Bard. This limited the scope of the findings, as it is unclear whether similar results would be observed with other advanced conversational agents. Additionally, due to resource constraints, only one human researcher played the role of the human chatbot, which could introduce variability and reduce the consistency of responses in the human-chatbot condition. For Experiment 2, Stable Diffusion was the sole image-generation model used, leaving unexplored how other cutting-edge models might perform in generating realistic or uncanny images. Expanding the range of AI models in future studies would allow for a more comprehensive understanding of the uncanny valley effect across technologies and platforms.

Longitudinal Studies Needed

This study focused on a snapshot in time, without accounting for how perceptions of AI might evolve as users become more familiar with these technologies. Longitudinal studies are needed to examine how attitudes toward AI change over extended periods and identify the factors influencing this shift. For instance, as AI becomes more integrated into daily life, perceptions of what constitutes "natural" or "human-like" interaction might change, potentially altering the boundaries of the uncanny valley. Tracking these changes over time would provide valuable insights into the evolving relationship between humans and AI.

Need for Deeper, Nuanced Analysis

While this study identified significant differences in participant perceptions, it did not fully explore the underlying causes of these differences. Future research should dive deeper into what specific aspects of chatbot or image interactions drive perceptions of naturalness or uncanniness. For instance, in text-based interactions, is it the presence of empathy, the occasional mistake, or learned human conditioning that fosters a sense of authenticity? Similarly, in image generation, a more nuanced analysis could reveal which features—such as symmetry, lighting, or texture—are most responsible for triggering discomfort. Such insights would help refine AI designs to better align with human expectations and preferences.

By addressing these limitations, future studies can enhance the robustness of findings, improve the generalizability of results, and provide more actionable insights for designing human-like AI systems.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The thesis began with an in-depth review of the uncanny valley phenomenon, revealing a notable gap in research on its effects in large language models (LLMs) and generative AI systems. To address this gap, two experiments were conducted: the first focusing on text-based conversations with LLMs and the second examining AI-generated images produced by image-generation models. Both experiments demonstrated that the uncanny valley effect extends beyond traditional visual stimuli into both text-based interactions and AI-generated images.

In the first experiment, three chatbots were compared: a GPT4o-based bot intentionally designed to exhibit uncanny characteristics, another GPT4o-based bot engineered to replicate human-like conversation, and an actual human serving as a control. Sixty participants were randomly assigned to interact with one of these three "chatbots." The GPT4o-Uncanny-Valley Bot consistently received lower ratings than its human and more "human-like" GPT4o counterparts, affirming that discomfort arises when users perceive bots as nearly but not convincingly human in text conversation. Furthermore, 60% of participants mistook human-human interactions for AI-driven chats, highlighting how shifting expectations can blur the line between human and machine—often leading to skepticism toward "perfect" or overly structured responses.

In the second experiment, Stable Diffusion was used to generate a series of human images at varying levels of realism. A survey was then distributed. 56 participants responded and rated each image's level of eeriness. Using human evaluation as the ground truth, correlations were drawn between image uncanniness and established metrics such as the Frechet Inception Distance (FID) and Kernel Inception Distance (KID). The results revealed a similar pattern in image-generation contexts. While Stable Diffusion produced images that were convincingly realistic or distinctly uncanny, it faced challenges with mid-level complexities. Moreover, objective measures like FID and KID metrics were closely aligned with subjective human ratings, demonstrating their usefulness in gauging realism and strangeness.

Despite these insights, the study faced limitations, including a relatively small participant pool, short interaction windows, and reliance on specific AI models. Slow bot responses and contextual inaccuracies further affected user perceptions. Nonetheless, the findings offer compelling evidence that uncanniness exists in generative AI. Looking ahead, longitudinal studies involving a broader user base and multiple AI models will be pivotal in examining how the uncanny valley evolves over time, informing more human-centered design strategies for both conversational agents and image-generation systems.

6.2 Future Work

The findings of this study open several avenues for future research, addressing critical gaps in our understanding of the uncanny valley and its implications for AI design, human-AI interaction, and human behavior in digital environments.

First, addressing the research gaps uncovered here will be crucial for advancing current knowledge. While this study corroborates earlier work on text-based uncanny valley effects (e.g., Ciechanowski et al. [4]), it goes a step further by examining advanced large language models (LLMs). Future studies may include a broader, more diverse participant pool and adopt longitudinal designs to explore how the linguistic uncanny valley changes over time and across cultures. Beyond text, the presence of uncanny valley effects in both conversational interactions and AI-generated images highlights the importance of cross-modal perspectives. Investigating modalities such as voice, virtual reality, or multimodal systems could clarify how people perceive and respond to AI's human-like attributes across various contexts. Moreover, this study reveals a rapid blurring of boundaries between human and AI behaviors: participants frequently mistook humans for bots and vice versa, suggesting a need to examine how digital communication norms may be reshaping human conversational patterns.

Second, enhancing design approaches for both conversational agents and image generation remains a key priority. Intentional imperfections—such as pauses, typos, or minor errors—show promise in fostering authenticity and mitigating user discomfort. Future research could systematically vary these "vulnerability factors" to find an optimal balance that reduces suspicion without compromising user experience. At the same time, refining mid-range realism in image generation poses a significant challenge for models like Stable Diffusion, which excel at highly realistic or distinctly uncanny outputs but struggle with subtle, semi-realistic nuances. Further work on specialized training methods and stylization techniques could help prevent unsettling artifacts. Additionally, we could build AI systems that evolve towards real-time detection and adaptation, learning to adjust language style, pacing, or image-rendering parameters by monitoring user feedback and sentiment cues.

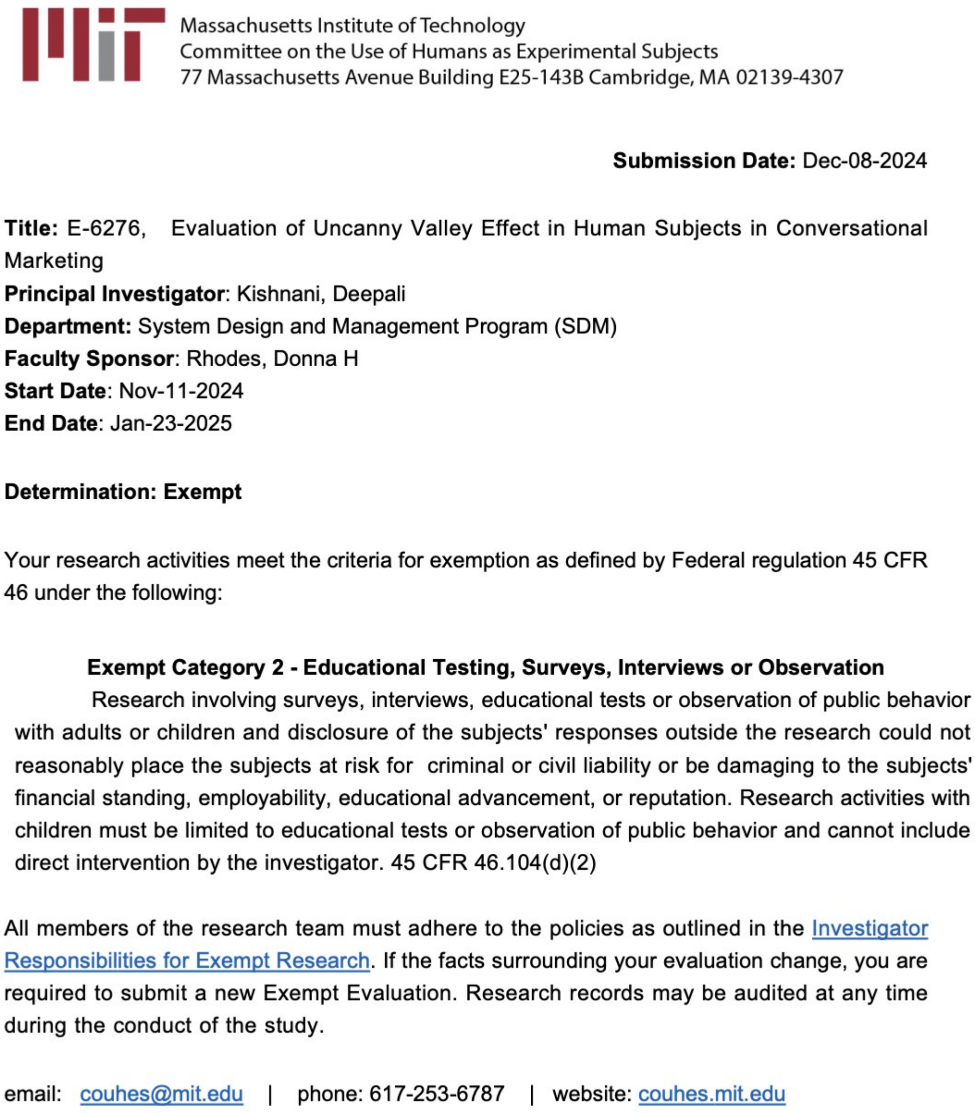
Interdisciplinary collaboration will be vital for translating these insights into practical advances. As the uncanny valley intersects with human psychology, cultural norms, and design principles, research teams comprising psychologists, UX designers, linguists, and computer scientists could more effectively dissect the emotional and cognitive aspects of user perceptions. Ethical and societal considerations also warrant attention: increasingly human-like AI raises concerns about transparency, manipulation, and user autonomy. Collaborations with ethicists and sociologists can help ensure that strategies for circumventing the uncanny valley do not inadvertently erode trust or encourage deceptive practices.

Longitudinal studies and real-world applications will further solidify our understanding of how AI acceptance and trust develop over time. By tracking users' evolving comfort levels and exploring how different interaction styles or design elements influence long-term engagement, researchers can clarify whether repeated exposure reduces or amplifies the uncanny valley effect.

By exploring these directions, researchers and practitioners can refine both conversational and generative AI technologies to be more trustworthy, relatable, and beneficial to users. The ultimate goal is not to create perfect AI but to develop systems that resonate with the imperfect human behaviors and emotional needs—while maintaining transparency and ethical integrity.

Appendix

Appendix A: COUHES Approval



Massachusetts Institute of Technology
Committee on the Use of Humans as Experimental Subjects
77 Massachusetts Avenue Building E25-143B Cambridge, MA 02139-4307

Submission Date: Dec-08-2024

Title: E-6276, Evaluation of Uncanny Valley Effect in Human Subjects in Conversational Marketing
Principal Investigator: Kishnani, Deepali
Department: System Design and Management Program (SDM)
Faculty Sponsor: Rhodes, Donna H
Start Date: Nov-11-2024
End Date: Jan-23-2025

Determination: Exempt

Your research activities meet the criteria for exemption as defined by Federal regulation 45 CFR 46 under the following:

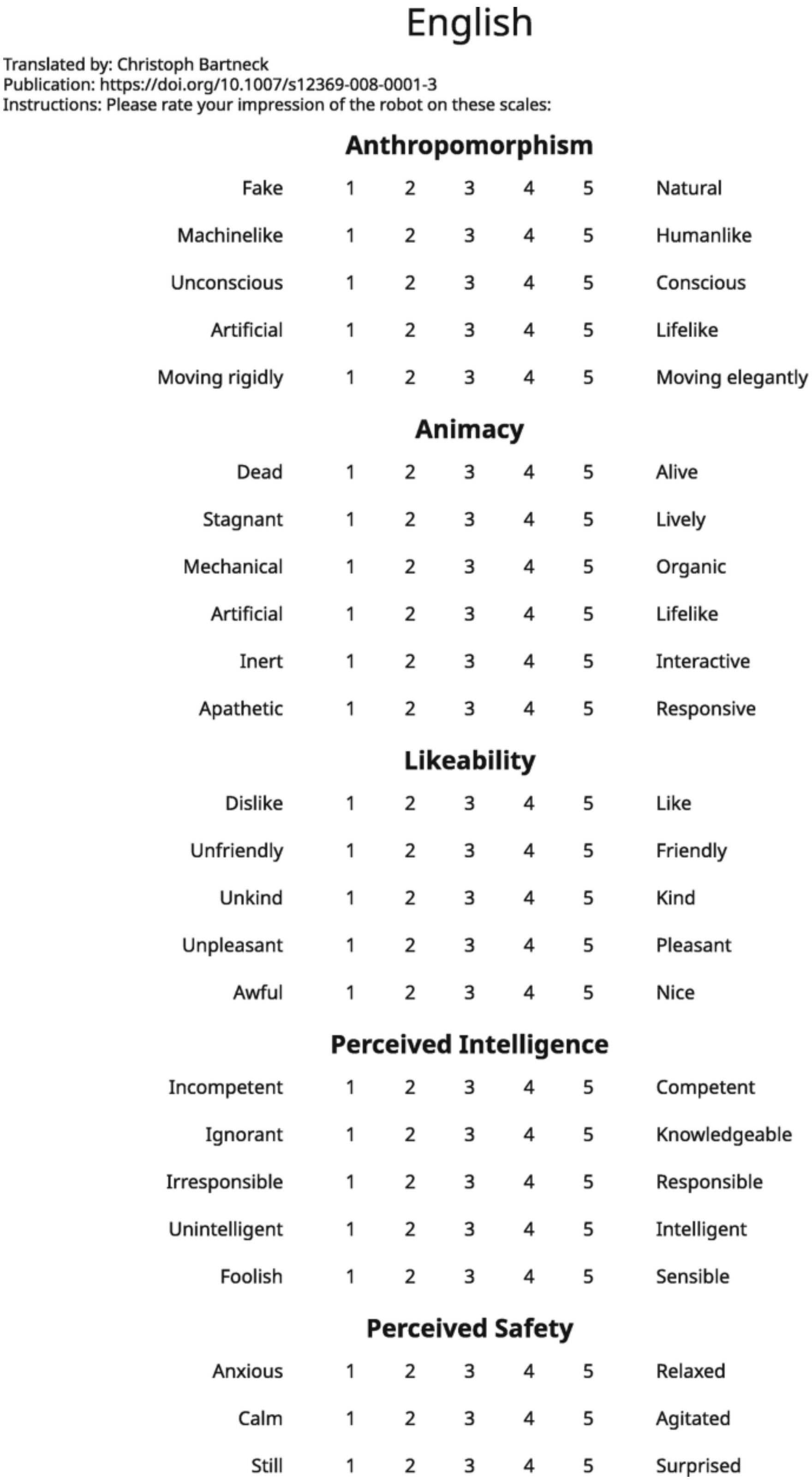
Exempt Category 2 - Educational Testing, Surveys, Interviews or Observation

Research involving surveys, interviews, educational tests or observation of public behavior with adults or children and disclosure of the subjects' responses outside the research could not reasonably place the subjects at risk for criminal or civil liability or be damaging to the subjects' financial standing, employability, educational advancement, or reputation. Research activities with children must be limited to educational tests or observation of public behavior and cannot include direct intervention by the investigator. 45 CFR 46.104(d)(2)

All members of the research team must adhere to the policies as outlined in the Investigator Responsibilities for Exempt Research. If the facts surrounding your evaluation change, you are required to submit a new Exempt Evaluation. Research records may be audited at any time during the conduct of the study.

email: couhes@mit.edu | phone: 617-253-6787 | website: couhes.mit.edu

Appendix B: Godspeed Questionnaire



English

Translated by: Christoph Bartneck
Publication: https://doi.org/10.1007/s12369-008-0001-3
Instructions: Please rate your impression of the robot on these scales:

Anthropomorphism

Fake	1	2	3	4	5	Natural
Machinelike	1	2	3	4	5	Humanlike
Unconscious	1	2	3	4	5	Conscious
Artificial	1	2	3	4	5	Lifelike
Moving rigidly	1	2	3	4	5	Moving elegantly

Animacy

Dead	1	2	3	4	5	Alive
Stagnant	1	2	3	4	5	Lively
Mechanical	1	2	3	4	5	Organic
Artificial	1	2	3	4	5	Lifelike
Inert	1	2	3	4	5	Interactive
Apathetic	1	2	3	4	5	Responsive

Likeability

Dislike	1	2	3	4	5	Like
Unfriendly	1	2	3	4	5	Friendly
Unkind	1	2	3	4	5	Kind
Unpleasant	1	2	3	4	5	Pleasant
Awful	1	2	3	4	5	Nice

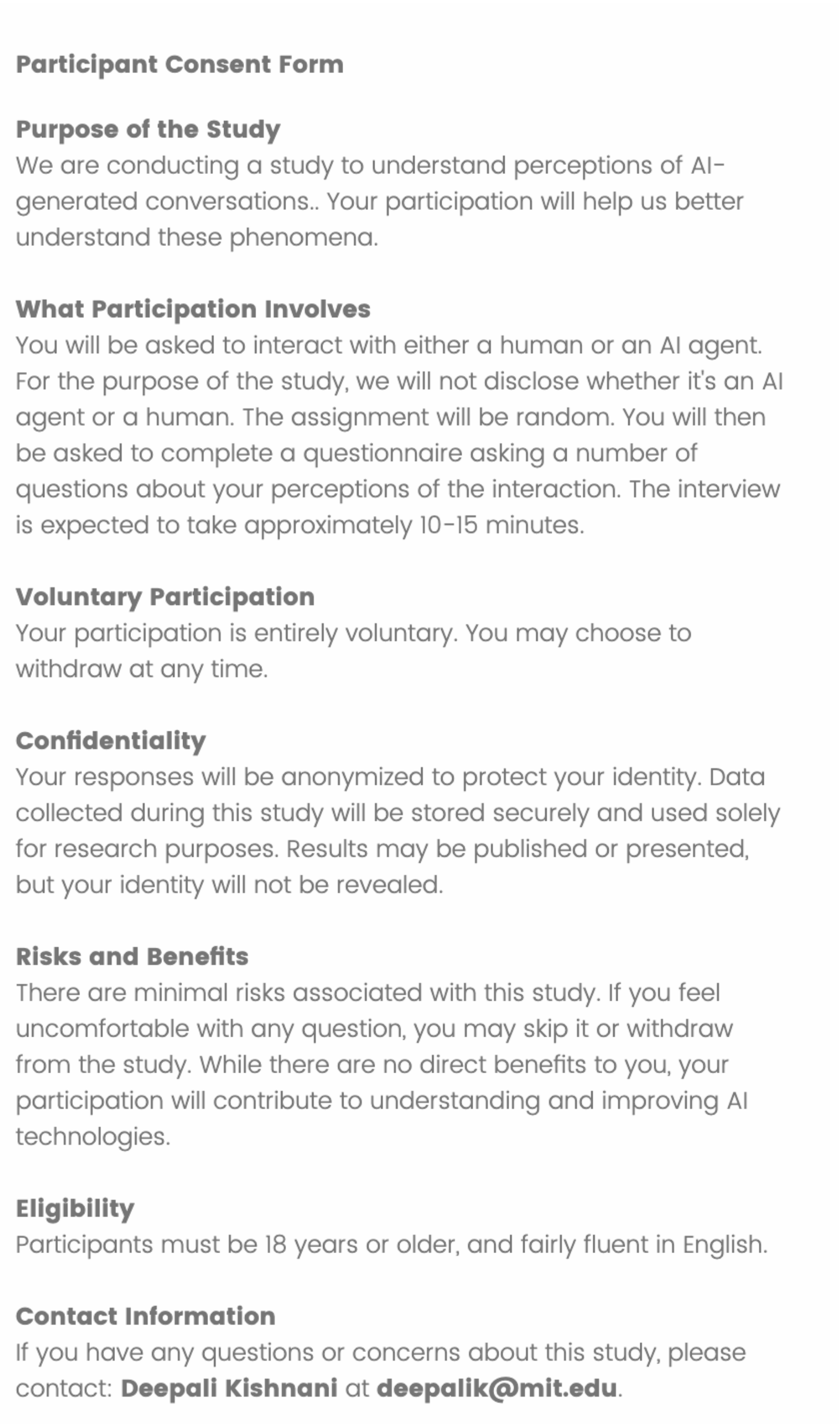
Perceived Intelligence

Incompetent	1	2	3	4	5	Competent
Ignorant	1	2	3	4	5	Knowledgeable
Irresponsible	1	2	3	4	5	Responsible
Unintelligent	1	2	3	4	5	Intelligent
Foolish	1	2	3	4	5	Sensible

Perceived Safety

Anxious	1	2	3	4	5	Relaxed
Calm	1	2	3	4	5	Agitated
Still	1	2	3	4	5	Surprised

Appendix C: Experiment 1 Survey



Participant Consent Form

Purpose of the Study
We are conducting a study to understand perceptions of AI-generated conversations.. Your participation will help us better understand these phenomena.

What Participation Involves
You will be asked to interact with either a human or an AI agent. For the purpose of the study, we will not disclose whether it's an AI agent or a human. The assignment will be random. You will then be asked to complete a questionnaire asking a number of questions about your perceptions of the interaction. The interview is expected to take approximately 10-15 minutes.

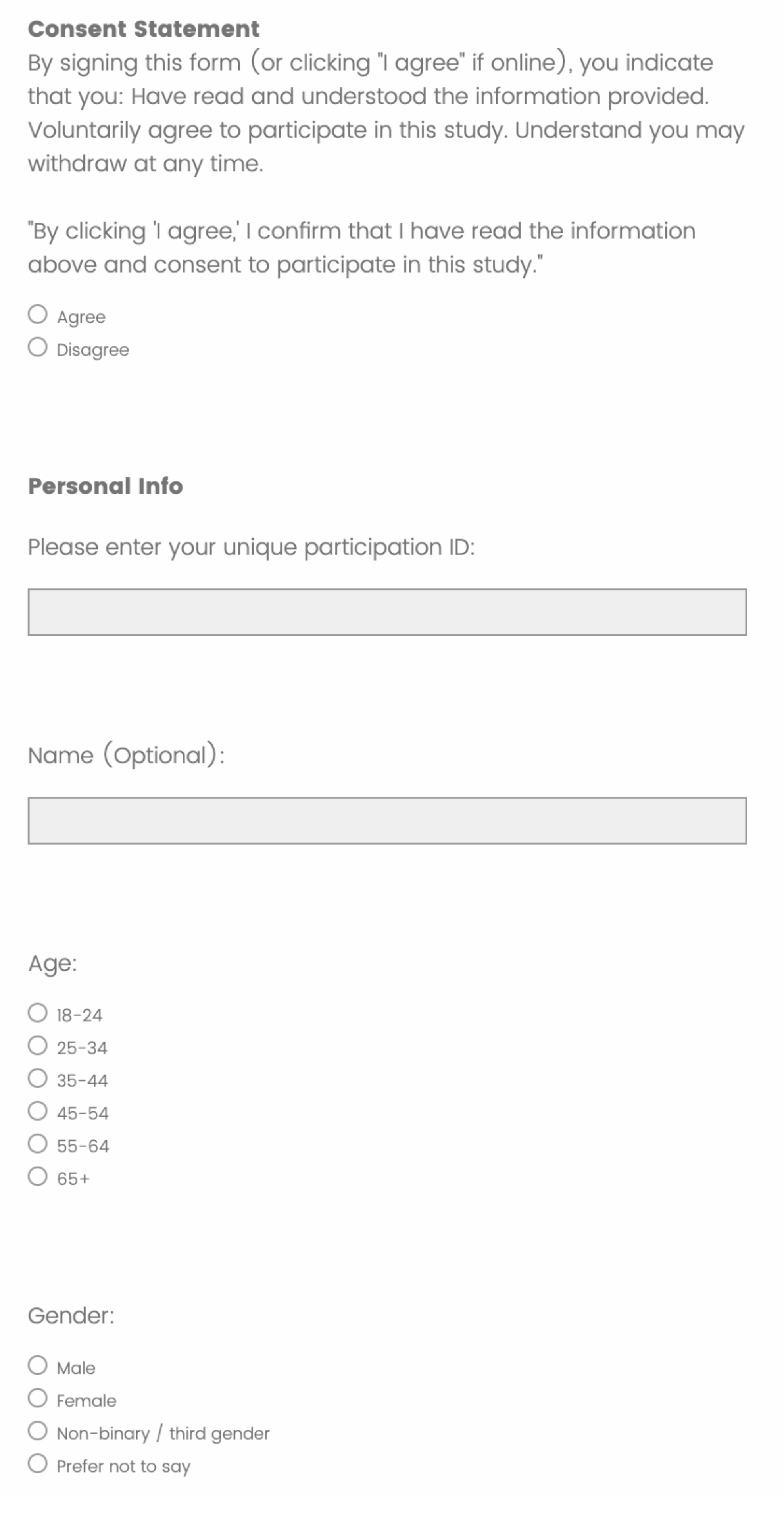
Voluntary Participation
Your participation is entirely voluntary. You may choose to withdraw at any time.

Confidentiality
Your responses will be anonymized to protect your identity. Data collected during this study will be stored securely and used solely for research purposes. Results may be published or presented, but your identity will not be revealed.

Risks and Benefits
There are minimal risks associated with this study. If you feel uncomfortable with any question, you may skip it or withdraw from the study. While there are no direct benefits to you, your participation will contribute to understanding and improving AI technologies.

Eligibility
Participants must be 18 years or older, and fairly fluent in English.

Contact Information
If you have any questions or concerns about this study, please contact: **Deepali Kishnani** at **deepalik@mit.edu**.



Consent Statement

By signing this form (or clicking "I agree" if online), you indicate that you: Have read and understood the information provided. Voluntarily agree to participate in this study. Understand you may withdraw at any time.

"By clicking 'I agree,' I confirm that I have read the information above and consent to participate in this study."

○ Agree
○ Disagree

Personal Info

Please enter your unique participation ID:

Name (Optional):

Age:

○ 18-24
○ 25-34
○ 35-44
○ 45-54
○ 55-64
○ 65+

Gender:

○ Male
○ Female
○ Non-binary / third gender
○ Prefer not to say

Educational Background:

○ High School
○ Associate's Degree or Equivalent
○ Bachelor's Degree
○ Master's Degree
○ Doctorate

On a daily basis, how often do you interact with AI agents such as GPT, Gemini etc.?

○ 0 - 1 hour
○ 1 - 3 hours
○ 3+ hours
○ I don't interact with AI agents at all.

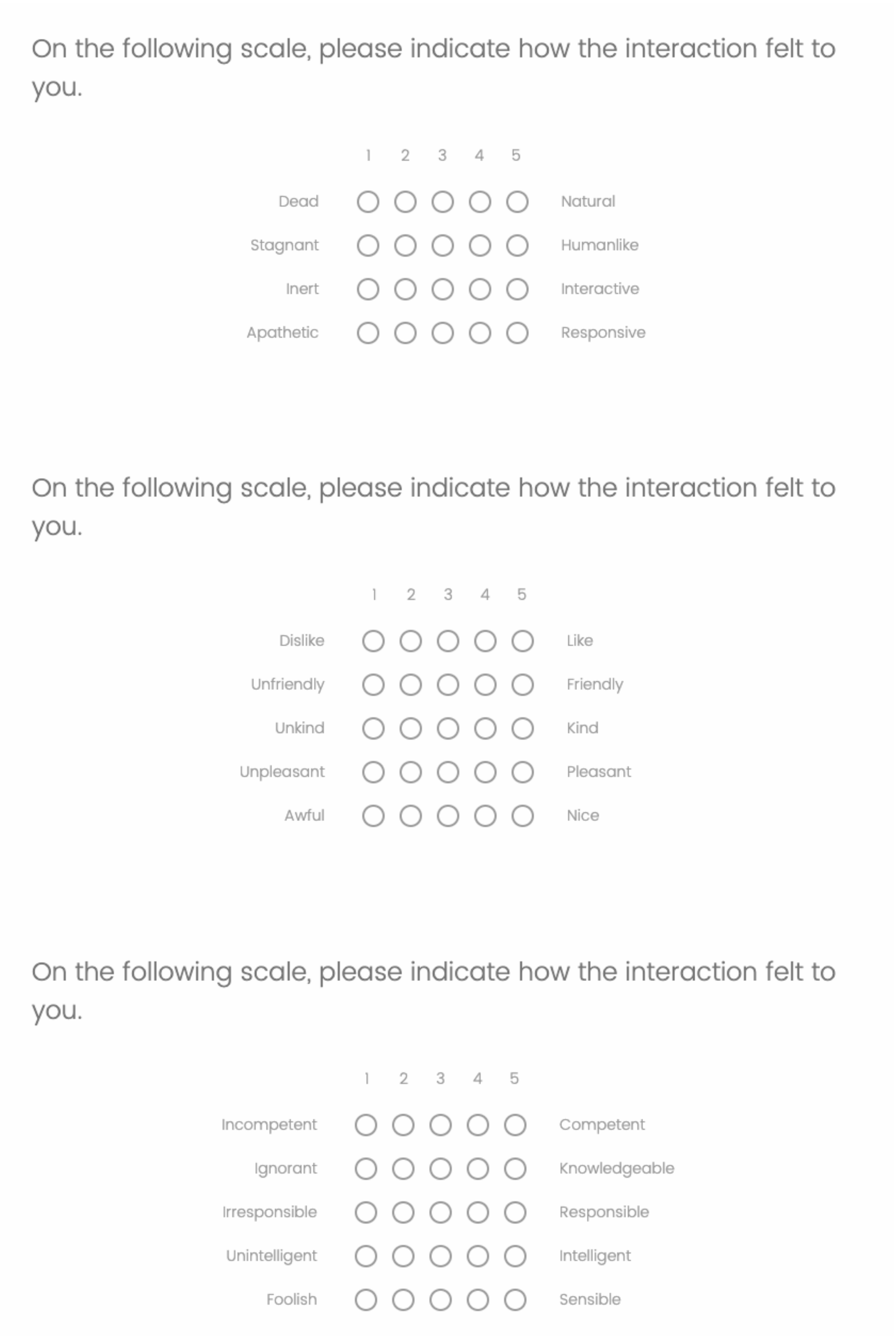
What is your native language?

Where are you currently located?

Default Question Block

On the following scale, please indicate how the interaction felt to you.

	1	2	3	4	5	
Fake	○	○	○	○	○	Natural
Machinelike	○	○	○	○	○	Humanlike
Unconscious	○	○	○	○	○	Conscious
Artificial	○	○	○	○	○	Lifelike
Awkward & Disjointed	○	○	○	○	○	Smooth & Natural



On the following scale, please indicate how the interaction felt to you.

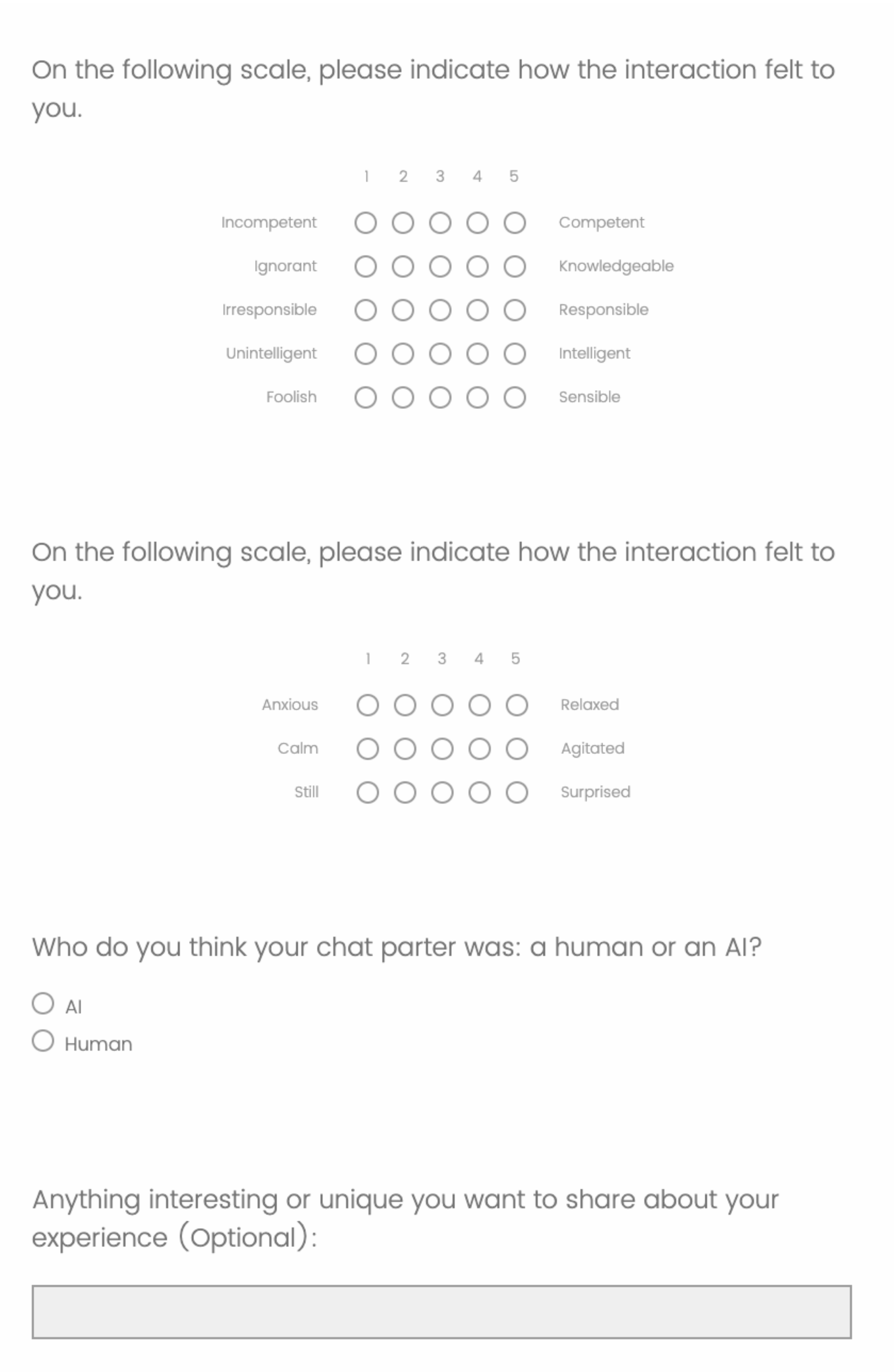
	1	2	3	4	5	
Dead	○	○	○	○	○	Natural
Stagnant	○	○	○	○	○	Humanlike
Inert	○	○	○	○	○	Interactive
Apathetic	○	○	○	○	○	Responsive

On the following scale, please indicate how the interaction felt to you.

	1	2	3	4	5	
Dislike	○	○	○	○	○	Like
Unfriendly	○	○	○	○	○	Friendly
Unkind	○	○	○	○	○	Kind
Unpleasant	○	○	○	○	○	Pleasant
Awful	○	○	○	○	○	Nice

On the following scale, please indicate how the interaction felt to you.

	1	2	3	4	5	
Incompetent	○	○	○	○	○	Competent
Ignorant	○	○	○	○	○	Knowledgeable
Irresponsible	○	○	○	○	○	Responsible
Unintelligent	○	○	○	○	○	Intelligent
Foolish	○	○	○	○	○	Sensible



On the following scale, please indicate how the interaction felt to you.

	1	2	3	4	5	
Incompetent	○	○	○	○	○	Competent
Ignorant	○	○	○	○	○	Knowledgeable
Irresponsible	○	○	○	○	○	Responsible
Unintelligent	○	○	○	○	○	Intelligent
Foolish	○	○	○	○	○	Sensible

On the following scale, please indicate how the interaction felt to you.

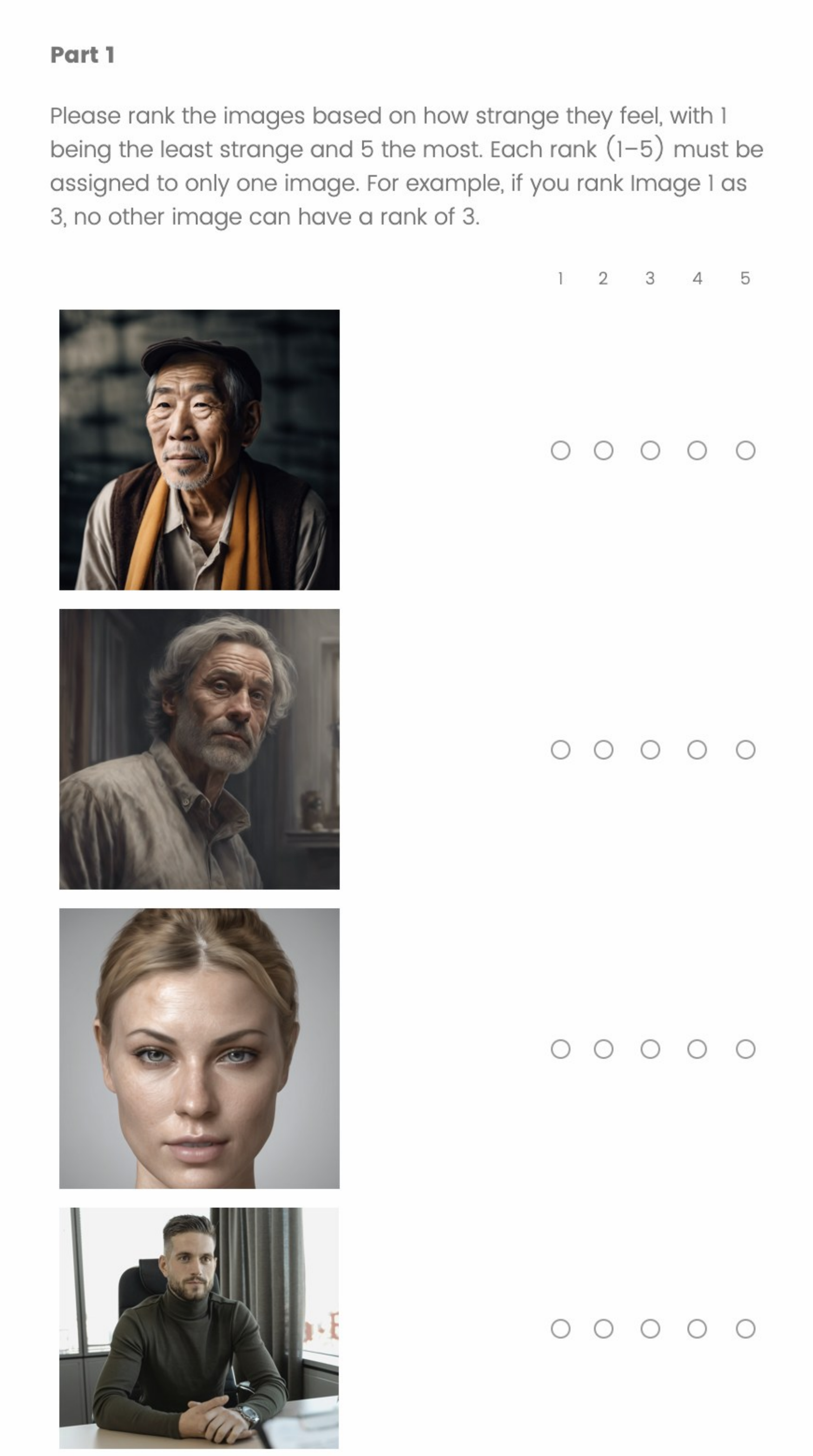
	1	2	3	4	5	
Anxious	○	○	○	○	○	Relaxed
Calm	○	○	○	○	○	Agitated
Still	○	○	○	○	○	Surprised

Who do you think your chat parter was: a human or an AI?

○ AI
○ Human

Anything interesting or unique you want to share about your experience (Optional):

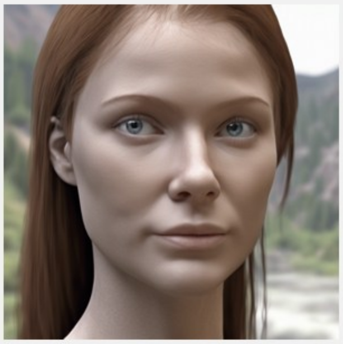
Appendix D: Experiment 2 Survey



Part 1

Please rank the images based on how strange they feel, with 1 being the least strange and 5 the most. Each rank (1–5) must be assigned to only one image. For example, if you rank Image 1 as 3, no other image can have a rank of 3.

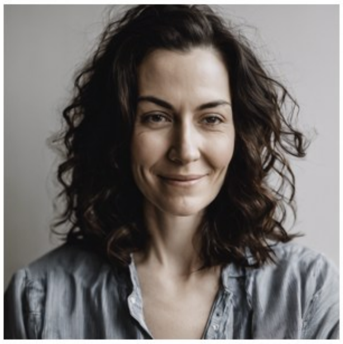
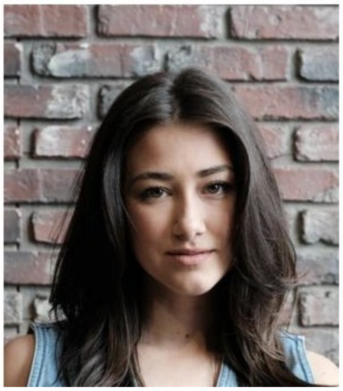
	1	2	3	4	5
	○	○	○	○	○
	○	○	○	○	○
	○	○	○	○	○
	○	○	○	○	○

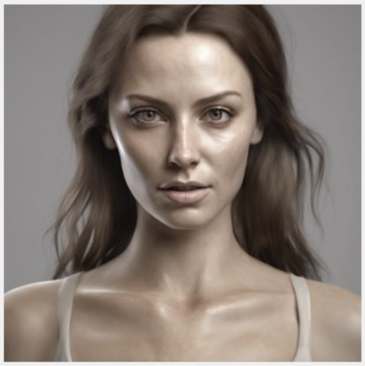
	○	○	○	○	○

Explain in 1-2 sentences: What features or factors influenced your ranking?

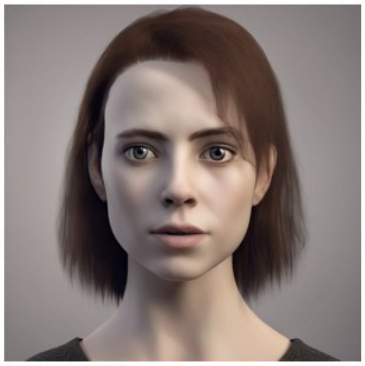
Part 2

Please rank the images based on how strange they feel, with 1 being the least strange and 5 the most. Each rank (1–5) must be assigned to only one image. For example, if you rank Image 1 as 3, no other image can have a rank of 3.

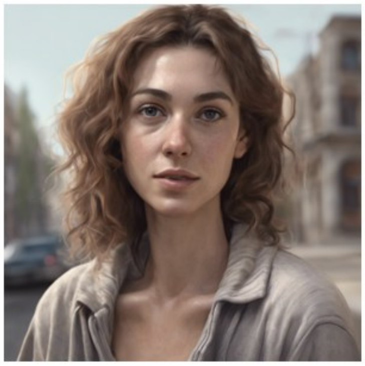
	1	2	3	4	5
	○	○	○	○	○
	○	○	○	○	○



○ ○ ○ ○ ○



○ ○ ○ ○ ○

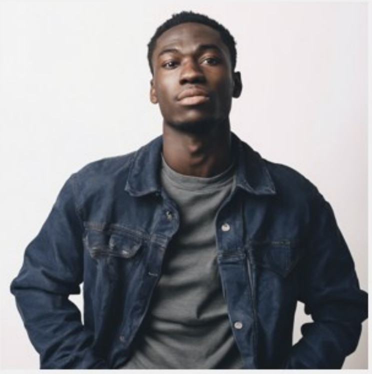


○ ○ ○ ○ ○

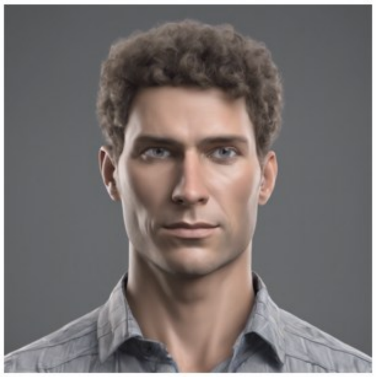
Explain in 1-2 sentences: What features or factors influenced your ranking?

Part 3

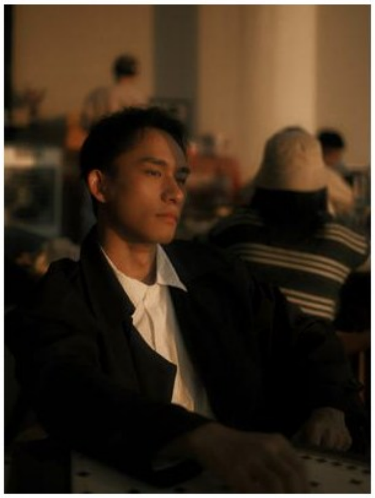
Please rank the images based on how strange they feel, with 1 being the least strange and 5 the most. Each rank (1–5) must be assigned to only one image. For example, if you rank Image 1 as 3, no other image can have a rank of 3.



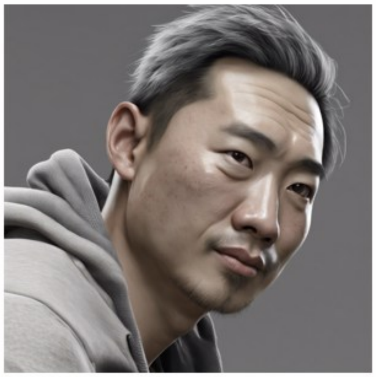
○ ○ ○ ○ ○



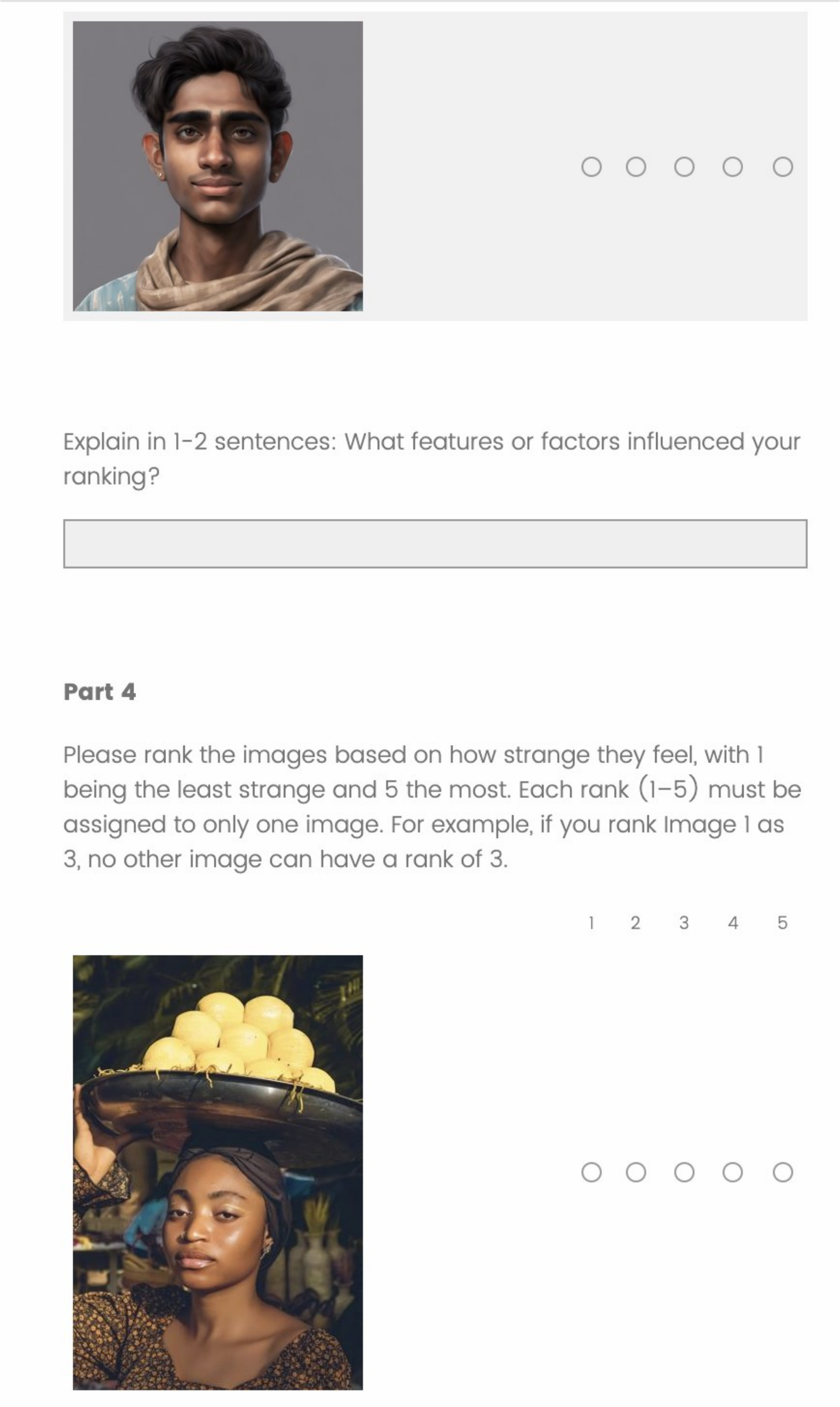
○ ○ ○ ○ ○



○ ○ ○ ○ ○



○ ○ ○ ○ ○



○ ○ ○ ○ ○

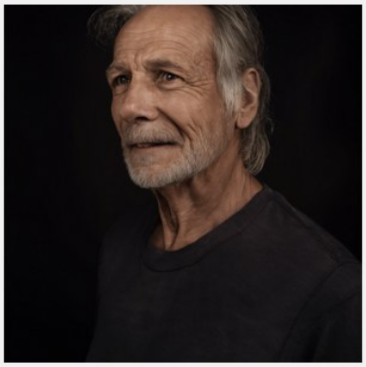
Explain in 1-2 sentences: What features or factors influenced your ranking?

Part 4

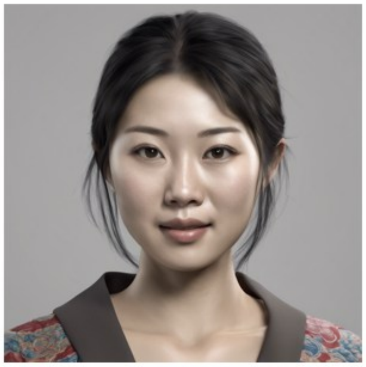
Please rank the images based on how strange they feel, with 1 being the least strange and 5 the most. Each rank (1–5) must be assigned to only one image. For example, if you rank Image 1 as 3, no other image can have a rank of 3.

1 2 3 4 5

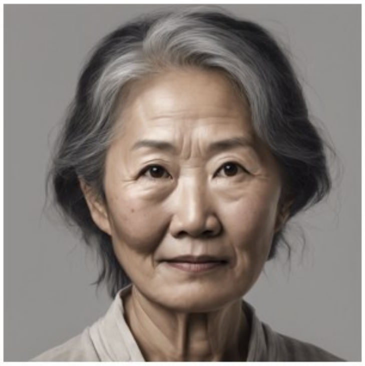
○ ○ ○ ○ ○



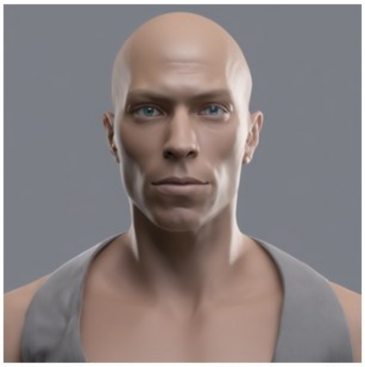
○ ○ ○ ○ ○



○ ○ ○ ○ ○



○ ○ ○ ○ ○

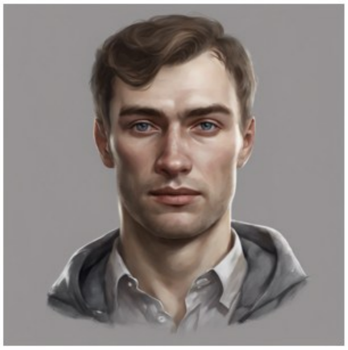
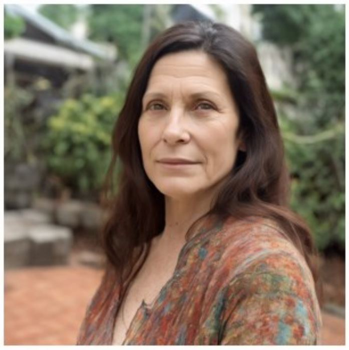
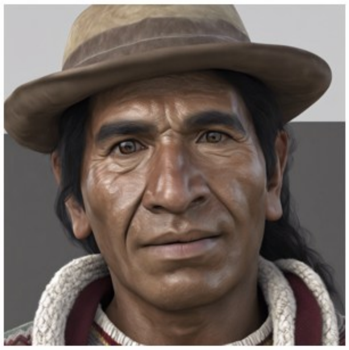
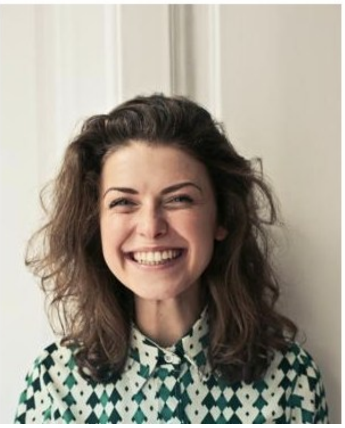


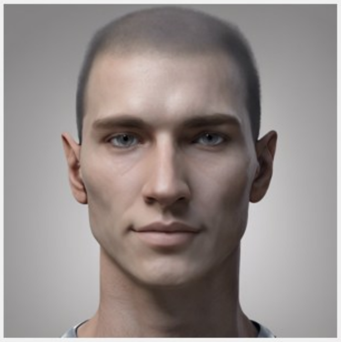
○ ○ ○ ○ ○

Explain in 1-2 sentences: What features or factors influenced your ranking?

Part 5

Please rank the images based on how strange they feel, with 1 being the least strange and 5 the most. Each rank (1–5) must be assigned to only one image. For example, if you rank Image 1 as 3, no other image can have a rank of 3.

	1	2	3	4	5
	○	○	○	○	○
	○	○	○	○	○
	○	○	○	○	○
	○	○	○	○	○



○ ○ ○ ○ ○

Explain in 1-2 sentences: What features or factors influenced your ranking?

Age

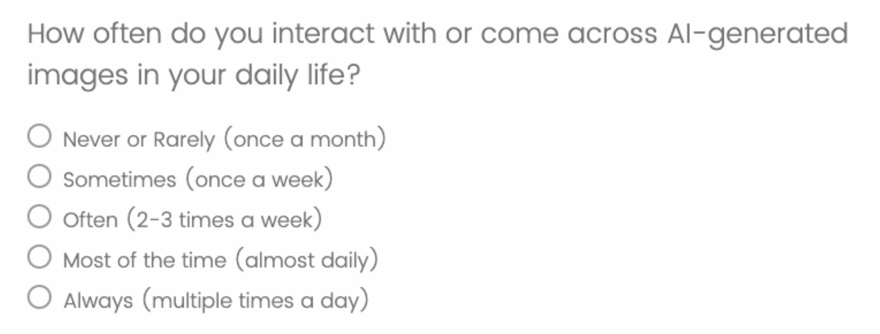
○ 18-24
○ 25-34
○ 35-44
○ 45-54
○ 55-64
○ 65+

Gender

○ Male
○ Female
○ Non-binary / third gender
○ Prefer not to say

Educational Background

○ High School
○ Bachelor's Degree
○ Master's Degree
○ Doctorate



How often do you interact with or come across AI-generated images in your daily life?

○ Never or Rarely (once a month)
○ Sometimes (once a week)
○ Often (2-3 times a week)
○ Most of the time (almost daily)
○ Always (multiple times a day)

Appendix E: Experiment 1 Sentiment Analysis Scores

Participant ID	Sentiment Score	Sentiment Label
2	0.459	Positive
4	0.313	Positive
7	0.215	Positive
10	0.260	Positive
13	0.342	Positive
17	0.367	Positive
18	0.341	Positive
20	0.190	Positive
25	0.133	Positive
27	0.407	Positive
29	0.085	Positive
31	0.123	Positive
33	0.325	Positive
34	0.158	Positive
37	0.191	Positive
40	0.192	Positive
44	0.372	Positive
46	0.223	Positive
45	0.246	Positive
38	0.249	Positive

Table 17: Results of Sentiment Analysis for Human-to-Human Bot

Participant ID	Sentiment Score	Sentiment Label
1	0.129	Positive
5	0.314	Positive
8	0.272	Positive
11	0.380	Positive
14	0.241	Positive
16	0.284	Positive
19	0.309	Positive
21	0.385	Positive
23	0.398	Positive
24	0.187	Positive
26	0.355	Positive
28	0.214	Positive
30	0.273	Positive
32	0.282	Positive
35	0.258	Positive
36	0.063	Positive
39	0.252	Positive
41	0.406	Positive
42	0.121	Positive
43	0.276	Positive

Table 18: Results of Sentiment Analysis for GPT40-Human-Like Bot

Participant ID	Sentiment Score	Sentiment Label
3	0.048	Positive
6	0.112	Positive
9	0.129	Positive
12	0.295	Positive
15	0.222	Positive
47	0.176	Positive
48	0.152	Positive
49	0.196	Positive
50	0.244	Positive
51	0.319	Positive
52	0.064	Positive
53	0.151	Positive
54	0.161	Positive
55	0.184	Positive
56	0.105	Positive
57	0.219	Positive
58	0.162	Positive
59	0.304	Positive
60	0.157	Positive

Table 19: Results of Sentiment Analysis for GPT40-Uncanny-Valley-Bot

References

[1] Mori, M. (1970). The uncanny valley: the original essay by Masahiro Mori. *Ieee Spectrum*, *6*(1), 6.

[2] MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3), 297-337

[3] Tinwell, A., Grimshaw, M., Abdel Nabi, D., & Williams, A. (2011). Facial expression of emotion and perception of the uncanny valley in virtual characters. *Computers in Human Behavior*, 27(2), 741-749

[4] L. Ciechanowski, A. Przegalinska, M. Magnuski, and P. Gloor, "In the shades of the uncanny valley: An experimental study of human–chatbot interaction," *Future Generation Computer Systems*, vol. 92, pp. 539–548, 2019.

[5] MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing the uncanny valley through exposure to autonomous social robots. *Computers in Human Behavior*, 37, 477-488

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

[7] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... & Dean, J. (2022). PaLM: Scaling language modeling with pathways.

[8] Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49, 245-250.

[9] Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., ... & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9), 1248-1258.

[10] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, *1*(2), 3.

[11] Bond-Taylor, S., Hessey, P., Sasaki, H., Breckon, T. P., & Willcocks, C. G. (2022, October). Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *European Conference on Computer Vision* (pp. 170-188). Cham: Springer Nature Switzerland

[12] MacDorman, K. F. (2024). Does mind perception explain the uncanny valley? A meta-regression analysis and (de) humanization experiment. *Computers in Human Behavior: Artificial Humans*, *2*(1), 100065.

[13] Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*, *7*(4), 413-422.

[14] Weizenbaum, J. (1966). ELIZA – A Computer Program for the Study of Natural Language Communication between Man and Machine. *Communications of the ACM*, 9(1), 36-45.

[15] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3), 379-423.

[16] Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2), 257-286.

[17] Jelinek, F. (1976). Continuous Speech Recognition by Statistical Methods. *Proceedings of the IEEE*, 64(4), 532-556

[18] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). (1986) DE Rumelhart, GE Hinton, and RJ Williams, Learning internal representations by error propagation, Parallel Distributed Processing: Explorations in the Microstructures of Cognition, Vol. I, DE Rumelhart and JL McClelland (Eds.) Cambridge, MA: MIT Press, pp. 318-362

[19] Bengio, Y., Simard, P., & Frasconi, P. (1994). *Learning Long-Term Dependencies with Gradient Descent is Difficult. IEEE Transactions on Neural Networks, 5(2), 157-166*

[20] Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735-1780.

[21] Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating Text with Recurrent Neural Networks. Proceedings of the 28th International Conference on Machine Learning (ICML)

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019.

[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI, Tech. Rep., 2019.

[25] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

[26] OpenAI, "GPT-4 technical report," *OpenAI*, 2023. [Online]. Available: https://openai.com/research/gpt-4. [Accessed: Jan. 12, 2025]

[27] Huang, Yinghui & Li, Lie & Dong, Wanghao & Dong, Yuhang & Huang, Yingdan & Liu, Hui. (2024). Human-AI Collaboration Supporting GPT-4o Achieving Human-Level User Feedback in Emotional Support Conversations: Integrative Modeling and Prompt Engineering Approaches (Preprint). 10.2196/preprints.65435.

[28] Asada, M. (2015). Development of artificial empathy. *Neuroscience research*, *90*, 41-50.

[29] Knickrehm, C., & Bauer, K. (2024). GPT, Emotions, and Facts

[30] Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). 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*, *29*(3), 759-771

[31] Nesse, R. M. (2005). Natural selection and the regulation of defenses: A signal detection analysis of the smoke detector principle. *Evolution and human behavior*, *26*(1), 88-105.

[32] Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*, *7*(4), 413-422.

[33] Tinwell, A., Grimshaw, M., Nabi, D. A., & Williams, A. (2011). Facial expression of emotion and perception of the Uncanny Valley in virtual characters. *Computers in Human behavior*, *27*(2), 741-749

[34] Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). 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*, *6*, 390

[35] Seyama, J. I., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence*, *16*(4), 337-351

[36] Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). 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*, *6*, 390.

[37] Wang, S., Lilienfeld, S. O., & Rochat, P. (2015). The uncanny valley: Existence and explanations. *Review of General Psychology*, *19*(4), 393-407

[38] Stein, J. P., & Ohler, P. (2017). Venturing into the uncanny valley of mind—The influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, *160*, 43-50

[39] Karras, T., Aila, T., Laine, S., Herva, A., & Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (ToG)*, *36*(4), 1-12.

[40] Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). 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*, *6*, 390

[41] Zhang, Y., Colman, B., Guo, X., Shahriyari, A., & Bharaj, G. (2025). Common sense reasoning for deepfake detection. In *European Conference on Computer Vision* (pp. 399-415). Springer, Cham

[42] Saygin, A. P., Chaminade, T., & Ishiguro, H. (2010). The perception of humans and robots: Uncanny hills in parietal cortex. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32)

[43] Cheetham, M., Wu, L., Pauli, P., & Jancke, L. (2015). Arousal, valence, and the uncanny valley: Psychophysiological and self-report findings. *Frontiers in psychology*, *6*, 981

[44] Zhang, J., Li, S., Zhang, J. Y., Du, F., Qi, Y., & Liu, X. (2020). A literature review of the research on the uncanny valley. In *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments: 12th International Conference, CCD 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22* (pp. 255-268). Springer International Publishing.

[45] Castelo, N., & Sarvary, M. (2022). Cross-cultural differences in comfort with humanlike robots. *International Journal of Social Robotics*, *14*(8), 1865-1873.

[46] Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). 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*, *6*, 390

[47] Bartneck, C., Kanda, T., Ishiguro, H., & Hagita, N. (2009, September). My robotic doppelgänger-A critical look at the uncanny valley. In *RO-MAN 2009-The 18th IEEE international symposium on robot and human interactive communication* (pp. 269-276). IEEE.

[48] MacDorman, K. F., & Chattopadhyay, D. (2017). Categorization-based stranger avoidance does not explain the uncanny valley effect. *Cognition*, *161*, 132-135.

[49] Janowski, K., Ritschel, H., & André, E. (2022). Adaptive artificial personalities. In *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application* (pp. 155-194)

[50] Haring, K. S., Mougenot, C., Ono, F., & Watanabe, K. (2014). Cultural differences in perception and attitude towards robots. *International Journal of Affective Engineering*, *13*(3), 149-157

[51] Beese, N. O. (2024). *The Virtual Experience-Examining Visual, Auditory and Haptic Capabilities and Aspects of Spatial Cognition and User Experience in Virutal Reality* (Doctoral dissertation, Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau)

[52] Roth, D., Stauffert, J. P., & Latoschik, M. E. (2019). Avatar Embodiment, Behavior Replication, and Kinematics in Virtual Reality. *VR Developer Gems*, *1*, 321-348.

[53] K. Dautenhahn, "Socially intelligent robots: Dimensions of human–robot interaction," *Philos. Trans. R. Soc. B: Biol. Sci.*, vol. 362, no. 1480, pp. 679–704, 2007.

[54] M. Coeckelbergh, *Human Being @ Risk: Enhancement, Technology, and the Evaluation of Vulnerability Transformations*. Dordrecht, The Netherlands: Springer, 2013.

[55] Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, *56*(1), 81-103.

[56] Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors*, *63*(7), 1196-1229.

[57] Gao, J., Xu, Z., Liang, Z., & Liao, H. (2019). Expected consistency-based emergency decision making with incomplete probabilistic linguistic preference relations. *Knowledge-Based Systems*, *176*, 15-28.

[58] Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: a survey. *International Journal of Social Robotics*, *5*, 291-308.

[59] Zhang, L., Li, W., Bai, Q., & Lai, E. (2021). Graph-based Self-Adaptive Conversational Agent. *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, 1791–179

[60] J. Feine, U. Gnewuch, S. Morana, and A. Maedche, “A taxonomy of social cues for conversational agents,” *Int. J. Hum.-Comput. Stud.*, vol. 132, pp. 138–161, 2019, doi: 10.1016/j.ijhcs.2019.07.009.

[61] F. Betriana, K. Osaka, K. Matsumoto, T. Tanioka, and R. C. Locsin, "Relating Mori’s Uncanny Valley in generating conversations with artificial affective communication and natural language processing," *Nursing Philosophy*, vol. 22, no. 2, pp. 1-8, 2021, doi: 10.1111/nup.12322.

[62] K. Gray and D. M. Wegner, “Feeling robots and human zombies: Mind perception and the uncanny valley,” *Cognition*, vol. 125, no. 1, pp. 125–130, 2012, doi: 10.1016/j.cognition.2012.06.007.

[63] A. Abubshait and E. Wiese, “You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction,” *Front. Psychol.*, vol. 8, p. 1393, 2017, doi: 10.3389/fpsyg.2017.01393.

[64] Sutherland, I. E. (1964, January). Sketch pad a man-machine graphical communication system. In *Proceedings of the SHARE design automation workshop* (pp. 6-329).

[65] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, *313*(5786), 504-507.

[66] Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103).

[67] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

[68] Radford, A. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

[69] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14* (pp. 694-711). Springer International Publishing.

[70] Prabakaran, N., Bhattacharyay, R., Joshi, A. D., & Rajasekaran, P. (2023). Generating Complex Animated Characters of Various Art Styles With Optimal Beauty Scores Using Deep Generative Adversarial Networks. In *Handbook of Research on Deep Learning Techniques for Cloud-Based Industrial IoT* (pp. 236-254). IGI Global.

[71] Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, *34*, 8780-8794.

[72] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021, July). Zero-shot text-to-image generation. In *International conference on machine learning* (pp. 8821-8831). Pmlr.

[73] Karras, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv preprint arXiv:1812.04948.

[74] Valentin Schwind, Katrin Wolf, and Niels Henze. 2018. Avoiding the uncanny valley in virtual character design. interactions 25, 5 (September-October 2018), 45–49.

[75] Tinwell, A., Grimshaw, M., & Williams, A. (2010). Uncanny behaviour in survival horror games. *Journal of Gaming & Virtual Worlds*, *2*(1), 3-25.

[76] Seyama, J. I., & Nagayama, R. S. (2007). The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence*, *16*(4), 337-351.

[77] Schwind, V., & Jäger, S. (2016). The uncanny valley and the importance of eye contact. *i-com*, *15*(1), 93-104.

[78] Broad, T., Leymarie, F. F., & Grierson, M. (2020). Amplifying the uncanny. *arXiv preprint arXiv:2002.06890*.

[79] Wu, H., Chen, Z., Huang, Y., & Tu, H. (2024). Research on the uncanny valley effect in artificial intelligence news anchors. *Multimedia Tools and Applications*, 1-26

[80] Igaue, T., & Hayashi, R. (2023). Signatures of the uncanny valley effect in an artificial neural network. *Computers in Human Behavior*, *146*, 107811

[81] L. Lee, "Measures of Distributional Similarity," in 37th Annual Meeting of the ACL, 1999, pp. 25–32.

[82] A. Ghildyal and F. Liu, "Attacking Perceptual Similarity Metrics," *arXiv preprint arXiv:2305.08840*, May 2023.

[83] MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). "Too Real for Comfort? Uncanny Responses to Computer Generated Faces." *Computers in Human Behavior, 25(3)*, 695-710

[84] Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, *1*, 71-81

[85] Loria, S. (2018). textblob Documentation. *Release 0.15*, *2*(8), 269.

[86] Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, *13*.

[87] Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning* (Vol. 242, No. 1, pp. 29-48)

[89] Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications

[90] Weisman, W. D., & Peña, J. F. (2021). Face the uncanny: the effects of doppelganger talking head avatars on affect-based trust toward artificial intelligence technology are mediated by uncanny valley perceptions. *Cyberpsychology, behavior, and social networking*, *24*(3), 182-187

[91] Yu, Y., Zhang, W., & Deng, Y. (2021). Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School*, *3*.