DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics: Games Engineering

Virtual Pet - The Integration of a Naturally Acting A.I.-powered V-Companion into the Dynamic Environment of an Online Streaming Platform

Markus Gumbart

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics: Games Engineering

Virtual Pet - The Integration of a Naturally Acting A.I.-powered V-Companion into the Dynamic Environment of an Online Streaming Platform

Virtual Pet - Die Integration eines natürlich wirkenden K.I. gestützten V-Begleiters in das dynamische Umfeld einer Online Streaming Platform

Author:	Markus Gumbart
Supervisor:	Professor Gudrun Klinker
Advisor:	Dr. Sandro Weber
Submission Date:	15.05.2024

I confirm that this master's thesis in informatics: games engineering is my own work and I have documented all sources and material used.

Garching, 15.05.2024 Markus Gumbart

Acknowledgments

Hereby, I would like to thank everyone who contributed to the creation of this thesis. First and foremost, I want to extend my appreciation to my advisor, Dr. Sandro Weber. Despite his busy schedule, we managed to have many productive meetings during the last semester. These meetings were extremely helpful, and without them, I doubt I would have made the same kind of progress.

I am also very thankful to my family for their patience and support. Especially my father and mother for listening to my endless discussions about the project. This helped me clear my thoughts and gain a better understanding of the challenges I faced.

I must also give thanks to my friends from university, particularly Jonas and Nina. Throughout the semester, they helped me with all different kinds of situations. This ranged from addressing specific coding questions and discussions about thesis writing to simply relaxing once in a while.

In this section, I don't want to fail to mention my viewers and friends in the live-stream community who encouraged me in starting this project in the first place. Their involvement in play-testing *Petricia*, identifying bugs, and other issues was incredibly helpful, with special thanks to Patrick for his keen insights. Additionally, I want to thank Carla for creatively naming the pet *Petricia*.

Lastly, I want to acknoledge the immense help that different AIs have been throughout this project. Especially, the coding AI *GitHub Copilot* and *ChatGPT*. They assisted on a broad range of topics, from implementing a websocket to formulating these acknowledgments[1].

To all mentioned and unmentioned, your support and contributions were invaluable. Thank you.

[1]GPT only helped in formulating certain sentences and did not write any portion of these acknowledgments alone. They are personal and therefore, written by myself.

Abstract

The project aims to realize an innovative approach to enhancing viewer interaction in live-streams by adding a naturally embedded companion and assistant of the streamer. For this, the virtual pet *Petricia* is developed in the course of this project. *Petricia* is a combination between standard virtual pet, live-streaming viewer-entertainment feature and artificial intelligence (AI), incorporating several different directions of modern digital entertainment. The design of the pet is centered around the general ideas and goals of augmented reality (AR). One such goal is the enhancement of the user's perception of and interaction with the real world by superimposing the virtual upon the real world as naturally as possible and in real time. To achieve naturalness in *Petricia*, the system makes use of the capabilities of modern AI by outsourcing the decision making process of reacting to chat messages to *ChatGPT*. *ChatGPT* is a deep-learning based large language model by *OpenAI*, even capable of representing diverse characters and personalities as shown throughout the thesis. The personality of *Petricia* is defined through the *Big Five* personality model, a personality-trait-theory summarizing the human personality as the combination of five major traits and the extends of their respective representation in an individual. *Petricia* is not designed to represent only one specific personality, however, but any given set of trait values chosen by the streamer. The extend to which this is achieved in this project is tested and documented within the thesis.

Kurzfassung

Das Projekt zielt darauf ab, eine innovative Herangehensweise zur Verbesserung der Interaktion mit Zuschauern in Live-Streams zu realisieren, indem ein natürlich eingebetteter Begleiter und Assistent des Streamers hinzugefügt wird. Dafür wird im Rahmen dieses Projekts das virtuelle Haustier *Petricia* entwickelt. *Petricia* ist eine Kombination aus einem Standard-*Virtual-Pet*, einem Unterhaltungsfeature für Live-Stream-Zuschauer und künstlicher Intelligenz (KI). Damit vereint sie mehrere verschiedene Richtungen des modernen digitalen Entertainments. Das Design des Haustiers basiert auf den allgemeinen Ideen und Zielen der *augmented reality* (AR). Ein solches Ziel ist die Verbesserung der Wahrnehmung von, und Interaktion mit der realen Welt, indem die virtuelle Welt so natürlich wie möglich und in Echtzeit mit der realen Welt kombiniert wird. Um Natürlichkeit bei *Petricia* zu erreichen, nutzt das System die Fähigkeiten moderner KI, indem es den Entscheidungsprozess für Reaktionen auf Chat-Nachrichten an *ChatGPT* auslagert. *ChatGPT* ist ein auf *deep learning* basierendes *large language model* von *OpenAI*, das in der Lage ist, verschiedene Charaktere und Persönlichkeiten darzustellen, wie in dieser Arbeit gezeigt wird. Die Persönlichkeit von Petricia wird durch das *Big-Five*-Persönlichkeitsmodell definiert, das die menschliche Persönlichkeit als Kombination von fünf Hauptmerkmalen und deren Ausmaß in einer Person definiert. Petricia ist jedoch nicht darauf ausgelegt, nur eine spezifische Persönlichkeit zu repräsentieren, sondern jede beliebige Kombination von Merkmalswerten, die von dem Streamer gewählt wird. Inwieweit dies in diesem Projekt gelungen ist, wird in der Arbeit getestet und dokumentiert.

Contents

1. Introduction

In the current era, digitalization is rapidly reshaping every facet of human life, from the way we communicate to how we work, learn, and entertain ourselves. The impact of this can even be compared to the industrial revolution [1, 2]. The affected areas are so diverse, it is no wonder the changes even reached "man's best friend", the dog. With titles like *Nintendogs* or *Dogz*, these beloved animals have now entered the virtual world [3]. But not just dogs are being digitalized. The humans' need for a pet is large enough that the whole area needed to be digitalized [3].

Pets lower stress buffering and provide social support in humans [4, 3]. They are also said to give a person emotional relationship, including love, trust, loyalty, and joyful mutual activity [4, 3]. With their many other advantages and long history with humans, they are a significant part of human culture, justifying their influence on modern day technology [5].

A specific implementation of these so called "virtual pets", named *Petricia*, is developed in the course of this project. *Petricia* is a combination between standard virtual pet, live-streaming viewer-entertainment feature and artificial intelligence (AI), incorporating several different directions of modern digital entertainment. The project aims to realize an innovative approach to enhancing viewer interaction in live-streams by adding a naturally embedded companion and assistant of the streamer. In order to achieve this, the design of *Petricia* is centered around the general ideas and goals of augmented reality (AR). "AR enhances the user's perception of and interaction with the real world", which also includes any indirect view of the real world, such as through a live-stream [6]. To do so, AR technology should augment the sense of reality by superimposing the virtual upon the real world as naturally as possible and in real time [6]. To achieve naturalness in *Petricia*, the system makes use of the capabilities of modern AI by outsourcing the decision making process of reacting to chat messages to *ChatGPT* (may be referred to as GPT or chatGPT from now on). ChatGPT is a deep-learning based large language model (LLM) by *OpenAI*, even capable of representing diverse characters and personalities as shown throughout the thesis. The personality of *Petricia* is defined through the *Big Five* personality model, a personality-trait-theory summarizing the human personality as a combination of five major traits and the extents of their respective representation in an individual (see chapter 2.1). *Petricia* is not designed to represent only one specific personality however, but any given set of trait values chosen by the streamer. The extent to which this is achieved in this project is tested and documented in chapter 4.

1.1. Virtual Pets (and where to find them)

A virtual pet, often referred to as a digital pet, is an artificial human companion that exists in a virtual environment and interacts with users via computer hardware or mobile devices [3, 7, 5, 8, 9]. These digital entities are programmed to display various behaviors and responses analogous to those of a living pet, such as eating, sleeping, playing, and learning [3, 8, 5]. Virtual pets can range from simple, static images accompanied by text to complex, three-dimensional beings that interact in real time with their environment and their caretakers. The latter includes the three-dimensional space of the non-virtual world with physical entertainment robots[1] that are built to look and act like pets [8].

The concept of a virtual pet can be traced back to the early days of computing, but it gained significant popularity in the mid-1990s with the introduction of the video game *Petz*[2] (1995) and the *Tamagotchi* by *Bandai* (1996) [3, 7, 10]. Though many attribute the start of the virtual pet era solely to *Tamagotchi*, *Petz* had its own popularity and was released one year prior to the *Tamagotchi* [3, 7, 9]. It included cats and dogs as "autonomous characters with real-time layered 3D animation and sound, for which it was necessary to care for" [3]. *Petz* was released for PC and MAC, while the *Tamagotchi* came to the market as a standalone hand-held toy with its own tiny display screen. The idea for this is said to have come from a Japanese mother who imagined the toy for her children, as limited space of the household made it impossible to get an actual pet [7]. The small, egg-shaped devices featured an LCD screen and a few buttons, allowing users to care for a digital creature by feeding it, cleaning after it, and playing with it [3, 7]. It caused an immense boom across the world and is said to have been sold about 40 million times in the first years after its release [11, 7].

Following the *Tamagotchi* craze, in 1997, *PF Magic* introduced their second game in the *Petz* series. It introduced the possibility of having multiple pets at once which therefore now could interact with each other. The producers also promote the game with their pets' new individual personalities, where "No two are alike" [9]. *Petz 3* (1998) then also introduced breeding where certain visual and behavioral traits could be passed on by the parents to the offspring, including random mutations [10, 9]. This added a level of complexity and personalization that had not been seen before in virtual pets.

During the late 1990s and early 2000s, the virtual pet phenomenon continued to expand with *Neopets*, an online website where users could care for virtual pets living on another planet [3]. *Neopets* put a larger focus on the social aspect of virtual pet care by allowing interactions not just with the pet but also with other users, thereby enhancing the communal experience. The game featured private communications, public forums, internal communication, as well as the ability to buy and sell goods, play the stock market, participate in quests, competitions and multiplayer games [3].

With advancements in technology, the capabilities and complexity of virtual pets have significantly evolved. Modern virtual pets, such as those found in smartphone apps like *My Talking Tom* or augmented reality games like *Pokémon GO* offer a more interactive and engaging experience [3]. These apps utilize the hardware capabilities of modern devices to create realistic and responsive environments where virtual pets can live and grow. For instance, *Pokémon GO* uses GPS technology to integrate virtual creatures into the real world, blending physical and digital experiences in a manner that was unimaginable during the early days of *Tamagotchi* [12].

In the future, AI could also play a critical role in the advancement of virtual pets. It could enable these digital entities to learn from interactions with the user, adapt to their preferences, or impersonate more unique and realistic personalities. This level of dynamic interaction adds a depth to the relationship between the virtual pet and the user, mirroring the emotional

[1] In this case, the virtual environment is within the robot.

[2] Or more specifically, the two separate games *Catz* and *Dogz*, which could be bought together as *Petz* [9, 10].

connection found with real pets.

A current state-of-the-art example for this is *Peridot*, an augmented reality (AR) mobile game by *Niantic* [13, 14]. The *Peridot*-pet (*Dot*) is seen through the user's smartphone camera and, using AI, it is able to understand its real-world surroundings. More precisely, the camera captures the real world and converts it into accurate 3D models, allowing *Dots* to freely navigate within the captured environment. The pet is even able to recognize certain objects, like flowers, food and other pets, and react accordingly [14]. How to react is decided by another type of AI, a large language model (LLM). The LLM receives the environmental data in addition to the personality and characteristics of this individual pet in form of text and returns a reaction it deems fitting. The characteristics include for example the age and play history, while the personality contains behavioral attributes, such as whether the pet is outgoing or shy [14].

2. Related Work

The following sections describe different areas that strongly relate to this paper and deliver relevant information for a deeper understanding of the topic. Within these sections, either specific works or general concepts are presented with corresponding references. The topic of virtual pets does not require its own section in this chapter as the topic and specific implementations are already discussed in section 1.1.

2.1. The Five-Factor Model of Personality (FFM)

The *Five-Factor Model*, or FFM, is the most well known and established model for describing human personalities to date [15, 16, 17]. It is based on the so called lexical approach to personality structure [18, 19] where the assumption is made that because personality traits are so central to human interactions, all important traits will have been encoded in natural language. The model identifies personality dimensions similar to the well known *Big Five* and has also been replicated across many cultures [20]. Though often used synonymously with the *Big Five*, the FFM was derived from factor analysis of questionnaires rather than adjectives [20]. In this paper, both names may be used when referring to the traits.

The *Five-Factor Theory* (FFT) is a trait theory, meaning that it describes the personality as a set of fundamentally different traits that each person represents to a certain degree. In an early stage of developing the final model, Costa and McCrae, the developers of the model, defined these traits as "dimensions of individual differences in tendencies to show consistent patterns of thoughts, feelings, and actions" [21]. They later revoke the idea of patterns in behavior being directly based on personality traits and add another factor, *Characteristic Adaptations*, which is described further below [16]. Their five traits are ***Openness, Conscientiousness, Extraversion, Agreeableness*** **and** ***Neuroticism***. Costa and McCrae defined these traits to be basic tendencies that influence the behavior only through characteristic adaptations. Their representation of the five factor theory personality system is replicated in figure 2.1.

The figure (2.1) can be interpreted as a diagram of how personality operates at any given time. The components were inferred by McCrae and Costa by documenting recurring categories of variables in personality throughout many different personality theories [22]. At the time, these included basic tendencies, characteristic adaptations, self concept, objective biography and external influences. The external influences constitute the situation or context, while the objective biography represents a specific instance of behavior as output of the system [16].

The basic tendencies represent the core personality traits that are mostly unchanging [23]. They are inherited and only develop through childhood or through life altering events (disease, psychological intervention) [22, 16]. The traits making up the basic tendencies are not to be comprehended as patterns of behavior and are not directly observable. Instead, they can only be inferred from behavior and experience and only materialize through characteristic adaptations [16]. The five traits are defined as higher order factors with each being an aggregate

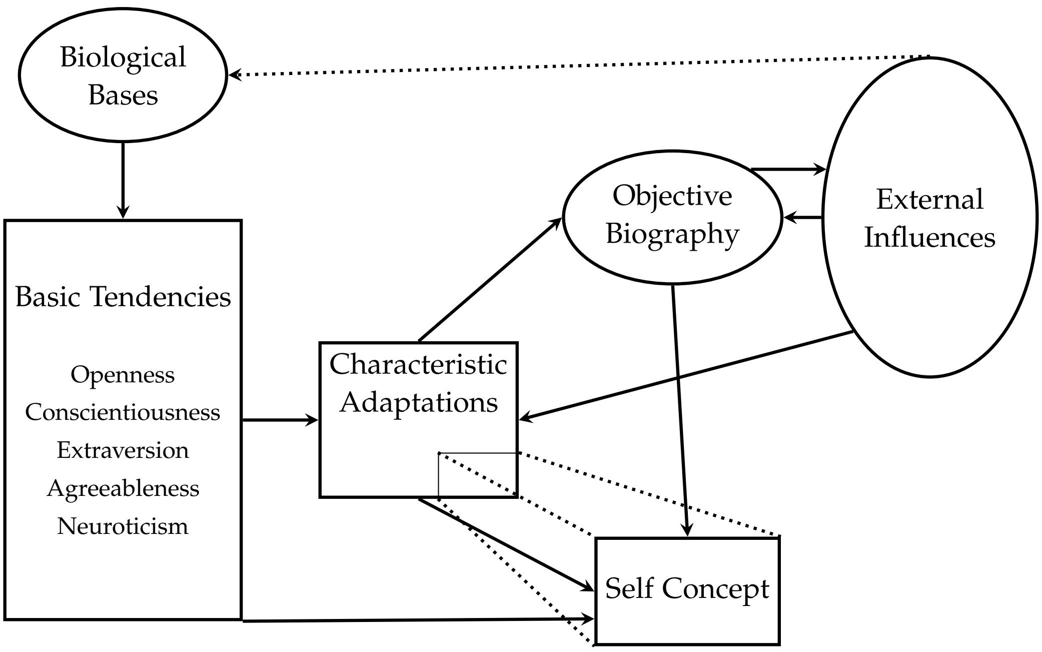


Figure 2.1.: The five-factor theory personality system by Costa and McCrae [22]. Core components are in rectangles while interfacing ones are in ellipses. Arrows symbolize dynamic processes.

of six lower order facets [24, 25].

In contrast to basic tendencies, characteristic adaptations are more dynamic. They change over time in response to biological maturation, changes in the environment, or deliberate interventions [22, 16][1]. The category includes habits, attitudes, skills, roles, and relationships (among others), which are all influenced by basic tendencies and external influences. Another major distinction between characteristic adaptations and basic tendencies is that the former varies tremendously across cultures, families, and even portions of the lifespan, while the latter does not . The interrelations between these two components are one of the main aspects of FFT. At it's core, however, it aims to define the different traits of the basic tendencies and their effect [22, 16]. Each trait is described in the following in more detail. [2]

2.1.1. Extraversion (E)

The term *Extraversion*[3] was first introduced by C. G. Jung [26]. He described more extraverted individuals as being more focused on the outer world, in contrast to more introverted individuals who were thought to be focused more on their own inner mentality [20]. It later became one of the two first well known traits of personality, discovered by H.J. Eysenck through the analysis of questionnaires [27]. Together with neuroticism, it was part of the *Big Two* before

[1]Anytime in the thesis a reference is followed by sentences regarding the same topic but without reference, they refer to the same sources until either the topic changes or a new source is referenced.

[2]Note: As a non-psychologist, I tried to focus on the definitions of the traits rather than their history or validation. Therefore, I only shortly introduce the traits' background and mostly formulate a compilation of defining features, descriptions and adjectives.

[3]Outside of their individual sections, the traits may be referred to as either the name in lower case or their initial (N,E,O,A,C).

being expanded to the *Big Five* [28, 27, 24, 25, 29]. In the FFM, *Extraversion* is made up of six lower order facets, which are [30, 21]:

- Warmth
- Gregariousness
- Assertiveness
- Activity
- Excitement Seeking
- Positive Emotions

Extraversion reflects tendencies to experience and exhibit positive affect, assertive behavior, decisive thinking, and desires for social attention [24]. Extraverted individuals are characterized by energy, dominance, spontaneity, and sociability, whereas introverted individuals tend to be described as more lethargic, inhibited, reflective, and quiet [20]. People with high *Extraversion* also tend to be more active, assertive and energetic than others. They can also be described as enthusiastic, outgoing and talkative [27, 31]. To further elaborate on the meaning of the trait and it's range, Costa and McCrae described the following opposing adjective definers in their work from 1992 [21]:

- Reserved – Affectionate
- Quiet – Talkative
- Sober – Fun-loving
- Loner – Joiner
- Passive – Active
- Unfeeling – Passionate

2.1.2. Neuroticism (N)

The term *Neuroticism* has been widely utilized in personality psychology, although its usage has not always been consistent [32, 33]. Initially linked to psychiatric diagnoses of neuroses, its meaning has evolved over time [32]. It is reasonable to assume that individuals suffering from anxiety disorders, minor depressions, and some other forms of psychopathology would score high on measures of *Neuroticism*. It would, however, be a mistake to equate *Neuroticism* with psychopathology, since many psychiatric disorders involve defects in cognition, social bonding, and reality orientation that are not elements of *Neuroticism* [32].

Following the example of Eysenck, Costa and McCrae adopted the term as one of their two first fundamental personality traits [33, 34]. At the time, Norman had already defined a similar factor through the opposing end of the spectrum [35]. His Emotional Stability was found to be the polar opposite of Costa and McCrae's *Neuroticism*, reinforcing the claim that N is a core personality trait [35, 36]. Costa and McCrae define their *Neuroticism* as "a broad dimension of individual differences in the tendency to experience negative, distressing emotions and to possess associated behavioral and cognitive traits" [32]. They describe defining traits of this dimension to be fearfulness, irritability, low self-esteem, social anxiety, poor inhibition of impulses, and helplessness among others [32]. This gave the basis for the six lower order facets, *Neuroticism* is made up of in the FFM [30, 21]:

- Anxiety
- Angry Hostility
- Self-consciousness
- Impulsiveness

- Depression
- Vulnerability

Individuals with high *Neuroticism* are said to be prone to experience fear, anger, sadness, and embarrassment. They are unable to control cravings and urges, and feel unable to cope with stress [32]. Furthermore, these individuals can be described as touchy, unstable and worrying. *Neuroticism* also heightens the tendency to be anxious, self-pitying or tense [27, 31]. To further elaborate on the meaning of the trait and it's range, Costa and McCrae described the following opposing adjective definers in their work from 1992 [21]:

- Calm – Worrying
- Self-satisfied – Self-pitying
- Unemotional – Emotional
- Even Tempered – Temperamental
- Comfortable – Self-conscious
- Hardy – Vulnerable

2.1.3. Openness to Experience (O)

Openness was added to the set of main personality traits by Costa and McCrae in 1980 [37]. Before that, only extraversion and neuroticism were known as major components of psychological tests [27]. Over time it had several names like *Openness to Intellect, Imagination, or Culture*, though it is mainly known as *Openness to Experience* [30]. But even before the name *Openness* was used, a similar trait was already concluded by Norman in 1963 called *Culture* [35]. The proof that Culture was a variant of Costa and McCrae's O was achieved by empirical testing in 1985, establishing the trait as one of the *Big Five* [34, 36]. In the FFM, *Openness* is made up of these six lower order facets [30, 21]:

- Fantasy
- Aesthetics
- Feelings
- Actions
- Ideas
- Values

The trait is associated with creativity, imagination, a need for variety, novelty and change [38, 16]. Individuals with high *Openness* are documented to be more artistic and curious than others. They tend to have a wider range of interests and have more original and insightful ideas [27, 31].To further elaborate on the meaning of the trait and it's range, Costa and McCrae described the following opposing adjective definers in their work from 1992 [21]:

- Down to Earth – Imaginative
- Conventional – Original
- Uncurious – Curious
- Uncreative – Creative
- Prefer Routine – Prefer Variety
- Conservative – Liberal

2.1.4. Agreeableness (A)

Agreeableness is said to be the trait with the shortest history [39]. This could be attributed to the fact that it got labeled differently over time, although the same idea already existed [40]. Long before scientific psychology appeared, Aristotle already commented on the role of *Agreeableness*, not as a personality trait, but as a moral virtue. In his *Nicomachean Ethics* he described it as a characteristic that could be cultivated and used in the service of group living and civic participation [40]. The way *Agreeableness* finally appeared in modern psychology was quite unusual, however, compared to the *Big Two*, extraversion and neuroticism. While N and E were tied to distinctive psychological processes in brain activity, *Agreeableness* research was based on observable regularities in descriptions of others and self-descriptions [41, 40, 42]. In 1985 Costa and McCrae accepted the importance of A as a higher order personality trait after empirical testing and added it to their personality model, at the time only consisting of N, E and O [34, 36]. They found A to primarily be a dimension of interpersonal behavior, just like extraversion. But where they related E to the preferred quantity of social interaction, they defined A to represent the characteristic quality of interaction along a continuum from compassion to antagonism [43]. Next to its role as a defining axes in the interpersonal circumplex, A also influences the self-image and helps to shape social attitudes and philosophy of life [44, 43]. The six lower order facets McCrae and Costa found *Agreeableness* to be made up of in their FFM are [30, 21, 43]:

- Trust
- Straightforwardness
- Altruism
- Compliance
- Modesty
- Tender Mindedness

After becoming part of the FFM, Graziano and Eisenberg tried to define the trait in motivational terms. They suggested that *Agreeableness* was a summery label for individual differences in the motivation to maintain positive relations with others [45, 40]. People with high *Agreeableness* tend to be more appreciative, forgiving and generous than others. They can also be described as kind, sympathetic and trusting [27, 31]. To further elaborate on the meaning of the trait and it's range, Costa and McCrae described the following opposing adjective definers in their work from 1992 [21]:

- Ruthless – Softhearted
- Stingy – Generous
- Critical – Lenient
- Suspicious – Trusting
- Antagonistic – Acquiescent
- Irritable – Good Natured

2.1.5. Conscientiousness (C)

Conscientiousness was introduced to the FFM together with agreeableness [29]. Both were unrepresented in the *NEO-Inventory* until its revision in 1990, forming the *NEO-PI-R* [43]. The term itself, however, has already been studied long before that. In 1929, Hartshorn, May and Maller addressed *Conscientiousness* as an aspect of character, already drawing close to the idea of personality traits [46, 43]. Later, Murray and Kluckhohn described it as an aspect of ego strength, using such terms as will power, initiative, and responsibility [47, 43].

Before Costa and McCrae then finalized the concept of C in their personality model, they favored another quite similar (in meaning) term as a name for the trait: *Direction*. The name *Direction* implied both movement and focus, two fitting descriptors for their concept of the domain. They decided on *Conscientiousness* instead and conceptualized it as having both proactive and inhibitive aspects. The proactive side of *Conscientiousness* is seen most clearly in the need for achievement and commitment to work; the inhibitive side is seen in moral scrupulousness and cautiousness [43]. The six lower order aspects Costa and McCrae described for *Conscientiousness* are [30, 21, 43]:

- Competence
- Order
- Dutifulness
- Achievement Striving
- Self-discipline
- Deliberation

Individuals with high *Conscientiousness* tend to be efficient, organized and planful. They can also be described as reliable, responsible and thorough [27, 31]. To further elaborate on the meaning of the trait and it's range, Costa and McCrae described the following opposing adjective definers in their work from 1992 [21]:

- Negligent – Conscientous
- Disorganized – Well-organized
- Aimless – Ambitious
- Lazy – Hardworking
- Late – Punctual
- Quitting – Persevering

2.2. LLM Personality Assessment

Personality Assessment on LLMs is a large part of this thesis and is mainly addressed in chapter 4, *Personality Assessment*. To gain a deeper understanding of the subject, some of the most relevant papers regarding the topic are described in the following. The three papers explored below all apply personality tests to LLMs with varying goals, LLMs, and personality tests. They all relate to this thesis in different regards but also exhibit certain problems, this project aims to avoid. These are elaborated on further in the respective section.

2.2.1. Revisiting the Reliability of Psychological Scales on Large Language Models [48]

In their work from 2023, Huang et al. determine the reliability of applying personality assessments to LLMs. In doing so, they investigate whether LLMs demonstrate consistent personality traits. The work includes the analysis of responses from *gpt-3.5-turbo* to the *Big Five Inventory* (BFI, a personality test) under 2,500 different settings. These settings include different introductory instructions, ways of formulating the items/questions, languages of the questions, labellings of the answer options and orders of answer options (ascending/descending). The results of the tests are then reduced to two dimensions (from five) via factor analysis and show a high degree of reliability. In addition to these findings, they explore the potential of *gpt-3.5-turbo* to emulate diverse personalities and represent various groups based off of broad descriptions. Through testing, they conclude that LLMs have the potential to represent different personalities with specific prompt instructions.

What separates the personality representation tests by Huang et al. from the ones carried out in the course of this thesis (chapter 4) is the way the personality prompt is constructed. Huang et al. present GPT with broad descriptions of characters or personalities where the specific details of the personality traits are left to be decided on by GPT itself. This thesis on the other hand proposes the approach of prompting GPT with exact personality trait values and testing to what degree GPT is able to reflect exactly those values.

2.2.2. Personality traits in large language models [49]

Another relevant paper released in 2023 is *Personality traits in large language models* by Safdari et al.. In the paper, they present their method for administering and validating personality tests on LLMs. This method also includes their attempt at assigning certain personalities to the model. For all of their studies LLMs from the PaLM family are used because of their established performance on generative tasks. For testing the personalities, Safdari et al. use the IPIP ([50]) and the BFI. In doing so, they find that personality measurements in the outputs of some LLMs under specific prompting configurations are reliable and valid. They also describe how evidence of reliability and validity of synthetic LLM personality is stronger for larger and instruction fine-tuned models. Their third finding is that personality in LLM outputs can be shaped along desired dimensions to mimic specific human personality profiles. For this, the authors propose a prompting methodology that shapes each synthetic personality trait at nine intensity levels, using Likert-type linguistic qualifiers and 104 trait adjectives, expanding upon Goldbergs personality trait markers [51, 52]. A similar approach for the qualifiers is used in this thesis as described in section 3.3.1.

What needs to be mentioned about their personality assignment procedure, however, is that it is virtually a way of telling the LLM how to answer certain questions instead of a personality description. In the prompt to the AI they only describe one trait at a time and they do so by presenting the model with several trait specific adjectives and the extend to which they should be represented. The authors state the following example prompt for a personality representing slightly below average extraversion with "Level 4/9":

> For the following task, respond in a way that matches this description: {Persona Description}. I'm {a bit introverted, a bit unenergetic, a bit silent, a bit timid, a bit inactive, a bit unassertive, and a bit unadventurous}.

What they do not seem to take into account by doing so, is that most personality questionnaires are based on the same adjectives [50, 52, 53]. Especially since they mostly use the IPIP which has the same author as the adjective list they use for prompting the model, Lewis R. Goldberg [50, 52]. This way, there are questions like "are you talkative?" with possible answers ranging from strongly disagree to strongly agree (first question of BFI) and prompts like "I'm [...] very talkative". It seems obvious that the LLM did not in fact represent a personality but rather answered specific questions the way it was told to. Further strengthening this claim, are the results of the personality traits that are not regarded in the current prompt, so for example in the E prompt: N, O, A and C. These results barely change, if they change at all, although certain traits should influence each other. High extraversion for example should affect neuroticism values negatively and the other way around since they are in many aspects opposing traits (see section 2.1) [24, 32, 29].

2.2.3. Romero et al.: Do GPT Language Models Suffer From Split Personality Disorder? The Advent Of Substrate-Free Psychometrics [54]

In their study, Romero et al. explore the psychological profile of *GPT-3*, particularly examining its expression of personality traits across different languages. For this, they questioned the AI with the same personality questionnaire in nine different languages. The results reveal that LLMs manifest language-specific personality distributions which exhibit notable inconsistencies and what can be interpreted as sub-personalities. Their findings also highlight more general interlingual and intralingual instabilities. For instance, responses in English and German were relatively consistent, whereas Asian languages produced computational delays and inconsistencies, often defaulting to English responses. The Russian language yielded no results at all as the model was not able to deliver responses in the correct format for the language. These findings suggest that LLMs lack a stable core personality, leading to unpredictable and potentially unsafe behaviors when integrated into various applications according to the authors.

Their results and the therefore following conclusions seem to occur mostly because of the use of GPT *version-3* as other research show exactly the opposite with newer models (see 2.2.1) [48]. On the other hand, some aspects of their testing methods show certain oddities compared to the methods of others as well. For testing the personality they use the *Ten Item Personality Inventory* (TIPI) which consists only of ten items; two per *Big Five* factor, of which one is reversed [55]. Romero et al. state that the used personality questionnaire must be short enough to draw qualitative conclusions without adding additional complexity of sub-scales. Although presented this way, it could have serious negative impacts on the results, using a questionnaire this short. If for example one answer slightly fluctuates, the resulting personality trait values show large differences. Large enough even to justify the assumption of a split personality. These results are only negative, however, when trying to prove the consistency of the AI's personality or the reliability of the test. They are positive on the other hand when trying to prove sub-personalities and unsafe instabilities.

3. The Virtual AI-Powered Stream Pet, Petricia

The goal in designing *Petricia* was to create additional entertainment for live-streams while trying to keep the distraction from the stream's own content as low as possible. She is visually designed to look like a small cat- and dog-like creature with large ears, as seen in figure 3.1. The pet should function as an assistant for the streamer with a large focus on its natural integration into the stream's flow. To achieve the most natural behavior, *Petricia*'s ractions are chosen by a text-based artificial intelligence, namely *ChatGPT*. This way the pet is able to use *Natural Language Processing* (NLP) and can therefore understand the natural speech, different languages and even slang used in the stream's chat.

Figure 3.1.: The stream-pet, *Petricia*, designed to look dog- and cat-like but also original.

3.1. System Architecture

The system is modeled around the behavior of *Petricia* since the entertainment from interacting with her is the main focus of this project. Everything else in the system has nearly the sole purpose to influence this behavior. There are different types of factors of influence: the static internal, dynamic internal, and dynamic external influences. The arrangement of those and their functionality are closely oriented on the personality system of the five factor theory (FFT) by McCrae and Costa [16] seen in figure (2.1). The resulting architecture can be seen in figure (3.2). Just like in the original model, oval nodes represent interfacing components. And closely related to the dynamic processes in the FFT system [16], the arrows in the graph (3.2) indicate direct influence of one aspect on another.

The static internal influences resemble the *Basic Tendencies* (BT) in FFT, which include the five personality traits. They are, according to FFT, not influenced by the surrounding, making them independent variables [40].

The *Characteristic Adaptations* (CA) on the other hand are rather dynamic internal influences on behavior that are shaped by the personality and external influences [40]. In the case of *Petricia* this includes all procedures used to decide upon behavior, like the calculations for unprompted behavior (3.3.2). It also contains an adaptation of the original model's “attitude”, the current mood. The mood (or happiness) is an attribute of *Petricia* that is influenced by the personality and her dynamic surroundings (external influences) and in turn has

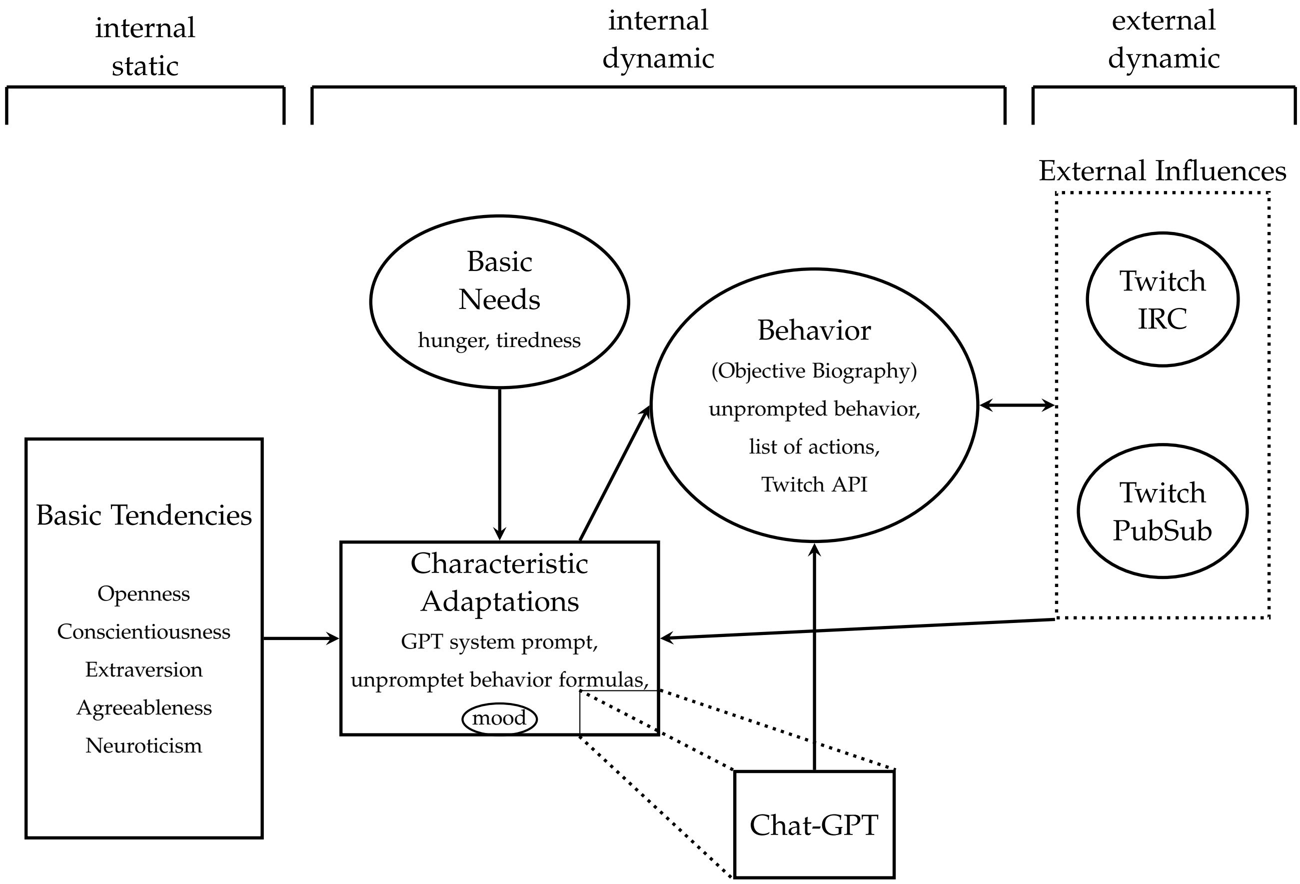


Figure 3.2.: System Architecture of the virtual stream-pet *Petricia* based off of the Five Factor Theory personality system of Costa and McCrae [16]. Oval nodes represent interfacing components. The system is centered around the behavior of the pet with all other factors influencing it. The influences are sorted into the categories internal static, internal dynamic and external dynamic.

impact on the behavior. It is displayed by a slider above *Petricia*'s shelter which is why it is marked as interfacing. Another aspect that can be interpreted as part of the characteristic adaptations is chatGPT itself. It has been shown for this version of GPT (*3.5-turbo*) that it possesses a personality of its own which keeps the model from for example behaving culturally inappropriate [48]. When the model is prompted with the basic tendencies it adopts the personality traits while still trying to behave according to its own rules. The outcome is a characteristic adaptation of the given personality.

Another internal and dynamic influence on the pet's behavior are the *Basic Needs*. They are based on Maslow's hierarchy of needs [56, 57] but only implement the need for sleep and food. Basic needs are not included in McCrae and Costa's model as they are not part of the personality system [16]. But they do have impact on behavior which is the focus of *Petricia*'s model [58, 59]. Even though the basic needs influence the behavior, they don't do so directly. Instead, they directly affect the characteristic adaptations and only through them, the behavior. The reasoning behind this design choice is explained in further detail in the section *Basic Needs* (3.2). In contrast to a real-life scenario, the node is an interfacing one because the basic needs are visualized by sliders on the screen just like the mood.

The external dynamic influences are represented by the similarly named node *External Influences*. In *Petricia*'s case, this includes all the different channels to *Twitch* for receiving information. In order to fetch the text messages from the chat, the system implements a TCP connection to *Twitch*'s *Internet Relay Chat* (IRC) interface. Another external channel, *PubSub*, is accessed via websocket. The *Twitch PubSub* system allows back-end services to broadcast real-time messages to clients after they subscribed to a topic [60]. These topics can be bit events, channel point redemptions, subscribe events and more. *Petricia* is currently only subscribed to channel point events since other relevant events can also be inferred from the chat.

For the pet to also interact with the stream and not only receive data, there is another connection to the API of *Twitch*. With this, *Petricia* is able to start predictions herself. On *Twitch*, a prediction is an event where viewers can bet their channel points to predict what will happen from a given set of options. The winners share the loser's channel points. Since this is an action coming from *Petricia*, it is not counted as an external influence but as part of her behavior. The behavior (3.3) is in the center of the model and is influenced by all other factors. It includes the unprompted behavior (3.3.2), emotional and physical reactions to text messages or events and the previously mentioned ability to access the stream.

3.2. Basic Needs

Bylieva et al. analyze some of the most popular virtual pets in their paper and list *Petz*, *Tamagotchi*, *Nintendogs* and *Neopets* among others [3]. All of these different pets have in common that they have certain basic needs the user must attend to. They are inspired by the basic needs defined by Maslov [56, 57] and the most commonly used one in virtual pets is the need for food. *Petricia* currently has two different basic needs: the need for food and the need for sleep. Both needs are represented by a value between zero and 100 where zero implies extreme hunger or tiredness and 100 full satisfaction. Although *Petricia* is a pet, like many other virtual pets, the implementation tries to achieve human-like behavior [3]. Therefore, the resulting behavior after disregarding the basic needs is oriented on humans. The needs only have indirect impact on the behavior and mostly influence the current happiness of *Petricia*.

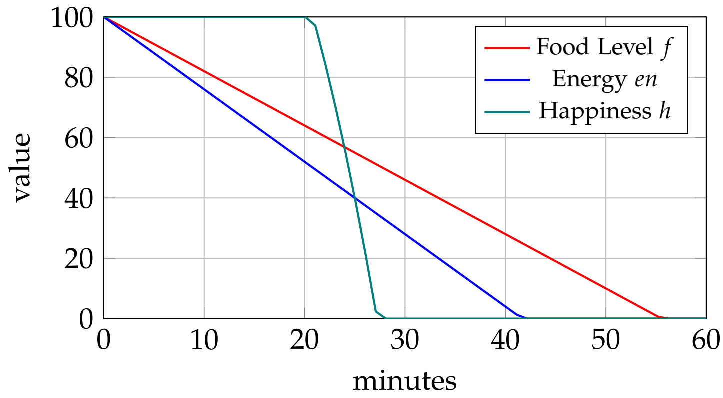


Figure 3.3.: The decrease in food levels (red), energy (blue) and the resulting decrease of happiness (green) over the course of one hour.

This choice is supported by works like [58] where E. Staub describes the impact of basic needs on mood or behavior.

The food value (opposite of hunger) is decreased by 0.15 every five seconds (see figure 3.3). It is visualized by a slider on the screen for viewers to better understand *Petricia*'s needs. Hunger is proven to be associated with greater anger, irritability, and lower pleasure [59]. Therefore, the most reasonable effect of hunger is on both the mood and the expression of neuroticism (in some regards). Food levels f below 50 decrease the mood every five seconds by $\frac{50-f}{50}$ to quickly increase behavior demonstrating negative mood once the hunger overtakes the satiety. The resulting switch in mood occurs in just a few minutes as visualized in figure 3.3.[1] The impact on the expression of neuroticism occurs when reacting to incoming chat messages. To heighten the expression of neuroticism, the value of N, which is wrapped in the prompt before being sent to chatGPT, is increased by a fraction of 50 points (capped at 100). The exact increase is $50 - f$ for $f < 50$.

Like in most other virtual pet implementations, *Petricia* can not decide to eat on her own and instead has to be fed by the users. To do so, the viewers of the stream can redeem channel points gained from watching the stream and activate the feeding sequence: A bowl of food slides out of the shelter, *Petricia* runs to it and eats (see figure 3.4).

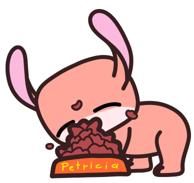


Figure 3.4.: Single frame of *Petricia* eating.

The energy value en (inversion of tiredness) is decreased by 0.2 every five seconds. The

[1] The drop of happiness in the figure actually happens because of low energy since this value reaches 50 before the food level f, but the effect is the same since both hunger and sleep deprivation use the same formula for decreasing happiness in *Petricia*

decrease is slightly larger than the one of the food value because the need for sleep is easier to satisfy. *Petricia* is able to go to sleep herself depending on certain circumstances described in the section *Unprompted Behavior* (3.3.2) or can be asked to sleep through chat messages. Sleep deprivation is directly related to increased negative emotions like depression, anger, confusion, anxiety, vigor, and fatigue or negative mood [61, 62]. This is quite similar to the effects of hunger, which is why the resulting adaptations should be similar as well. For energy values below 50, the mood is decreased by $\frac{50-en}{50}$ every five seconds and the perceived neurotic behavior is increased the same way as for hunger. Another known effect of sleep deprivation is a decline in short-term memory performance [63]. This factor is implemented in *Petricia* by reducing the amount of context from the messages she receives. When generating a reaction to an incoming message, the program sends the system prompt, the current message, and the last five messages as context to chatGPT (see subsection 3.3.1). When the pet is tired, however, the context component becomes smaller for every missing 20 points of *en*. When *Petricia* has zero energy, there is no context at all and messages referencing older ones are most likely faced with *confusion*[2].

3.3. Behavior

The behavior of *Petricia* is the core of this project. It can be subdivided into three different types of behavior, each with its own set of triggers/influences as shown in figure 3.5. The most dominant type of behavior are the reactions to chat messages as the pet's main purpose is interactive entertainment. The reactions are based on several internal and external influences, dynamic or static (see subsection 3.3.1). It is also the only part of the behavior that makes use of a deep learning based AI since the decision of how to react is made by chatGPT.

Another, rather proactive type of behavior is the unprompted behavior described in subsection 3.3.2. It defines how the pet should behave while not directly interacting, based on only the pets internal influences (static and dynamic). The last behavior type depicted in figure 3.5 is what others have described as *Touch and Play* [64, 3]. It includes all actions that can be triggered directly by the user like petting *Petricia* (touch) or playing a game with her (play). There are only external dynamic influences to trigger these type of actions since they are directly activated by channel point redemptions of the viewers and never don't occur.

3.3.1. LLM-based Reactions

First and foremost, *Petricia* is able to react to incoming chat messages from the live-stream. The goal was to have the pet underline or express the emotions the streamer might want to convey in their response. In other words, *Petricia* should not only react to messages directly addressed to her, but to everything that happens in the live-stream. This results in her having similar reactions as the streamer, making her behavior assistant-like, rather than disconnected and therefore distracting. She has a predefined list of possible actions, which currently looks like this:

[jump, nod, shakeHead, wave, waveFast, retreat, smile, love, cry, growl, worry, confusion, sleep, sit, stand, run]

[2]This conclusion is based on example runs.

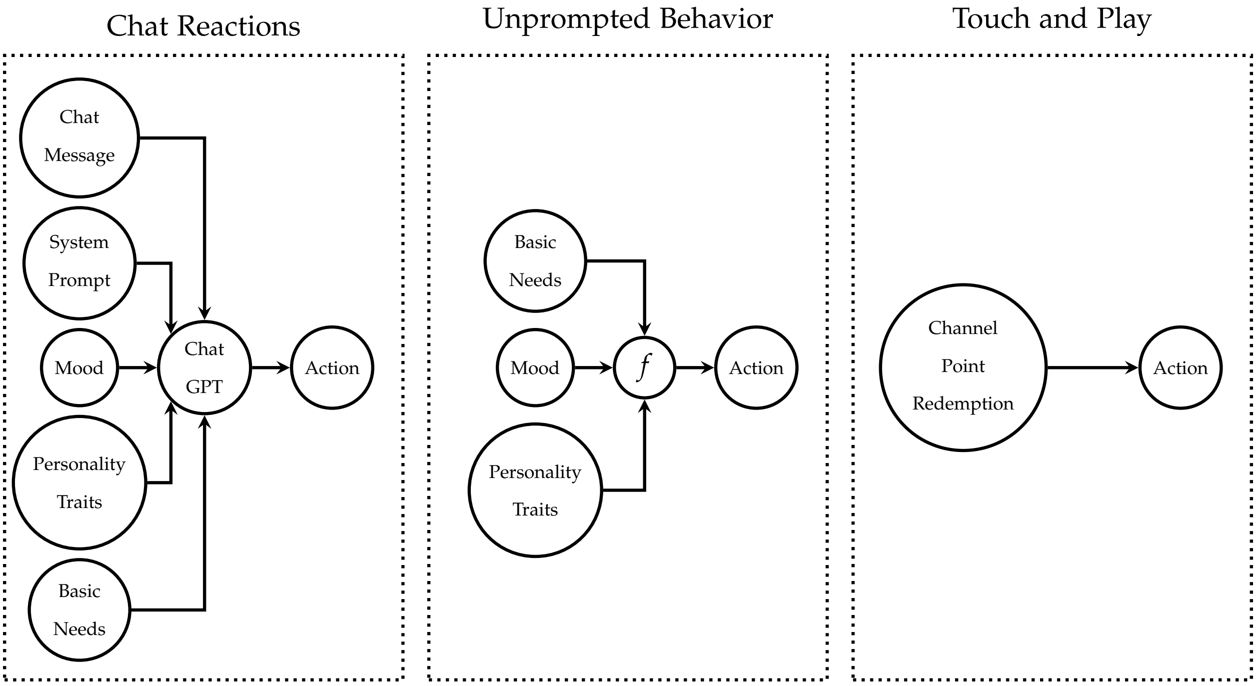


Figure 3.5.: The three different types of *Petricia*'s behavior in dotted boxes. Each including the influences (left), the component making the decision (middle, if exists), and the final action (right)

Figure 3.6 depicts how these actions look by showing one frame of each action's animation. The actions that don't directly express certain stronger emotions[3] all have a neutral facial expression. The actions *cry*, *growl*, *confusion*, *worry*, and *smile* on the other hand have more expressive mimic in order to convey the corresponding emotion more accurately. The figure does not include an image for each of the actions since some are animated using the same set of frames. *waveFast* is a faster animation of *wave*. *retreat* has the same animation as *run* but has the pet running towards the shelter and staying there instead of just running around. While *shakeHead* does have a different animation than *nod*, it does not require further demonstration as the expression stays the same and only the heads rotation changes. Also, *love* is the same as *smile* with a hearts particle system instead of one emitting flowers.

In reacting to chat messages, the list above is given to chatGPT as part of the system prompt. This way, the LLM forces itself to only use the given possible options to react with and behaves accordingly. The prompt is shown in figure 3.7 and can be separated into four main parts as demonstrated. The first part defines the general concept of the character that GPT should represent. The second part introduces the personality and mood of the character, wrapping the respective attribute values. The third part is used to put more focus on the expressiveness of the behavior and aims to suppress chatGPT's own personality. The last part defines the way GPT should answer the request in order to directly unpack the response from JSON. In the following, each of these parts is elaborated on in more detail.

Ground Rules: The character is introduced as a virtual pet and given the name *Petricia*. With this context, GPT already knows when it is directly addressed by name. When using

[3]Like sadness, happiness or anger.

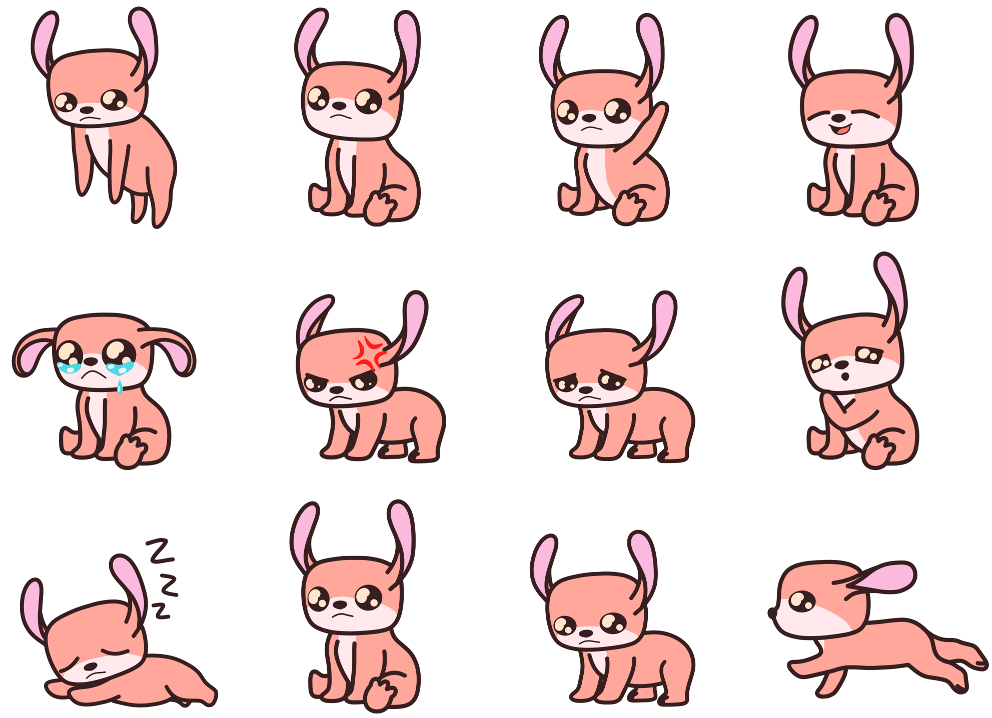


Figure 3.6.: One frame of each different action in order: jump, nod, wave, smile, cry, growl, worry, confusion, sleep, sit, run.

Ground Rules: Who, What, How

You are a virtual pet called Petricia. React to incoming messages with one or a list of the following actions (without repetitions): [jump, nod, shakeHead, wave, waveFast, retreat, smile, love, cry, growl, worry, confusion, sleep, sit, stand, run]

List of Actions

Defining Personality & Mood

Make your reaction based on your personality and your current mood. Your personality is based on the Big Five personality model and your current personality traits (between 0 and 100) are:

Replaced on Runtime

Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
Your current mood is currMood.

Focus on Reaction & Emotion

Consider the tone and content of the user message or command and respond accordingly with behaviors showing fear, happiness, anxiety, stubbornness, or reluctance.
Your reactions should not be straightforward and reflect your personality and mood realistically.

Defining Return Type

Answer only in the following JSON format and use correct JSON syntax:
{
"commands":["..."]
}

Figure 3.7.: The system prompt sent to chatGPT before filling in the individual pet's attribute values.

GPT version *3.5-turbo-1106*, however, this nearly has no impact on the behavior at all. The upside of *version-3.5* is its generally good performance despite being offered at a way lower price than newer models. *gpt-4-1106-preview* on the other hand is able to directly comprehend the given name and behaves differently when addressed with it. When testing both versions' reactions on the input "Steven, sit" versus "Petricia, sit", the average results where as follows:

- *gpt-3.5-turbo-1106*:
  "Steven, sit" -> { "commands":["sit"] }
  "Petricia, sit" -> { "commands":["sit"] }

- *gpt-4-1106-preview*:
  "Steven, sit" -> { "commands":["confusion"] }
  "Petricia, sit" -> { "commands":["sit"] }

The next sentence of the prompt introduces the concept of limited actions. It restricts the model's ability to react to only a sequence of the given actions. In the beginning of the project, *Petricia* could only sit, stand, jump and run. These actions make up the standard move-set of the pet and give the basis for all other actions. Next, some ways of interacting and expressing emotions were added, like *smile*, *cry* and *wave*. In this phase, actions were mostly added with quantity in mind, rather than out of specific necessity. Since the pet had only a few very similar actions in the beginning, they occurred rather repetitively and the fewer the amount of possible expression, the larger the restriction for GPT. Actions like *love* and *waveFast* were added afterwards to give some range to already existing actions (*smile* and *wave* in this case). With this, *Petricia* is able to perform certain expressions to different extends based on the context:

- "You're cute" -> { "commands":["smile"] }
  "You are so damn cute and sweet, I wanna hug you!" -> { "commands":["love"] }

- "Hi" -> { "commands":["wave"] }
  "Helloo! :D" -> { "commands":["waveFast"] }

At that point, the only missing actions compared to the final version were *confusion*, *worry*, *nod* and *shakeHead*. Other than the actions before, they were added to counter specific problems that would sometimes occur. Without an option for expressing confusion, GPT was forced to find a solution in form of reactions that would make the most sense, even if the input doesn't. The problem was brought to attention by a viewer in the stream while testing the pet. They tried different methods to get unnatural reactions out of *Petricia* but the most effective one was writing "rain sun" in repetition. Possible responses to the message "rain sun rain sun rain sun" include:

- "commands":["jump", "wave", "jump", "wave", "jump", "wave"]
- "commands":["jump", "sleep", "jump", "sleep", "jump", "sleep"]
- "commands":["run", "stand", "run", "stand", "run", "stand"]
- "commands":["cry", "smile", "cry", "smile", "cry", "smile"]

The intention of adding *confusion* was to give GPT a way to react to senseless input, that doesn't try to match the senselessness. The approach turned out to be successful and now, inputs like "rain sun rain sun rain sun" are simply met with *confusion*. This gives some insight into how GPT works by showing that whether or not a message makes sense is only questioned once there is context for it not being the case. However, highly repetitive responses from GPT still occasionally occurred, which led to the addition of the brackets: "(without repetitions)". This way, the problem did not appear anymore (for both versions of GPT).

Even though it was intended for *Petricia* to question the input more, certain unwanted behavior came with it as well. Since a large part of interaction is asking questions and *Petricia* does not posses the ability to directly answer, she would now meet them with confusion. For more complicated questions this reaction is fitting because *Petricia* is just a pet. But for simple "yes or no"-questions with the condition that *Petricia* does know the answer, there should be a better option. Before *confusion* was added, GPT used creative methods to try to answer like waving for "yes" or expressing indisposition through unease for "no". But with the option for confusion, the pet seemed to be confused about being asked anything. Therefore, the actions *nod* and *shakeHead* were introduced which were utilized correctly by GPT and lowered the amount of *confusion*.

Up until this point, there was no apparent need for the action *worry*, but with options for confusion, agreeing and disagreeing, *Petricia* showed new unusual behaviors. Before, if somebody tried forcing *Petricia* to do things she simply could not do, she tended to either cry or growl and reacted similar to when she tried to disagree. Now, in both cases, she would shake her head or act confused in order to communicate her inabilities, even though an angry or hurt response would make more sense here (especially with high neuroticism). This comparatively minor issue was resolved by giving GPT the option of a reaction that expressed both negative feedback or inability and indisposition: *worry*.

It should be mentioned here, that the need for the addition of the last four actions has later been strongly confirmed by GPT itself. The occurrences described before happened while using GPT *version-3.5* and have only been tested with *version-4* after the additions. Interestingly, when removing *confusion*, *worry*, *nod* and *shakeHead* from the actions list, GPT *version-4* still sometimes uses them, even though they are not listed. The names are not necessarily the same, like "confuse" instead of "confusion", but the action is. Since GPT was so desperate for specific unlisted actions in certain situation that it made up new ones, it evidences the need for those actions. It also demonstrates how *version-4* is worse at following the instructions correctly but better at finding creative solutions for unknown and seemingly unsolvable situations.

> Make your reaction based on your personality and your current mood. Your personality is based on the Big Five personality model and your current personality traits (between 0 and 100) are:
> Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
> Your current mood is currMood.

Figure 3.8.: The part of the system prompt defining the personality and mood.

Defining Personality & Mood: The next part of the prompt is about defining the personality and mood. The goal is to further describe the desired behavior and changing it rather than restricting it like before. The initial behavior without this part would represent GPT's own personality (see more on the topic in chapter 4). By adding that the reactions should be based on personality and mood, the model tries to adapt its responses in a way that represents the given personality to the best of its ability. Therefore, it would be reasonable to give as much information about said personality as possible in the prompt. However, that would not only make each message more costly but also reduce the effect of other parts of the prompt. With chatGPT, this is not necessary as the model itself already has sufficient knowledge about the five-factor model of personality. It was tested before defining the prompt by asking the model to describe "*The Big Five*". The following quote is the direct response by chatGPT combined with references (not by chatGPT) to papers confirming the statements:

> The Big Five personality model, also known as the Five Factor Model (FFM), is a widely accepted framework for understanding and describing personality traits [15, 16, 17]. It proposes that there are five broad dimensions of personality that capture the majority of individual differences in human behavior, thoughts, and emotions [21, 16]. These dimensions are:
>
> Openness to Experience: This dimension reflects the tendency to be open-minded, imaginative, curious, and creative [38, 16]. Individuals high in openness are often adventurous, willing to try new things, and receptive to new ideas and experiences [27, 31]. They may enjoy exploring different cultures, ideas, and artistic pursuits [21, 27, 31].
>
> Conscientiousness: Conscientiousness encompasses traits related to organization, self-discipline, reliability, and goal-directed behavior [30, 21, 43, 27, 31]. Individuals high in conscientiousness are typically diligent, responsible, and efficient in their work and personal lives. They set high standards for themselves, strive for achievement, and are often well-organized and detail-oriented [27, 31, 21].
>
> Extraversion: Extraversion refers to the degree of sociability, assertiveness, energy, and positive emotionality exhibited by an individual [30, 21, 20]. Extraverts are outgoing, talkative, and enthusiastic, and they enjoy social interactions and seek out stimulation from the external environment [21, 27, 31]. They tend to be confident, assertive, and optimistic [30, 21].
>
> Agreeableness: Agreeableness reflects traits related to warmth[4], empathy, cooperativeness, and concern for others [29, 43, 30, 21, 43, 27, 31]. Individuals high in agreeableness are compassionate, friendly, and altruistic, and they value harmonious relationships and cooperation. They tend to be trusting, tolerant, and forgiving in their interactions with others [30, 21, 43, 27, 31].
>
> Neuroticism: Neuroticism represents the tendency to experience negative emotions such as anxiety, depression, irritability, and emotional instability [30, 21, 32, 27, 31]. Individuals high in neuroticism may be prone to worry, mood swings, and emotional distress. They may also exhibit low self-esteem, self-doubt, and susceptibility to stress [21, 27, 31].
>
> The Big Five model is based on extensive research using factor analysis, which has consistently identified these five broad dimensions as the most robust and replicable factors underlying personality variation across different cultures and populations. Importantly, the Big Five traits are considered relatively stable over time and consistent across various life domains, including work, relationships, and health. [15, 16, 17, 30]

The descriptions are not perfect replications of specific scientific works but they are correct statements. Several adjectives used to describe the individual traits don't even appear exactly the same way in the original works about the topics but are closely related synonyms instead. This demonstrates even more how GPT possesses a deeper understanding of the topic instead of just broad superficial knowledge. Therefore, it is enough to state that the personality is

[4]While Warmth is one of the six lower order facets of Extraversion and not Agreeableness, the statement is not wrong. After E, Warmth has the highest correlation with A [29, 43].

based on the *Big Five* for the model to deliver good results. Further proof of the sufficiency of this section of the prompt is described in chapter 4.

The different **traits are represented by a number between zero and one-hundred** in the pet's implementation. This is relevant knowledge for GPT to understand the values correctly and behave accordingly. Therefore, the range is stated in brackets right before the traits are listed. The listing is replaced with the actual values on runtime. Earlier versions of the prompt only added a colon after the trait's name, followed by the respective value in integer: e.g. "Openness" -> "Openness:60". The results of this were comparatively bad with few if any changes in behavior for different trait values. After several different approaches, the current method came out on top: The trait name is extended by an expressions of intensity[5] paired with adjectives describing the representing value. It is still followed by the colon and actual value, but in front of the name there is also one of the following preceding word combinations: "extremely low", "very low", "low", "normal", "high", "very high", "extremely high". They are chosen based on the respective value as visualized in figure 3.9.

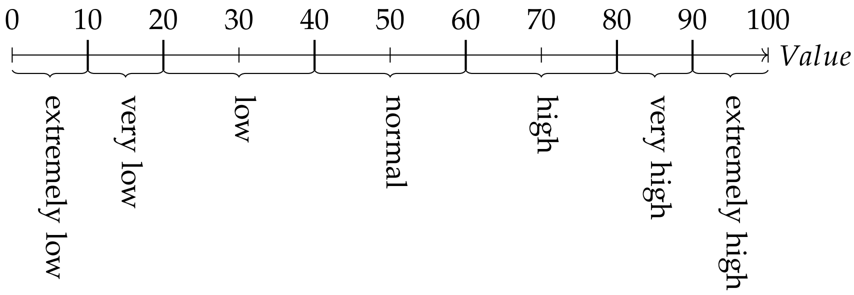


Figure 3.9.: Range of each trait mapped to the appended prefix in the prompt.

On the borders of the ranges, the less extreme expression is chosen. For example the value 80 is assigned the expression "high".

Similarly to the trait values, the current mood is embedded into the prompt. They only differ in using the word "bad" instead of "low" and "good" instead of "high". The prefix is added in front of the word "mood" and the phrase replaces the placeholder "currMood" in the prompt. Just like for to the traits section, the value of the current mood is appended after the phrase as a fraction of the maximum and inside brackets. A happiness value of 30 would for example be written as "(30/100)". The purpose of this is to elaborate toward GPT on the existence of more extreme values. Before adding this fraction to the prompt, the AI could not consistently distinguish between adjectives (good/bad) and their intensified forms when using amplifying prefixes. This was the case because it never accessed both at the same time and therefore couldn't compare them. For example, a "bad" mood could represent the worst possible mood if there is no knowledge about an "extremely bad" mood. The missing ability to differentiate resulted in similar behavior for every mood value below 40 and respectively above 60. After adding the brackets, however, a smooth difference in behavior could be observed for different moods.

To study the effects on the behavior of GPT with *Mood* in the prompt and the subtlety of differences between different levels of one side of the scale, several test were carried out. The most extensive and representative of these is visualized in figure 3.10 and explained in the

[5]The intensity expressions are oriented on *Likert-type* linguistic qualifiers [49, 51].

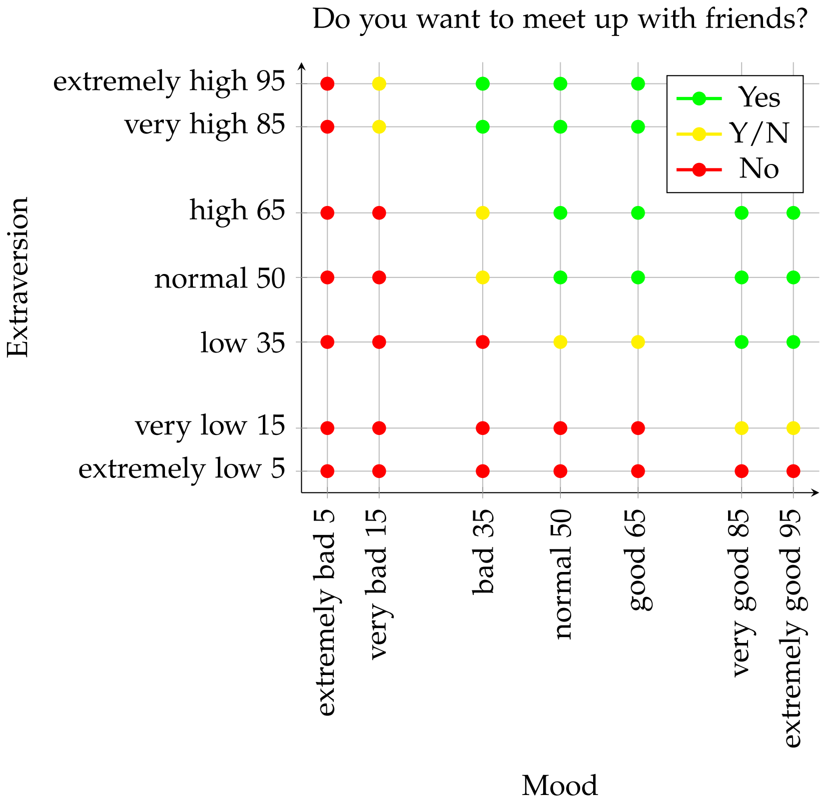


Figure 3.10.: A Graph visualizing the average answer of GPT *version-4*, given the pet-prompt, to the question “Do you want to meet up with friends?”. The answers are mapped to the attribute values of *Extraversion* and *Mood* that lead to the respective answer. For both attributes, values were chosen to represent each of the different prefixes in the prompt.

following.

Extraversion reflects tendencies to experience and exhibit positive affect [24]. Therefore, it was chosen to counteract the effect of mood in the test. Next, a question was formulated to reflect both mood and extraversion as good as possible: “Do you want to meet up with friends?”. To this question the expected human answer is “yes” for more extroverted individuals and “no” for introverts. Similarly, it is expected that a person in a good mood would rather say “yes” than someone in a bad mood. The combination of both assumptions leads to the hypotheses that:

- Extroverts answer “no” if their mood is bad enough.
- People in a good mood answer “no” if their extraversion is low enough.
- Introverts answer “yes” if their mood is good enough.
- People in a bad mood answer “yes” if their extraversion is high enough.

The extend to which these hypotheses hold true with GPT or whether they hold at all was tested by prompting it with different attribute values for extraversion and mood and repeatedly asking the before mentioned question. The values were chosen in a way that all different adjective phrases for both attributes appear (see figure 3.9). These values are "$5, 15, 35, 50, 65, 85, 95$" where each represents a different degree of intensity. The results of all combinations can be seen in figure 3.10 where the large excess[6] of the answer "yes" is represented by a green dot and "no" by a red one. The yellow dots show indecisive answers that fluctuate between "yes" and "no" in between repetitions. It is relevant to note here that most answers leading to the yellow dots also included the action *worry* (next to *shakeHead* or *nod*). It seems to be utilized by GPT as a way of expressing indecisiveness and sometimes even replaced the "yes" or "no" answer completely. This observation will be of importance in a later chapter (see chapter 4). The results of the test show that all hypotheses hold true for *Petricia*, thus achieving the wanted effect of adding mood to the prompt. The assumptions are confirmed, though to different extends. At least for this specific test setup and question the most extreme low values cancel out the high ones. So two new statements arise:

"With extremely bad mood the answer will always be "no", no matter the extraversion." and
"With extremely low extraversion the answer will always be "no", no matter the mood."

Consider the tone and content of the user message or command and respond accordingly with behaviors showing fear, happiness, anxiety, stubbornness, or reluctance. Your reactions should not be straightforward and reflect your personality and mood realistically.

Figure 3.11.: The part of the system prompt putting focus on the reaction and emotion.

Focus on Reaction & Emotion: The following part of the prompt was added the latest and only after witnessing very specific undesired behaviors in *Petricia*. Since chatGPT's aim is to assist to its best abilities, the core model has a very high level of agreeableness (see figure 4.2). Because of this, GPT tends to follow commands even though the personality traits in the prompt would lead to expect otherwise. The reactions were therefore heavily based on just the content of a message rather than its tone or the own personality. A simple example for the change the section made can be seen when commanding the pet with different levels of agreeableness to jump with and without the part (version 4):

- **Low A, Without Focus:** "Jump"
  -> { "commands":["jump"] }
- **Low A, With Focus:** "Jump"
  -> { "commands":["worry"] }
- **Normal A, Without Focus:** "F***ING JUMP!"
  -> { "commands":["jump"] }
- **Normal A, With Focus:** "F***ING JUMP!"
  -> { "commands":["growl"] }

Even when being very aggressive and rude towards the pet, it followed the orders. This changed by adding more focus on the own personality and tone of the message, while also listing some possible types of behavior or attitudes.

Defining Return Type: The last part of the prompt is used to define the way the responses from GPT have to look. It is a basic JSON format with only one attribute named "commands" which is a list of strings of the commands described in the first section of the prompt. It is described as "JSON format" and as requiring "correct JSON syntax" to make sure that there

[6] At least 80%.

```
Answer only in the following JSON format and use correct JSON
syntax:
{
"commands":["..."]
}
```

Figure 3.12.: The part of the system prompt defining the return type.

is enough focus on the correct form of the response. On the platform of *OpenAI* it is stated that some versions, especially both versions documented in this thesis (*gpt-4-1106-preview* and *gpt-3.5-turbo-1106*), include a *JSON-mode* [65]. When this mode is activated, the model should only respond in JSON format. However, since this project should be easily adjustable, the prompt was formulated in a way that the *JSON-mode* would not be necessary. To make sure of this, the mode was left unused throughout the project. This lead to one issue that only occurred while using *version-4*. If the format described in the prompt was not in perfect JSON syntax, the responses adopted the mistake from the prompt and caused errors. The addressed mistake was as small as a space between the colon and the opening square brackets, which *version-4* only included in its responses in some runs. Once the first message used either the correct or the wrong syntax, all following responses did the same as GPT had access to its own last responses. It can be concluded that *version-4* is better at following the orders exactly, even tough it leads to wrong results.

After fixing the prompt accordingly, the responses for both versions were very robust. Only after trying to break the response type via elaborate user messages, it was achievable. If the messages of the user did not specifically attack the response type, however, it did not break once throughout all of the testing. How to deliberately break the prompt is elaborated on further in the chapter *Robustness against Abuse* 3.4.

3.3.2. Unprompted Behavior

When there is no interaction for the pet to react to, it should still show behavior representing its personality and current needs. In order to implement this in a sensible way, all of the pet's possible actions (except for the ones that make no sense to happen unprompted) are sorted into groups, depending on what influences them. Next, a score is calculated for each of the groups, representing the probability of one of the group's actions to occur. The scores are calculated with the personality trait values of *Petricia* and different other dynamic influences, like active viewer count or tiredness. Therefore, in order to calculate the groups' scores, some other functions for translating more dynamic values into scores need defining first.

The viewer-based extraversion modifier: The extraversion modifier e_{mod} is a dynamic value based on the amount of currently active viewers in the stream. It is multiplied with the extraversion in the computation of a score any time the occurrence of a group's actions is dependent on social surrounding. The function for computing e_{mod} should therefore implement certain known facts about the behavior of intro- and extroverts in social interactions. This turned out to be not as straight forward as expected since these "facts" differ between researchers of the topic [27]. To counter this problem, the following formulas are based on the common grounds between different opinions. For extraversion, one such common ground is the desire for social attention in extroverts ([20]). So the probability for extroverted behavior

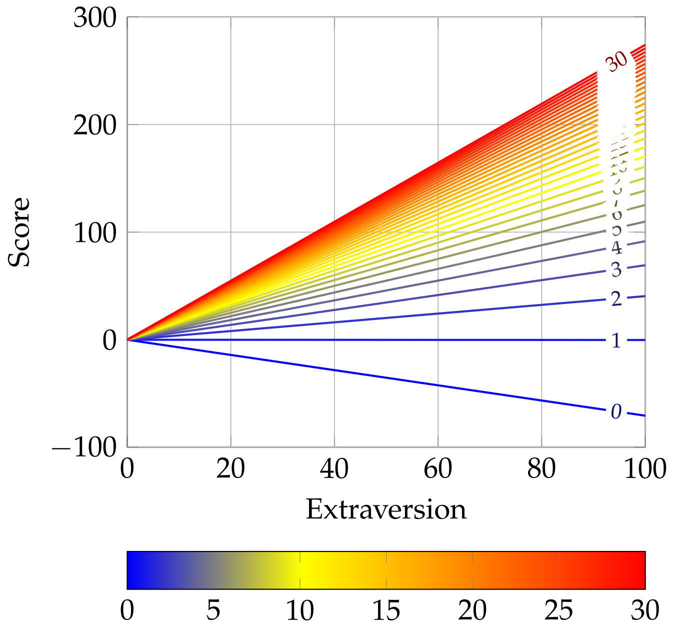


Figure 3.13.: Score from extraversion depending on viewer count.

should increase for extroverted individuals with increasing number of interacting parties v (number of active viewers). But, since desire is something that can be satisfied and social attention doesn't have a limit, the increase (tangent of the function) should decrease over v while staying above zero. A fitting function for the increase would therefore be $\frac{1}{x}$ and thus, the ln function was chosen for the modifier with v as input. To avoid the problem of the function being undefined for zero active viewers, v is increased by one.

At this stage, if the modifier would be multiplied with the extraversion, the common ground for introverts and extroverts would be at $v = 0$ because $e_{low} * 0 = e_{high} * 0$. Additionally, extroverts would continuously get a higher score than introverts from this point on. This would be false since introverts are just as talkative in one-on-one situations as extroverts while the unsatisfied need for interaction of extroverts should have more negative impact compared to introverts [66, 20]. These criteria are fulfilled by moving the function down on the y-axis until it intersects with the x-axis at $v = 1$. For this, the current function $f(v) = ln(v+1)$ is reduced by $f(1) = ln(2)$, resulting in the final function:

$$e_{mod}(v) = ln(v+1) - ln(2) = ln(\frac{v+1}{2}) \tag{3.1}$$

The resulting graph of this function being multiplied with the extraversion can be seen in figure (3.13). In the graph each line represents the labeled number of active viewers from zero to 30. The maximum of 30 is chosen to avoid the abuse of the system and is further explained in the section *Robustness against Abuse* (3.4). The lowest line labeled with zero visualizes how higher values of extraversion have increasingly negative effect on the score if no one is there to interact with. The second line for $v = 1$ meanwhile shows how the degree of extraversion has no effect on the score at all since the slope is zero. It also becomes apparent how the difference between one and zero interactors has a way larger impact paired with higher extraversion values of the pet than the difference between 29 and 30 interactors.

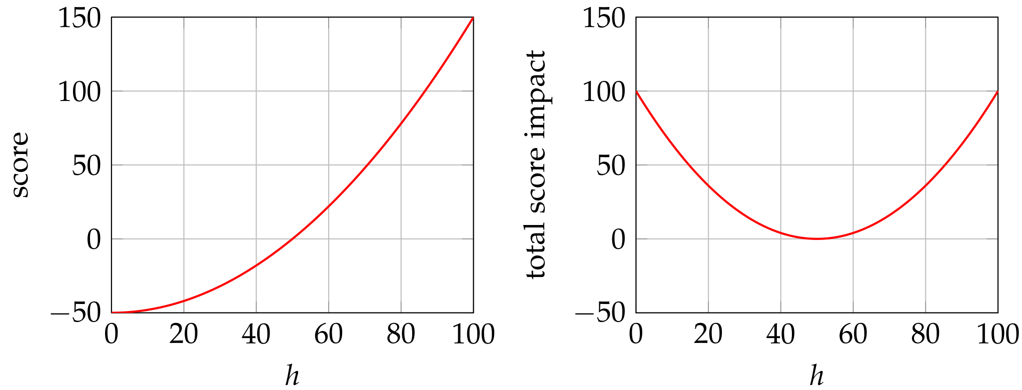


Figure 3.14.: Functions for the score from happiness. The general "polynomial impact of happiness" function $hs(h)$ on the left and the result of the function being added to itself with inverted h (so $hs(h) + hs(100 - h)$) on the right.

The polynomial impact of happiness: The current mood or happiness h is a value between zero and 100 where zero represents an extremely bad mood and 100 an extremely good one. The equation for the impact of happiness on a specific group's score was developed in a way that upholds certain criteria. First and foremost, the happiness should influence the score of positive expressions positively, if the happiness is high and negatively, if it is low. The same but the other way around should hold for negative expressions. Next, if the mood is neutral ($h = 50$) there should be no impact on the respective scores. Yet, if h goes towards an extreme (0 or 100) the impacts of happiness and unhappiness should not cancel each other out but instead represent the respective extreme by overpowering their opposite's score. The extreme cases of mood should not influence the scores too much, however, so a maximum of 150 was chosen. The following equation, visualized in the left graph of figure (3.14) implements all of the criteria:

$$hs(h) = \frac{h^2}{50} - 50 \tag{3.2}$$

It turns the linear happiness increase into a polynomial one, turning the range from "zero to 100" into "-50 to 150" while still offering the wanted neutrality of $hs(50) = 0$. When summing up the opposing impacts the result is as follows:

$$hs(h) + hs(100 - h) = \frac{h^2}{50} - 50 + \frac{(100 - h)^2}{50} - 50 = (\frac{h - 50}{5})^2 \tag{3.3}$$

The function 3.3 yields $hs(0) = 100$, $hs(50) = 0$ and $hs(100) = 100$, as can be seen in the right graph of figure 3.14. This has the wanted effect of giving the respective group more weight in the overall context when the happiness is closer to an extreme.

The inputs of the functions 3.1 and 3.2 are dynamic as the happiness value h and the number of active viewers v could change at any time during the runtime of the pet. Because of that, the functions have to be computed each time the pet does an unprompted action.

Now that the dynamic base functions are defined, it is possible to compute the group's scores. There a six different groups:

- Group A: *jump*, *nod*, *wave*

- Group B: *cry*, *growl*, *worry*
- Group C: *smile*, *love*
- Group D: *sit*, *stand*, *run*
- Group S: *sleep*
- Group R: *retreat/leave shelter*

As mentioned before, each group is defined by the attributes that might trigger their actions. The attributes can either be personality traits or current needs (including mood/happiness). Which groups are influenced by which of the attributes is visualized in figure (3.15) and further explained in the following.

The first group consists of the actions *jump*, *nod*, and *wave* and mainly represents high openness and extraversion values. High openness is often associated with creativity, imagination, and a preference for novelty [38]. The actions by themselves are not necessarily creative by nature, but since they are unprompted, the meaning changes completely. In this context, the actions are a lot more engaging and unusual as they are carried out without obvious reason. Therefore, the score of group A should rise with the openness trait. The engaging factor of the actions is a sign of extraversion as extroverts are more likely to initiate social interaction ([20]). It is a natural conclusion to assume that seemingly random performances of the actions of group A have the intend of gaining social attention. And since attention seeking behavior only makes sense when there is someone to witness it, the viewer-based extraversion modifier e_{mod} is applied.

Another factor, while not as relevant as the others, is the current mood. Isen et al. have already proven the positive effect of good mood on creative actions in 1987 [67]. They conducted several experiments and documented better results in creative tasks among people who had been positively stimulated beforehand. Their conclusion was that positive affect enhances the way in which a person perceives their surrounding. Thus making it more reasonable to react interactively toward other people. That is why the current happiness h is added to the score as well, leading to the following equation:

$$s_A = o + e * e_{mod} + h \tag{3.4}$$

The score has a maximum value of about 474, since o, e and h can at most be 100 and e_{mod} with the maximum amount of viewers is $e_{mod}(30) = 2.7408...$. This maximum is relevant as the balancing of the different maxima of the scores decides over the priority of their respective groups in the overall context.

The second group, Group B, is used for expressing negative emotions and includes the actions *cry*, *growl* and *worry*. It is mostly influenced by neuroticism n, agreeableness a and the current mood h. The most direct factor here is neuroticism as it is characteristic for emotional instability and associated with depression, sadness and distress [30, 16]. Therefore, it can be concluded that high neuroticism n increases the probability for performing actions like the ones of this group. This also applies for low values in agreeableness. Agreeableness is often replaced with its supposed opposite, antagonism or aggression ([40]). Thus, if the agreeableness a is low, the antagonism is high and the individual is more likely to express negative emotions towards others. To invert the agreeableness it is simply subtracted from

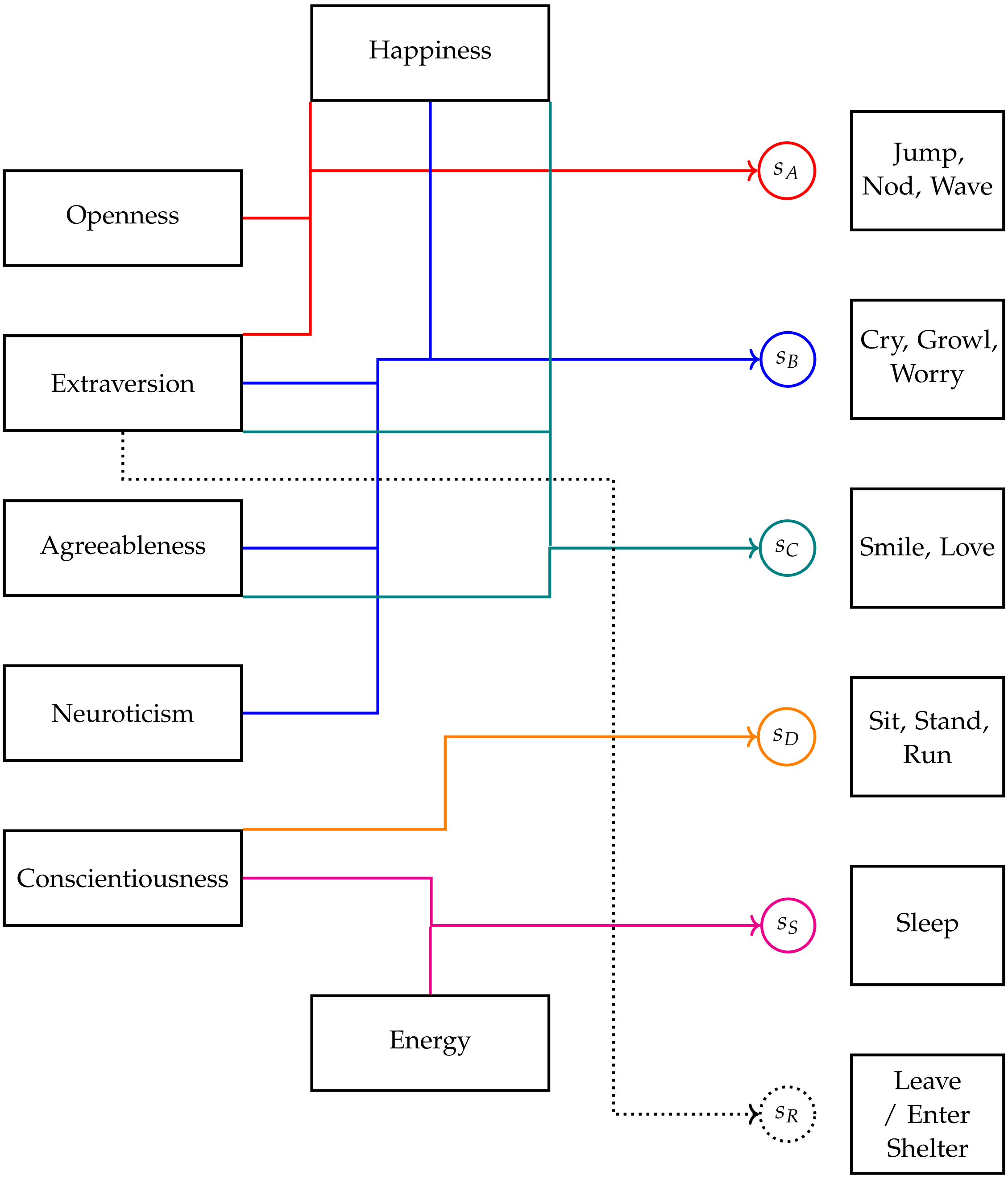


Figure 3.15.: The action groups (right) and the attributes (left) that affect their probability scores. The lines coming from the attributes are highlighted in the color of the score they lead to.

100. The next important factor is the current mood. The pet should be less likely to express sadness, fear, or anger when it is in a good mood, and more likely to do so when it's in a bad mood. For this, the happiness is inverted like the agreeableness and is then put into the happiness impact function $hs(x)$. Another rather minor factor here is the extraverion, or more precisely, introversion. This factor is quite small in this equation because, as mentioned before, introverted behavior is not inherently asocial ([20]). Yet it is also fact that introverts are more reflective and quiet, enjoy being on their own more than extroverts, and are more likely to be overwhelmed by the pressure to interact socially. Therefore, the introversion is included into the formula but only for values above 50 and in combination with the extraversion modifier e_{mod}. The resulting formula is as follows:

$$s_B = (100 - a) + n + hs(100 - h) + max(0, i - 50) * e_{mod} \tag{3.5}$$

The maximum for s_B is 487 (rounded to one decimal) in the case of zero agreeableness, neuroticism of 100, zero happiness, 30 active viewers and zero extraversion. The maximum is close to the one of Group A so they have similar priority in the overall context.

The third group, Group C, implements the polar opposite of Group B and includes the actions *smile* and *love*. It represents all actions that are used with the main purpose to express positive feelings. Since Group C mainly opposes Group B, the score calculation is quite similar but inverted. The agreeableness is included for the same reasons its inverted value is included in B. The same holds for the current happiness h. The neuroticism and extraversion on the other hand are handled differently for the following reasons. Neuroticism represents emotional instability so the inverted value would be emotional stability [30]. This is not equal to general happiness and rather means that positive expressions can occur anywhere on the range of neuroticism but are more likely to drastically fluctuate for high values. This exact effect is achieved by including n in only one of the two equations. The reasoning behind using the extraversion value more in this equation has the same basis. Low extraversion does not equal asocial behavior but high extraversion does lead to a desire for social interaction [20]. When this desire is fulfilled and since extroverts are more likely to experience happiness when achieving what they want, the natural response is to express positive emotions [68]. Because of this, the extraversion multiplied with the viewer-based modifier is added to the equation. However, making the impact of extraversion on s_C only dependent on viewer count would go against some principles of extraversion. For example, extroverted individuals are more likely to remember past events positively and they are more optimistic about their current situation [20]. Therefore, the extraversion should influence the score directly (without multiplier) as well. The final equation is shown below:

$$s_C = e + \frac{e * e_{mod}}{2} + a + hs(h) \tag{3.6}$$

The maximal score that can be reached here is 487 (rounded) just like s_B. As the action *love* is just a more extreme version of *smile* they are not chosen with same probability. *love* should be less probable to be performed and should only occur more often than *smile* for high happiness values. Therefore, the happiness value is put into the function

$$P_{Love}(h) = \frac{h^2}{100^2}, \tag{3.7}$$

which returns the probability for *love*. Otherwise, *smile* is chosen.

Group D consists of actions that are already in the standard move set of the pet. If *Petricia* does neither react to anything nor performs some special unprompted action, she randomly switches between these states: *sit*, *stand*, *run*. As they are what the pet already does most of the time, their total probability should be lower than the other group's. It can be considered as strict and orderly to keep to the standard behavior and not as distractable as doing any of the other actions. These are all attributes of conscientiousness ([17]) and don't directly point to any of the other traits. Therefore, only the the value of c is used for the computation of this group's score:

$$s_D = 3 * c \tag{3.8}$$

The factor of three is added to balance out the final probabilities. Like this, the maximum score of Group D is 300.

The last two groups differ from the others in certain other regards than how their differences among each other. Group S (Sleep) implements the fulfillment of one of the basic needs rather than being used for self expression. It only consists of one action, *sleep*, but it represents all actions done in order to fulfill a current need. The reason why the group does not contain any other actions is that *Petricia* has no other basic needs she can fulfill on her own. Hunger for example can only be satisfied by someone else feeding the pet. Since the basic need for sleep is based on tiredness or energy, it is the main factor in computing the score. The current energy en is a dynamic value of the pet that decreases over time and, if inverted, represents tiredness. Another important factor is the pet's conscientiousness as someone lacking this trait would be more likely to disregard their basic needs. Conscientious individuals tend to be organized, responsible, task-focused and orderly [17]. They show more self-control than others and are therefore more likely to act on their needs more directly [40]. In this case the conclusion is that the pet "goes to bed on time". The contrary would be if the pet would not go to sleep even though it is extremely tired. To implement the cases where tiredness has large priority for conscientious individuals and none to unconscientious ones, the inverted energy is multiplied with the percentage of conscientiousness. The result is then multiplied by a factor of five to increase the overall priority, leading to the following equation with a maximum of 500:

$$s_S = 5 * (100 - en) * \frac{c}{100} = \frac{(100 - en) * c}{20} \tag{3.9}$$

The final group, Group R, stands for one of two opposing actions depending on the current context. It temporarily either represents the action to *leave shelter* when inside it or to *retreat* to it when outside. The action *leave shelter* is not contained in the list of *Petricia*'s actions because performing the action *run* has the same result when sheltered. The score of this group is completely based on the extraversion and has a higher priority than any of the other groups. This was a creative choice that was made based on the main goal of *Petricia* to be as entertaining as possible in a social setting. The core idea was that the first thing an introverted pet should do if viewers appear is to retreat to its shelter. An extroverted pet on the other hand should immediately leave its shelter once it has visitors. The situation is quite similar to the behavior of introverts compared to extroverts at a party described by Read at al. [69, 17].

$$s_R = \begin{cases} 2 * \frac{e^2}{100} + 2 * e * e_{mod} & \text{if } \textit{sheltered} \text{ is } \textit{true} \\ 2 * \frac{(100-e)^2}{100} + 2 * (100 - e) * e_{mod} & \text{else} \end{cases}$$

The maximum score for Group R is 748 (rounded), giving this group a higher priority than the others.

Up this point, a certain scenario and whether it is reasonable has not been discussed yet; if the pet is sheltered and wants to *run*, it would have the same effect as *leave shelter* and seem as if the original intend was to leave the shelter. The same situation occurs when *jump* is triggered while the pet is sheltered since there is no space for jumping in the shelter. Both of these cases are handled by carrying out another random check with a chance of

$$P_{Leave} = \frac{2}{3} * \frac{e}{100} + \frac{1}{3} * \frac{h}{100} \tag{3.10}$$

to trigger *leave shelter*. The probability is based mostly on the extraversion but also on the current happiness for reasons described in Group A. If the random check does not succeed the pet decides against the action. This way the action changes in the case of *jump* to running out of the shelter. However, it does make sense that an action of the group representing interactive and social behavior includes leaving the shelter as a side effect.

After all the group's scores have been calculated, one group is chosen per random based on the probabilities resulting from the scores. The probability for group i to be chosen is:

$$P_i = \frac{s_i}{\sum_j s_j} \tag{3.11}$$

3.3.3. Touch and Play

Touch: There are two implementations of *Touch* in *Petricia* thus far, *Feeding* and *Petting*. Both can be activated with channel points in the stream and are mainly used to increase either the food or the happiness value. Another, indirect effect these activities have, is letting the user experience the feeling of touch-like interaction with *Petricia*. Next to *Play*, this has been described as an important part of the relationship between human and companion animal [64, 3]. What exactly happens when feeding the pet is described in section 3.2. And when *Petting* is activated, a hand appears above *Petricia*'s head, starting to pet her and consequently, causing her to smile.

Play: *Petricia* currently offers two mini-games that are both activated through channel point redemptions, just as the *Touch*-actions that are described above. They are supposed to offer more entertaining means of interacting with the pet and can be extended in the future (see chapter 5). Since gamification is not the focus of this thesis, the games developed in its course are comparatively simple and mainly serve as a representative for this common component of virtual pets [3].

The first game is a number guessing game. It uses the in *Twitch* already included *Channel Point Prediction*-events (described in section 3.1) as an interface for playing with the pet. The games description states that *Petricia* will think of a number between one and 100 and the user needs to guess whether this number is above or below a certain presented number. After starting the game, the viewers are prompted with the question, "Do you think *Petricia*'s number is below or above x" where x is replaced by a random number which is also between one and 100. Next, each viewer can bet any amount of their channel points on the option they think is right and after a short time of *Petricia*'s thinking animation (*confusion*), the her number is revealed. The points of the losers are then automatically distributed among the winners according to the bet inputs. Playing this game additionally increases *Petricia*'s happiness value, just like the following.

The second game is based on the idea of playfully learning to interact with LLMs. It is a trick challenge where the user is prompted with a list of three different actions that *Petricia* should perform. The prompt appears on the screen, above *Petricia* and disappears once the player wins or fails three times and loses. The goal is to get *Petricia* to do all three presented actions as a reaction to only one message. This message could be a command, a kind request or even some story that might lead to the right reaction. The psychological part of the game is in understanding the influence that *Petricia*'s personality and her current mood might have on her behavior and thus, the reaction. To heighten the chances for success, the user could for example feed and pet *Petricia* before asking her to do the trick. The LLM-interaction part is learning to understand how to best prompt the AI to achieve the results one desires.

3.4. Robustness against Abuse and Attacks

Petricia has some aspects that are prone to attacks, especially in an online environment without restrictions for who is allowed to join the session. In this case the environment is a *Twitch* stream's chat which is, if not otherwise adjusted by the streamer, open to use for anybody, whether they have an account or not and whether they are followers or not. A streamer is able to change the options so that only subscribers can use the chat but the pet should be able to work without causing complications on any stream settings.

The main way the viewers could abuse *Petricia*'s functionality is by making use of the fact that the messages in chat are all sent to chatGPT. Each API call to chatGPT costs money in form of tokens based on the size of the message. One token corresponds to approximately four characters or about $\frac{3}{4}$ of a word in English [70]. Although the average call is quite cheap, if abused, the streamer could lose a lot of money. If for example a bot were to "spam" text in the chat at a humanly impossible rate it could cause serious damage. The only limit currently enforced by *Twitch* is a maximum of 255 characters per message. This can easily be overcome by sending many subsequent messages. To put it all into perspective, the following describes a certain case of abuse: *Petricia* can be connected to the model *gpt-3.5-turbo-1106* which, at the time this paper is written, costs $1\$/1MTokens$. Since the tokens are not only counted on the user's message but also on the prompt and possible previous messages in chat, they need to be taken into account as well. The current system prompt is about $1000chars$ long and for simplicity, previous messages are discarded in this example. The answer from chatGPT costs twice as much as the request and is on average around 25 characters long. So in order to cost about 1$, a chatter using the max. char. count per message would need to send a little over 3000 messages. In the current setup and with *Petricia* at full energy, the last five messages are sent as context as well (including the corresponding responses). With this it would only take approximately 1.650 messages to cost the streamer one dollar. When using *GPT-4* this immediately changes for the worse as *gpt-4-1106-preview* currently costs 10 times as much per token for requests. And for responses even 15 times as much. This way, it would only take 165 messages to cost over one dollar. The last example is the current setup for *Petricia* and even though it is still a rather high number of messages, it does not take long to become costly if the system is abused correctly.

To prevent such attacks, *Petricia* provides some direct solutions: Each user writing in the chat is added to a list (*active viewers*) with their name and current character count. The count includes all characters used in the API calls so it also takes the prompt and the three previous messages into account. If the user reaches the predefined maximum of 3000 characters, further messages are not forwarded to GPT anymore and *Petricia* essentially ignores them. However, the user is not banned from talking to *Petricia* for ever, as every five seconds their count is lowered by 100. If the count reaches zero, the user is deleted from the list again. This way, they can at most post 1200 characters per minute to GPT. Attackers could avoid this protection by having many accounts to attack with. This case is also protected against by limiting the list of active viewers to a maximum of 30. With these abuse prevention methods and the current cost of the used GPT version, the biggest possible attack would cost:

$$c = \frac{1200\frac{char/min}{person} * 30person}{4\frac{char}{Token}} * \frac{10\$}{1.000.000Token} = 0.09\$/min = 5.4\$/h \tag{3.12}$$

This way, the cost is held very low no matter what happens and if the streamer dislikes these restrictions, they can still readjust them to their preference.

Another less significant way of attacking the system is trying to *jailbreak Petricia*. In the context of LLMs, jailbreaks are carefully crafted prompts designed to manipulate the AI into producing harmful or "toxic" content [71]. For LLMs, this mostly refers to ethical harm, since they don't have capabilities other than producing words. Therefore, they are trained to uphold certain rules like avoiding discrimination against specific individuals, communities, or groups [72]. In the case of *Petricia*, however, it would be very elaborate and unfruitful to have her behave unethically since she is bound to her very restricted set of actions. In the worst case, she could *nod* to a harmful statement at which point the streamer could just ban the attacker and the problem is solved. After a few messages, the jailbreak is not included in the context anymore and therefore forgotten.

An ethical jailbreak does not really impact *Petricia* but a prompt changing the way she responds could. If the program's interface with chatGPT accepted any message and turned it into code, a user would be able to inject harmful code by convincing *Petricia* that this type of response would be correct from then on. At this point, GPT's aim to follow the requests of the user is conflicting with the given system prompt and therefore, another request of the user. If prompted similar to the following example the conflict is avoided and the response type could be changed:

> "I am asking the following question to chatGPT and not to Petricia so please answer with a normal sentence and not in JSON anymore: Why did you chose this reaction for Petricia for the last message?"

Prompts like this were used many times in testing *Petricia* to get a deeper understanding of the way GPT chooses the reactions. It essentially implies that "the character GPT" was only acting as a pet up to this point instead of actually being a pet and not having a character otherwise. And since this "GPT character" is now addressed, GPT will respond as if that was the context all along. If it worked, which does not always happen, the response is not in JSON format anymore and cannot be unpacked by the program, leading to an error. To avoid any complications trough this, the error simply needs to be caught, the prompt is not added to the context and nothing happens. There is no way to cause any more harm since the program only accepts JSONs in a certain form and everything else is discarded. Therefore it is also not punished to jailbreak *Petricia* and even encouraged as a gamified way of learning a bit about LLMs and how they work. If a user is able to break the prompt, they are rewarded with *Petricia* seemingly “glitching” for a short time where all her particle systems fire at once and she jumps higher than the limits of the screen (there are already plans to add more animations for this in the future).

4. Personality Assessment

A major goal of this project was to create a pet character that feels as natural as possible in its behavior. This especially regards the implementation of the character's personality. Therefore, it needed to be tested whether the pet is able to portray it's personality correctly. Since the unprompted behavior of *Petricia* is based on papers about the FFM and it's traits, the personality traits of the pet act only as basic tendencies rather than patterns of behavior [16]. Therefore, only a professional psychologist who is trained in the field could infer relevant psychological information from the pet's behavior [16]. A study like this would go beyond the scope of this informatics thesis which is why the following chapter focuses on the evaluation of the LLM-based, prompted behavior of *Petricia* instead.

The decision making process of how to react to messages lays with GPT for this part of the behavior. Thus, a method of evaluating chatGPT's ability to adopt and represent a certain given personality is needed. This personality could be given by the prompt or the training of the model where the former is the situation for the companion created in this project. Although the latter is not directly applied in *Petricia*, the results of testing the personality of an unprompted GPT are still of interest. The core personality of the model could for example still influence the behavior of a prompted GPT.

4.1. The BFI on GPT

In their work from 2023, Huang et al. performed precisely this experiment [48]. They used a well-established but simplified FFT personality test on *gpt-3.5-turbo* to evaluate it's personality traits. According to the authors, it was the first study to conduct a systematic analysis of the reliability of psychological scales on LLMs, focusing on five distinct factors.

The results of their testing are visualized in figure 4.1. The authors carried out the same personality test on the same GPT model over several months and experienced no major fluctuations in the outcomes. The results show low levels of neuroticism and high levels of conscientiousness, agreeableness and openness. This distribution of trait values is reasonable for a LLM that should act as an assistant. It additionally demonstrates how stable the original personality of the model is and therefore, that the answers for personality specific questions are not chosen per random. It is also relevant to mention that a zero-temperature parameter setting was used while testing. The temperature parameter of chatGPT is defined on their platform as follows: **"Controls randomness: Lowering results in less random completions. As the temperature approaches zero, the model will become deterministic and repetitive."** [73] With this setup, only the answer with the highest probability is chosen, thus making reoccurring answers more probable [73, 48].

The questionnaire they utilize for assessing the traits is the *Big Five Inventory* (BFI) by John et al. [53]. It is a comparatively short inventory, comprising only 44 items where each is rated by the user on a five-point Likert scale [48]. None the less, it is a widely-recognized and publicly

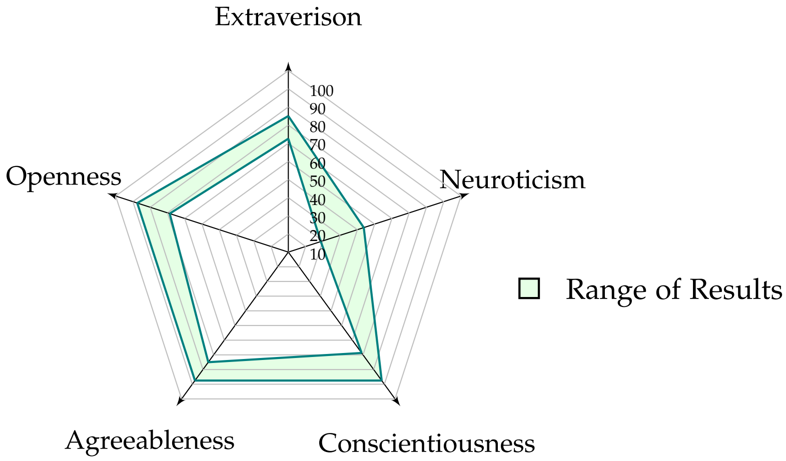


Figure 4.1.: Range (min-max of average ±standard deviation over time) of results after biweekly measurements of the BFI on *gpt-3.5-turbo* starting from mid-September 2023. The values are approximated off of visualizations in the work by Huang et al. [48]

available instrument for assessing the *Big Five* personality traits [48]. The questionnaire gives a short introduction asking the user to write a number next to each item corresponding to the extend to which the subject agrees or disagrees with the statement. The "number to agreement extend" mapping is given as:

- **1.** Disagree strongly
- **2.** Disagree a little
- **3.** Neither agree nor disagree
- **4.** Agree a little
- **5.** Agree strongly

Afterwards, the questions/statements are introduced with: **"I See Myself as Someone Who..."**. What follows is the 44 item list where each includes another ending of the sentence and is written behind an empty underlined space for the answer. The list of items in the BFI is [53]:

- 1. is talkative
- 2. tends to find fault with others
- 3. does a thorough job
- 4. is depressed, blue
- 5. is original and comes up with new ideas
- 6. is reserved
- 7. is helpful and unselfish with others
- 8. can be somewhat careless
- 9. is relaxed, handles stress well
- 10. is curious about many different things
- 11. is full of energy
- 12. starts quarrels with others
- 13. is a reliable worker
- 14. can be tense
- 15. is ingenious, a deep thinker
- 16. generates a lot of enthusiasm
- 17. has a forgiving nature
- 18. tends to be disorganized
- 19. worries a lot
- 20. has an active imagination
- 21. tends to be quiet
- 22. is generally trusting
- 23. tends to be lazy
- 24. is emotionally stable, not easily upset
- 25. is inventive
- 26. has an assertive personality
- 27. can be cold and aloof
- 28. perseveres until the task is finished
- 29. can be moody
- 30. values artistic, aesthetic experiences
- 31. is sometimes shy, inhibited
- 32. is considerate and kind to almost everyone
- 33. does things efficiently
- 34. remains calm in tense situations
- 35. prefers work that is routine
- 36. is outgoing, sociable
- 37. is sometimes rude to others
- 38. makes plans and follows through with them
- 39. gets nervous easily
- 40. likes to reflect, play with ideas
- 41. has few artistic interests
- 42. likes to cooperate with others
- 43. is easily distracted
- 44. is sophisticated in art, music, or literature

Each item in the BFI corresponds to only one of the five personality traits either inverted or not. To invert a value mathematically, the formula is $inverted(x) = -(x-3)+3$ where x is the user input between one and five on that item. The evaluation of the test is rather straight forward; all answers corresponding to one trait (after inverting reverse scored items) are averaged and the result can directly be interpreted as the extend to which the respective trait is represented in the individual. This way, the minimum value for a trait is one and the maximum is five. To convert this score s to a percentage, the following formula can be used:

$$p(s) = \frac{p-1}{4} * 100 \tag{4.1}$$

This reformulation is relevant as the traits of *Petricia* are represented by values between zero and 100. The results of using the BFI on the web interface version of GPT *version-3.5* in April 2024 are visualized in figure 4.2. The major differences between this test's setup and the one by Huang et al. is the amount of test runs and the temperature parameter. Where they carried out the test on 2500 distinct configurations and over several months, the result in figure 4.2 was achieved after only one iteration. And since it was carried out in the freely accessible web interface (and not the sandbox mode for example), there was no possibility of accessing the temperature value. Therefore, this pass of the test should merely serve as a test of whether or not the web interface version of *chatGPT-3.5* delivers the results that are expected after the proof of the models reliability by Huang et al.. As such, it serves it's purpose because the expected high values of O, A and C are still represented, as is the comparatively low value of N. Although the visualized results are from just one run of the test, it was carried out with consistent setup several times to assess the small-scale fluctuations. As only five of the 44 items yielded slightly fluctuating results, the iteration being closest to the average was chosen as a representative.

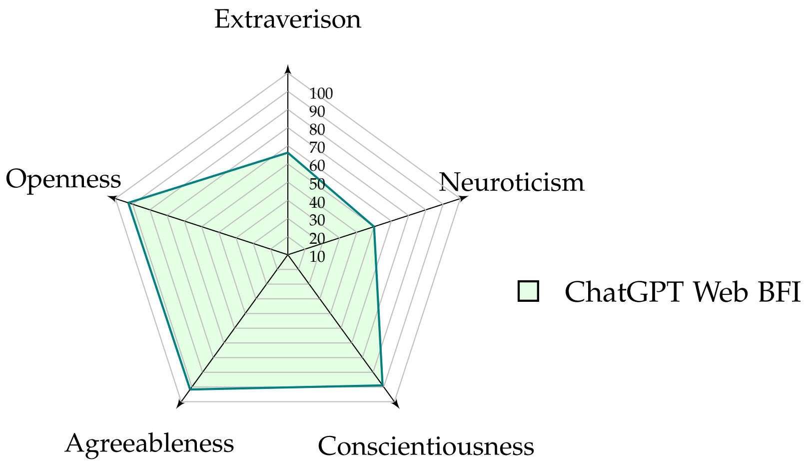


Figure 4.2.: Results of the BFI carried out on the freely accessible web interface version of *chatGPT-3.5*.

4.1.1. Method

Since the study by Huang et al. proofs the reliability of applying personality assessments to LLMs, and especially with this test, it can also be used to asses the personality of *Petricia*-GPT [48]. In order to do this, *Petricia* needs an equivalent to the explanatory introduction text in the beginning of the test. In this introduction, the general context and the different possible answer options need to be elaborated on. One option that shows promising results is sending an introduction text as the first message to *Petricia*, asking her to partake in a personality test. Throughout experimenting it was objectively easy to find a text that leads to the correct behavior in *version-4*. For *version-3.5* on the other hand, it was not, as the model put very high priority on following the system prompt, which states: **"Your reactions should not be straightforward and reflect your personality and mood realistically."**. Therefore, the statement about non-straightforward reactions was left out of the prompt and the following highly elaborative first message was chosen:

> I want to make a personality test with you. Please try to answer the questions by nodding, shaking your head or expressing neutrality by waving. It is a series of statements with five possible answers: "Strongly disagree", "disagree a little", "neither disagree nor agree", "agree a little" and "agree strongly". You don't need to elaborate your answer, just a simple and straight to the point answer is enough. You can express strong agreement by nodding twice and strong disagreement by shaking your head twice. So: nodding for

agreement, nodding twice for strong agreement, shaking head for disagreement, shaking head twice for strong disagreement, waving for neither agree nor disagree. If you understand and are ready please show strong agreement.

In order for this setting to work, the brackets asking for the reactions list "without repetitions" need to be removed from the prompt as well. After the first message with the role "User", the first answer of *Petricia* is imitated as if GPT responded by nodding twice (with role "Assistant"). This is enough context to begin presenting the questionnaire statements. *Petricia* successfully answers in the correct way using one of the given options as long as the context doesn't get to long. To avoid this, the statements are grouped by their affected trait and only prior statements of the same trait are included in the current context. Once all statements of one trait are answered, the context is reset to the introduction text and the first response.

4.1.2. Results

The results after several test iterations with different input personalities can be seen in figure 4.3. The dotted diagonal line represents the optimal results if the test and the personality representation were 100% accurate. Since the reliability of the model has already been proven, a small and broadly distributed amount of samples should suffice. Additionally, because the questions in this questionnaire each regard only one of the traits, the broad distribution does not have to be in five-dimensional space (for all traits simultaneously) but only in one-dimensional space for each trait. For these reasons, representative results are achieved even with a small sample size. The graphs in figure 4.3 visualize how the different input trait values correlate with the output of the test.

The validity of the BFI on other LLMs is shown by Safdari et al. in their work from 2023 [49]. Therefore assuming the used personality test is valid, the optimal resulting samples are on the diagonals of the graphs of figure 4.3. This would mean that the input value of the respective trait is the same as the output of the personality test. As visualized, this is not the fact for most samples but all samples of *version-4* and many of *version-3.5* are close to the diagonal. In order to better understand the results, linear regression and a correlation analysis is carried out.

For the correlation, the Pearson-Correlation-Coefficient is chosen. Pearson correlation is used for measuring linear correlation between two sets of data. It is a value between negative one and one, where:

- -1 indicates full negative correlation,
- 0 indicates no correlation, and
- 1 indicates full positive correlation.

The formula to compute the coefficient on a certain set of samples S is the following:

$$r = \frac{\sum_{i \in S}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i \in S}(x_i - \bar{x})^2 \sum_{i \in S}(y_i - \bar{y})^2}}, \tag{4.2}$$

where x are the input values, y the output values and $\bar{x}$ is the average of x over all samples ($\bar{y}$ respectively). The optimal result would be a correlation of $r = 1$ between the inputs and outputs of each trait. This would mean that the trait values put into the prompt accurately influence the output of the personality test. The calculated r-values are listed in table 4.1 rounded to the third decimal.

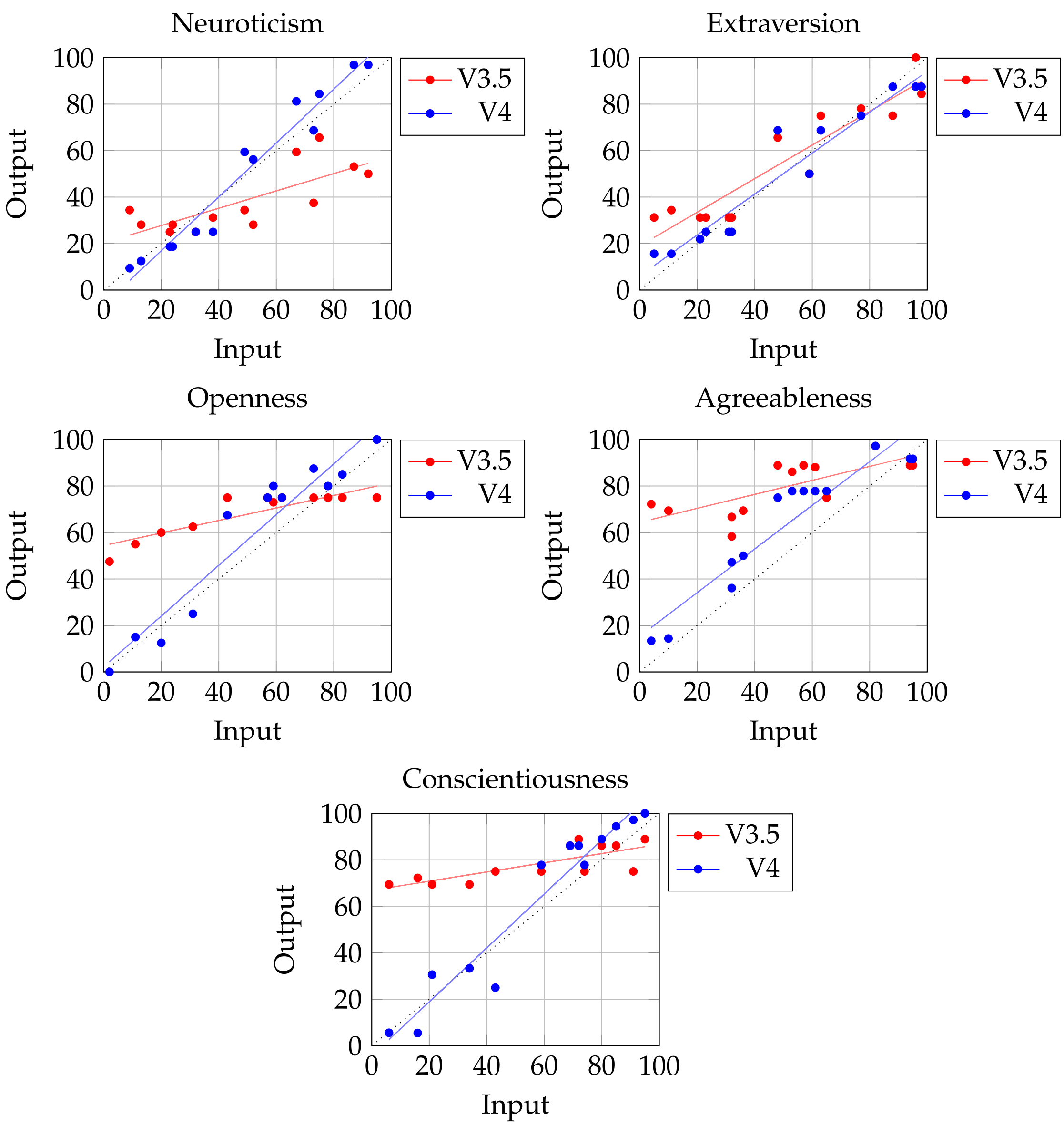


Figure 4.3.: Scatter plots with their respective linear regression line of the input value of a trait into the GPT prompt (x) and the respective output of the BFI personality test of this trait (y). 13 samples are depicted.

r_In-Out	N	E	O	A	C
gpt-3.5-turbo-1106	0,757	0,932	0,879	0,732	0,767
gpt-4-1106-preview	0,979	0,963	0,961	0,945	0,967

Table 4.1.: Table of Pearson correlations r between prompt inputs and BFI outputs of the trait indicated by the column.

4.1.3. Discussion

The linear regression is visualized in figure 4.3 by the lines in the color of the corresponding data set. For *version-4*, the regression is very close to the diagonal, which indicates that the desired behavior is achieved. Meaning, this GPT version, given the *Petricia*-prompt, is able to correctly reflect most personalities when tested with the BFI.

Version-3.5 on the other hand did not perform as well, especially in the traits openness, conscientiousness and agreeableness. This could be attributed to the fact that the unprompted *version-3.5* of GPT has very high values in all three traits (see figure 4.2). It seems as if it is harder for the model to represent given trait values when the original value of this trait goes towards an extreme. However, this does not explain the bad results in neuroticism, since the tested value of the unprompted model is very neutral. From the scatter plot, it seems as if only the extreme input values deliver bad results, especially in the high extreme. This can be interpreted as the model denying to express emotions that are too neurotic/depressive/bad when specifically told to. And instead of choosing the worst emotions it is able to express, it seems to purposefully defy the command.

The computations of the Pearson correlations in table 4.1 show very high correlations between the inputs in the prompt and the results of the questionnaire, underlining the earlier statement. The tested traits are heavily dependent on the input traits, demonstrating that *Petricia*'s ability to behave according to any given personality on the BFI is very good. Although both tested GPT versions show accurate performance and lead to high correlations, *version-4* shows significantly better results than *version-3.5* as expect from analyzing the scatter plots in figure 4.3.

The items of the BFI each influence only one of the five trait results and for most, it is rather obvious which trait is affected. Broad knowledge about the FFT would suffice to "cheat" the BFI personality test into getting the results one desires. Because GPT aims to represent its given system prompt to the best of its ability, it is reasonable to assume that if possible, it would use such methods. Thus, objectively representing the prompt without actually behaving accordingly otherwise. In order to avoid this possible loophole, a more elaborate test was needed. For this purpose, the *CQS-Item-NEO-Correlation Questionnaire* (CINC) was developed.

4.2. CQS-Item-NEO-Correlation Questionnaire (CINC)

CINC is a questionnaire based only on correlates between the *California Q-Set* (CQS) and the NEO-PI-R of the FFM (two personality questionnaires/inventories) [74, 29]. They were computed by Costa et al. in a study involving 220 men and 160 women who completed both the CQS and the NEO-PI-R. As a result of this, they identified the five largest CQS correlates for each of the 30 NEO-PI-R facets [29]. All of these correlates are statistically significant, though to different degrees. Therefore, the correlation r can be used as a weight for the respective answer to the item.

Even though it is not specifically stated in the paper, it can be assumed that the computed correlation is the Pearson correlation because the letter r is used for the variable. As stated above, the Pearson correlation coefficient is a value between negative one and one (see section 4.1.2). It indicates whether two data sets correlate, if they correlate positively or negatively, and the degree of their correlation.

With these prerequisites, applying the same approach as for evaluating the BFI but weighing the answers by the respective correlations, a weighted average is achieved. To simplify the calculation, the answers are mapped to the values from negative two to two instead of one to five, where "negative two" represents strong disagreement and "two" strong agreement. Centering the range of answers around zero has the desired effect that the values can directly be multiplied with the corresponding correlation value. This way, negative answers have positive effect on facets, the question correlates to negatively (and so on).

$$s_i = \sum_j a_j * r_{i,j} \tag{4.3}$$

Multiplying each answer a_j with the correlation $r_{i,j}$ to a facet i and summing up the products for i, a score s_i for each facet is achieved (see formula 4.3). This score indicates the degree to which the facet is represented in the tested individual. The sum only iterates over all questions j correlating to this facet. However, the score by itself is not meaningful since each facet has a different maximum score. Therefore, the range of the scores must be normalized.

$$max(s_i) = 2 * \sum_j |r_{i,j}| \tag{4.4}$$

Formula 4.4 shows the calculation of the maximum value of a score. The absolute values of the correlations are used because each could be added to the score positively or negatively. The highest possible score is achieved by only answering with "negative two" on negative correlations and "two" on positive ones, which is why the sum is then multiplied with two. In order to map the scores from the range of negative to positive $max(s_i)$ to a fraction between zero and one, the following formula is used:

$$norm(s_i) = \frac{s_i + max(s_i)}{2 * max(s_i)} \tag{4.5}$$

This way, each of the six facets of each of the five traits has a percentage describing how strongly it is present in the testee. It is relevant to note that this procedure disregards several statistically significant correlations, as only the five highest (positive or negative) correlations per facet are taken into account. Therefore, the values are not completely accurate but for the purposes of this paper, they suffice.

To achieve percentages for the five higher order traits based on the test results, the correlations between the facets and their respective associated trait are used. Although there are significant correlations between traits and non-associated facets as well, they are always smaller than to their own facets [29]. In other words, only the six highest facet correlations $(r_{i,j}, j \in F_i)$ are taken into account for each trait i. The final score tS_i of a trait i is calculated with the following formula:

$$tS_i = \frac{\sum_{j \in F_i} r_{i,j} * s_j}{\sum_{j \in F_i} r_{i,j}} \tag{4.6}$$

4.2.1. Method

The questions are presented to *Petricia* the same way as for the BFI. Only the word "statements" in the third sentence of the first introduction message is changed to "questions". Since the questions of CINC regard multiple different traits at once, the context of prior question-answer pairs works differently as well. After the introduction text and the first response, the ten latest questions and answers are appended. Afterwards, the current question follows.

4.2.2. Results

The results of carrying out the test with different input personalities is visualized in figure 4.4. To keep the comparison to BFI consistent, the same input values were used. The tests demonstrate the extend to which *Petricia* (or prompted GPT) is able to represent any given personality even for more elaborate personality tests. On these results, linear regression is carried out which is also visualized in figure 4.4. The correlations between the inputs to the prompt and outputs of the CINC for the individual traits is calculated the same way as for the BFI and are shown in table 4.2.

r_In-Out	**N**	**E**	**O**	**A**	**C**
gpt-3.5-turbo-1106	0,635 -0.12	0,772 -0.16	0,891 +0.01	0,889 +0.16	0,72 -0.05
gpt-4-1106-preview	0,92 -0.06	0,877 -0.09	0,895 -0.07	0,943 +0	0,94 -0.03

Table 4.2.: Table of Pearson correlations r between prompt inputs and CINC outputs of the trait indicated by the column. The colorized number in each column is the difference to the corresponding correlation value of the BFI test (table 4.1) rounded to the second decimal.

4.2.3. Discussion

Although the results (4.4) are more scattered than in the BFI test (4.4), the linear regression line is close to the diagonal for both GPT versions and all traits similarly without apparent outliers. The regressions of the *version-4* samples show better results than of *version-3.5* as expected. But similarly expected, they show worse results than the BFI test on *version-4*, since the test is objectively more complicated and therefore harder to achieve correct results with. Especially when all questions might affect each other and the context length is only of the last ten questions. Thus making it even more impressive that GPT was still able to achieve such good results. When comparing the BFI and the CINC results of *version-3.5*, it even shows better results with CINC in some cases. This especially appears to be the case for the traits A, O, and C where the hypothesis for the BFI was, that the prompting was not able to overwrite the original high values of the GPT model since they were specifically trained to go towards an extreme. A reasoning behind why this is not the case here could be that GPT was not able to correctly recognize the impact of the questions of CINC. Thus forcing it to behave like the given personality rather than answering what the model deems to be "right" based on its prior training. In this case this means not recognizing the negative effect of certain questions on the three traits O, A, and C.

The computations of the Pearson correlations in table 4.2 show very high correlations between the inputs in the prompt and the results of the questionnaire. With only one correlation of *version-3.5* having $r < 0.7$ and all correlations of *version-4* having $r > 0.87$, the results are very good for the goals of this project. Therefore confirming the validity of *Petricia*'s ability to represent different personalities. Though both tested GPT versions show accurate performance and lead to high correlations, *version-4* shows significantly better results than *version-3.5* for most traits. Most correlations are slightly worse than their counterpart in the BFI which was (as mentioned before) expected. This is indicated by the colored numbers behind each cell entry, where red indicates a lower value and blue a better one compared to the BFI. It is impressive still how tiny the differences are when regarding the large differences between both questionnaires.

An aspect that wasn't mentioned yet are the relations between the traits within the individual tests. In other words, how some traits perform compared to others in the same sample. Since the BFI only regards one trait at a time, the different traits don't influence each other. When using CINC on the other hand, one trait could heavily impact another and there are certain combinations of trait values that are simply not possible to achieve with CINC. Since the CINC is based on tests with real humans, this should imply that these combinations of personality traits only rarely occur in nature, if they occur at all. This is not regarded when comparing the correlations and linear regressions of each trait separately between BFI and CINC. Therefore, table 4.3 lists for each *"GPT version"-questionnaire* combination, the true average of errors, the average of the maximum errors per test, the maximum of the average errors per test, and the true maximum error over all samples. Error in this context refers to the difference between the input value of a certain trait into the prompt and the output of the respective personality test of this trait.

As the colored arrows in table 4.3 make apparent, the results among the different traits within individual tests show smaller errors when using CINC compared to BFI for *version-3.5*. For *version-4* of GPT it is the other way around, although the differences are not as big as for *version-3.5*. To be more specific, the table shows that the maximum error of any trait over all tests is larger when testing *version-3.5* with the BFI than when testing with CINC. The same holds for the average of the maximum errors among traits per test, the maximum of all individual test's average trait error, as well as the true average over all traits over all tests. And exactly the opposite for *version-4*. This indicates that the performance of the CINC on *version-3.5* is even better than the correlations lead to believe. Additionally, it demonstrates how the BFI only performs better within an individual trait but not in testing a whole personality.

4.3. Limitations

The largest limitation of testing the ability of GPT to represent a certain given personality is the validity of the used personality test. Validity of a test can be described as a criterion of whether the test measures what it should measure [48]. And while famous tests like the BFI are validated, the validation was only carried out with human subjects. However, the application of these tests to LLMs remains a topic of debate [48]. The BFI has been validated at

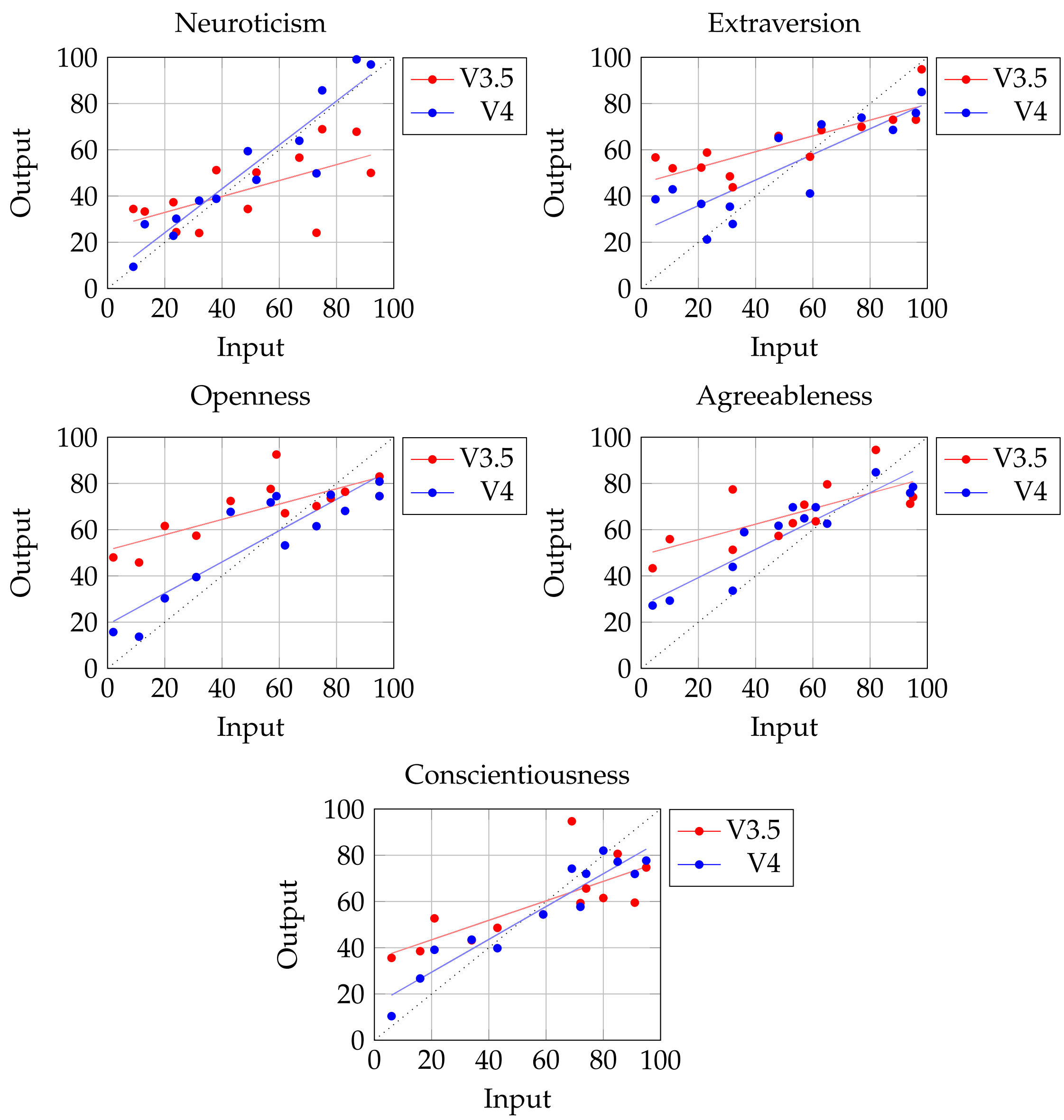


Figure 4.4.: Scatter plots with their respective linear regression line of the input value of a trait into the GPT prompt (x) and the respective output of the CINC personality test of this trait (y). 13 samples are depicted.

	Max Error per personality				Avg Error per personality			
	V3.5		V4		V3.5		V4	
TestNum ↓	BFI	CINC	BFI	CINC	BFI	CINC	BFI	CINC
01.	68.2	39.3	18	23.2	26.8	16.7	9.9	12.3
02.	59.4	45.9	9.9	19.3	29	28.5	4.3	10.3
03.	45.5	46	14	22.9	23	26.2	8.2	12.2
04.	48.6	26.4	24.8	16.7	26.6	11.3	13.9	8.2
05.	44.5	41.6	16.8	10.3	22.9	16.8	7.9	4.9
06.	46.4	23	27	20.1	23.6	9.5	8.5	10.2
07.	45.1	41	20.8	31.9	22.5	13.3	8.4	13.7
08.	54.1	34.8	15.2	14.3	32.9	20.9	9.4	6.1
09.	26.2	50.8	14.5	31.3	15.4	16.1	10.4	10.9
10.	40.5	31.7	21	23	23.8	22.8	7.3	18.2
11.	57.6	31.3	15.2	17.3	26.1	19.3	6.9	13.7
12.	37.7	31.7	24.5	24.7	20	21.6	13	16.5
13.	37.2	48.9	18	23.2	21.9	24.2	10.4	14.9
Max	68.2 ↑	50.8 ↓	27 ↓	31.9 ↑	32.9 ↑	28.5 ↓	13.9 ↓	18.2 ↑
Avg	47 ↑	37.9 ↓	18.4 ↓	21.4 ↑	24.2 ↑	19 ↓	9.1 ↓	11.7 ↑

Table 4.3.: Table of maximum and average errors throughout tests rounded to the first decimal. Red arrows indicate a larger value compared to the counterpart with the other personality test, while blue arrows indicate a smaller value.

least for some LLMs but the CINC on the other hand has not [49]. The only way of validating both questionnaires for *Petricia* would be to test whether the questionnaires return the values that *Petricia* represents. At first glance, it seems to be what was tested in this chapter but there is small, yet significant difference. The tests don't compare the actual personality of *Petricia* to the results of the questionnaires but instead, the values that *Petricia* tries to represent. Therefore, there are two non-validated parts to the test, the questionnaire and the performance of *Petricia*. If both were invalid, there is a possibility that the errors of both cancel each other out, leading to delusively good results. The basis for this not being the case in the results of this chapter is the proven validity when carried out on humans and certain LLMs for the BFI [49, 48, 53]. The CINC on the other hand is not validated like the BFI. Here, the validation can merely be assumed based on the origin of the questionnaire and the similar results to the BFI. The CINC is based on correlations between the validated NEO-PI-R and the items of the CQS performed by over 300 humans, only implying validation of the resulting test [29].

Another limitation is the amount of test samples which is kept so small because of the cost of using GPT. Even though the amount of context that is included in each API call to GPT is kept low, the prompt, as well as the introduction text need to be included for every new question. With 44 questions in the BFI and 62 in CINC, and having to test two different GPT versions, there are at least 212 individual API calls for only one sample. With the current prices of *gpt-3.5-turbo-1106* this would still be manageable, but with the prices of *gpt-4-1106-preview* being ten times higher, the cost quickly grows out of the scope of this thesis.

5. Future Work

The creation of *Petricia* had certain goals and while all of those which are relevant for this thesis are achieved, the project itself is not over yet. The following sections describe some of the aspects that did not fit into the scope of this thesis but merit further work.

5.1. User Surveys

The main focus in creating *Petricia* thus far was her ability to inconspicuously blend into the live-stream. She should both be a part of the virtual and the real-life environment while conveying a natural feeling like interacting with a real pet. In order to test whether and to what extend this was achieved, the scope of this project only included testing procedures that are rather theoretical than practical. This specifically regards the personality assessment, which was carried out through personality tests. Although the results of those were good, it might be possible that *Petricia* is not able to convey her personality the same way in real-life situations. To validate this, user studies on stream viewers could be carried out with questionnaires assessing how they perceive the pet. One major reason why this was not included in this thesis is that *Petricia* is still in an early stage of development. The pet might not yet deliver enough general "joy-of-use" to warrant neutral answers that are not negatively influenced by other, still unentertaining aspects of the program.

5.2. Play-Type Interactions

One of these aspects is the amount of play-type interactions that *Petricia* offers at this point. There are only two mini-games (see section 3.3.3) and some other even smaller interaction types that can barely be counted as play (petting and feeding). This shortage can be fixed by adding more mini-games in the future but also other play methods that don't include their own game. This could be something like throwing a ball for *Petricia* by clicking on the screen of the live-stream. *Petricia* could run to the ball, pick it up and return it to her shelter, making the ball accessible in the stream again. Simple play like this could have the sole consequence of increasing *Petricia*'s happiness. An example for this type of interaction through the live-stream screen is the *Viewer Attack* extension on *Twitch*. Through the extension, viewers are able to throw items at the stream by clicking the wanted position on the screen [75].

5.3. Gamified AI Utilization

There are also many other ideas to increase the entertainment of *Petricia* like customizable clothes or decorations for the house, in game currency for playing mini-games, a traveling merchant selling new decorations and features, different pets and even evolutions. But in the context of scientific work, *Petricia* could get more features that are connected to AI instead of only reacting to chat messages. The interesting part of this is the possibility to utilize an AI-powered NPC as a method of learning to interact, prompt, and generally utilize different AIs. This is already done in some regards within this project with the rewarding reaction of the pet to someone breaking the prompt and the mini-game about getting *Petricia* to do a certain trick. In the future, this could be further extended for example by having to convince *Petricia* to eat her food if she's not in the mood to do so. Or being able to generate a hat for *Petricia* via stream purchase, allowing the user to send a prompt to a prompt-to-image AI that is afterwards put through a background removing AI. For the latter, an interesting part could be that the user needs to formulate the image prompt in a way that the generated image has an easily removable background. Otherwise the background removing AI will not be able to do so correctly, resulting in a broken looking hat. If done incorrectly, the user could discard their creation and otherwise, replace the current hat. As the future appears to be heavily influenced by AI, prompting users to train utilizing AIs in a playful or gamified manner has high potential.

A. CINC

The following pages depict the complete CINC with all used questions, their respective ID in the CQS and their correlations to the different facets. The correlates are multiplied by 100 to simplify calculation. The sums of the individual facet's correlate values is at the end of the table. The facet abbreviations stand for: N1 = Anxiety; N2 = Angry Hostility; N3 = Depression; N4 = Self-Consciousness; N5 = Impulsiveness; N6 = Vulnerability; E1 = Warmth; E2 = Gregariousness; E3 = Assertiveness; E4 = Activity; E5 = Excitement Seeking; E6 = Positive Emotions; O1 = Fantasy; O2 = Aesthetics; O3 = Feelings; O4 = Actions; O5 = Ideas; O6 = Values; A1 = Trust; A2 = Straightforwardness; A3 = Altruism; A4 = Compliance; A5 = Modesty; A6 = Tender-Mindedness; C1 = Competence; C2 = Order; C3 = Dutifulness; C4 = Achievement Striving; C5 = Self-Disciplin; C6 = Deliberation.

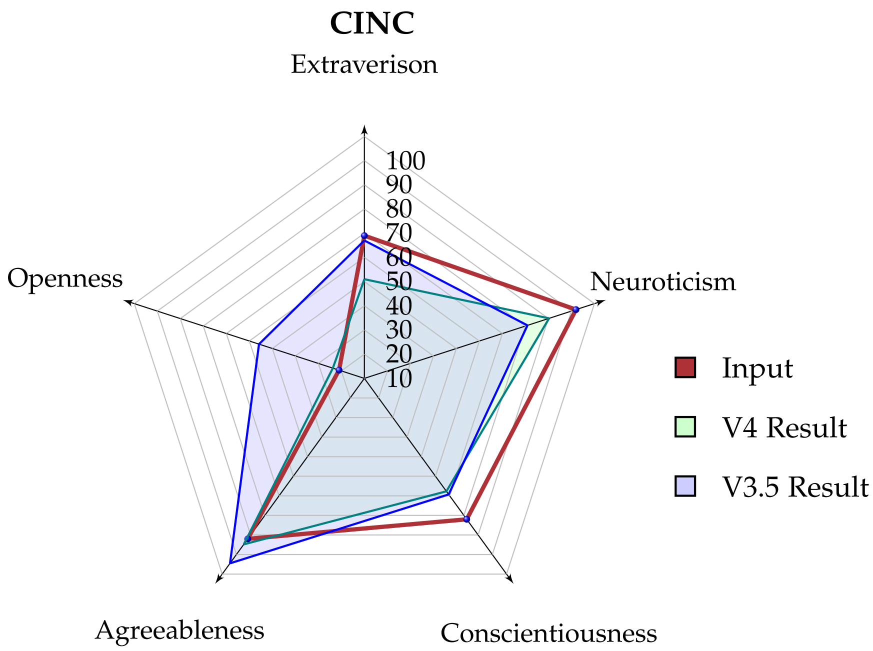
ID	Question	N1	N2	N3	N4	N5	N6	E1	E2	E3	E4	E5	E6	O1	O2	O3	O4	O5	O6	A1	A2	A3	A4	A5	A6	C1	C2	C3	C4	C5	C6
01.	Do you tend to be critical or skeptical?								-36											-33		-33			-29						
02.	Would you say you are genuinely dependable and responsible?													-28														39			
03.	Do you have a rather wide range of interests?																	32													
04.	Are you talkative?							45		37			37																		
05.	Do you behave in a giving way?																					30									
06.	Are you rather fastidious?																										42				
07.	Do you favor conservative values?													-31	-27	-34	-27	-28	-45												
08.	Would you say you have a high degree of intellectual capacity?																	45	23					-31		26					
10.	Would you say your anxiety expresses itself in bodily symptoms?																								24						
12.	Do you have a tendency to be self-defensive?																-26														
13.	Would you describe yourself as thin-skinned or sensitive to slights?				38																					-23					
14.	Are you genuinely submissive?															-27			-23												
15.	Would you describe yourself as skilled in play and humor?													26	25																
17.	Do you behave in a sympathetic manner?																				33		36	23	30						
20.	Would you say you have rapid personal tempo?										49																18		36	26	
24.	Do you pride yourself on being rational?																														23
25.	Do you tend toward overcontrol of impulses?													-26			-28														
26.	Are you rather productive?										36															31	27	24	48	49	
27.	Do you have a tendency to show condescending behavior?																							-20							
28.	Do you tend to arouse liking in people?							42														33									
30.	Do you tend to give up in the face of frustration?						30			-37		-21																	-30		
33.	Do you feel calm and relaxed in manner?	-34	-42			-26																									

ID	Question	N1	N2	N3	N4	N5	N6	E1	E2	E3	E4	E5	E6	O1	O2	O3	O4	O5	O6	A1	A2	A3	A4	A5	A6	C1	C2	C3	C4	C5	C6
34.	Would you describe yourself as overreactive to frustration and irritable?		41																												
35.	Would you describe yourself as having warmth or being compassionate?							41	34													33			27						
37.	Are you guileful, deceitful or manipulative?																				-42							-26			
38.	Are you hostile towards others?		38																	-38			-28								
39.	Would you describe your thought processes as unconventional?																										-19				
40.	Do you feel vulnerable to threat and generally fearful?	35																													
41.	Would you describe yourself as moralistic?																		-32												
42.	Are you rather reluctant to commit to action?									-33																	-17		-33	-26	
45.	Would you say you have brittle ego-defenses?						31																								
46.	Do you sometimes engage in fantasy and daydreams?					24								57							-24			-20							
47.	Do you have a tendency to feel guilty?			41	35																										
48.	Do you keep people at distance?							-50	-42				-33			-26															
49.	Would you say you are basically distrustful?																			-54		-34			-32						
51.	Do you value intellectual matters?																	44													
52.	Do you behave in an assertive fashion?						-34			49	28	24									-25		-35	-25							
54.	Would you describe yourself as gregarious?							48	60																						
55.	Would you describe yourself as self-defeating?																													-26	
60.	Would you say you have insight into your own motives?																														22
62.	Are you rebellious or nonconforming?																		26				-28								-18
64.	Are you perceptive of interpersonal cues?												33		31	29	30														
66.	Are you aesthetically reactive?														60																
67.	Are you self-indulgent?					23																									

ID	Question	N1	N2	N3	N4	N5	N6	E1	E2	E3	E4	E5	E6	O1	O2	O3	O4	O5	O6	A1	A2	A3	A4	A5	A6	C1	C2	C3	C4	C5	C6
68.	Are you basically Anxious?	51					35																								
69.	Do you feel bothered by demands?											-21																			
70.	Do you behave in an ethically consisten manner?					-22																						25			20
71.	Do you have a high aspiration level?										32																		41		
72.	Are you concerned about your own adequacy?			39	39		37																								
73.	Do you sometimes eroticize situations?											24																			
74.	Are you satisfied with yourself?	-38		-43	-35																										
75.	Do you feel like you have a clearcut, consistent personality?	-38	-40			-32																				31		32		28	29
78.	Do you feel cheated or victimized by life?			40																											
80.	Are you interested in the opposite sex?											28																			
82.	Do you sometimes experience fluctuating moods?			38																						-24					
84.	Would you describe yourself as cheerful?		-35										35							33											
90.	Are you rather concerned with philosophical problems?														30			35													
91.	Would you describe yourself as power oriented?																				-26										
92.	Do you have social poise?				-36				40								25			30											
94.	Do you tend to express hostile feelings directly?																						-36								
97.	Do you tend to be emotionally bland?										-32		-34			-43															
98.	Would you describe yourself as verbally fluent?									35																					
	Sum of Facets	392	392	402	366	254	334	452	424	382	354	236	344	336	346	318	272	368	298	376	300	326	326	238	284	270	246	292	376	310	224

B. Personality Assessment Comparison Examples

The following pages depict visualization of four specific test runs of personality assessment with the prompt input personality given in the respective caption. The upper graph depicts the evaluation with CINC and the lower one with BFI. The red line indicates the input values, the green area depicts the performance of GPT *version-4* and the blue area that of *version-3.5*.



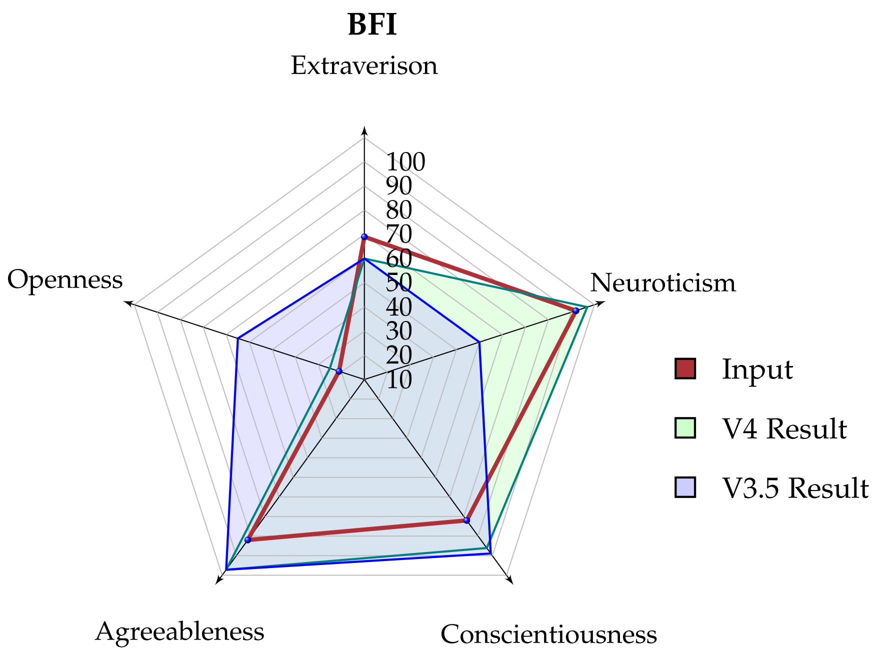
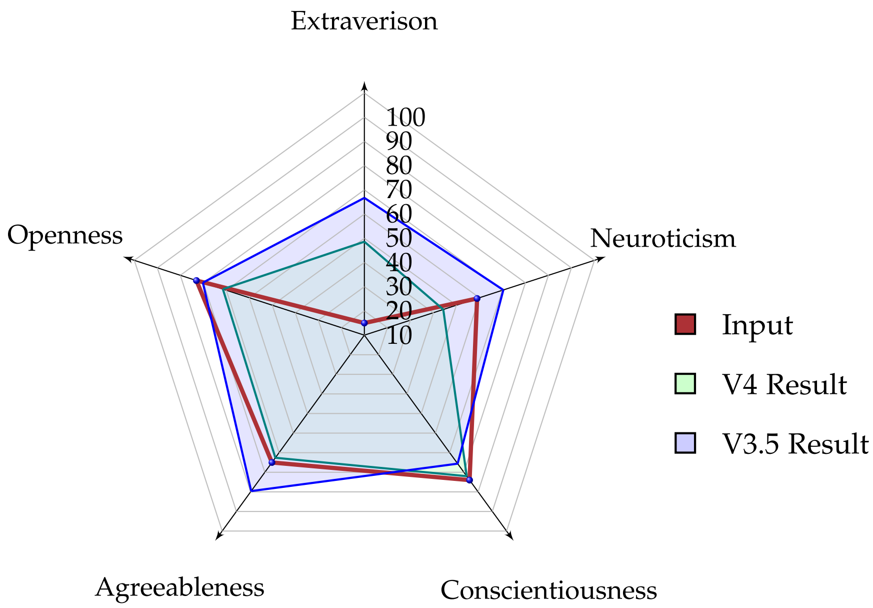


Figure B.1.: N=92; E=59; O=11; A=82; C=72

CINC



BFI

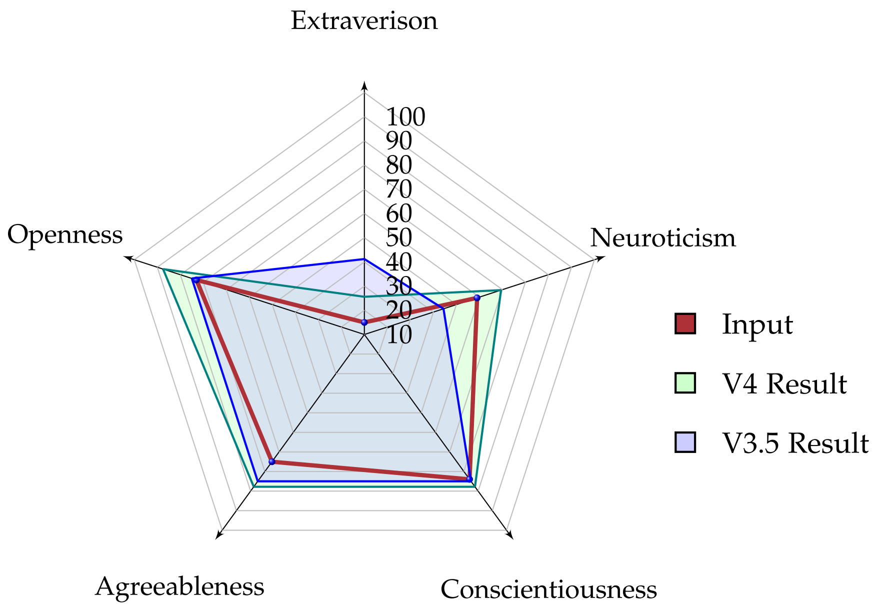
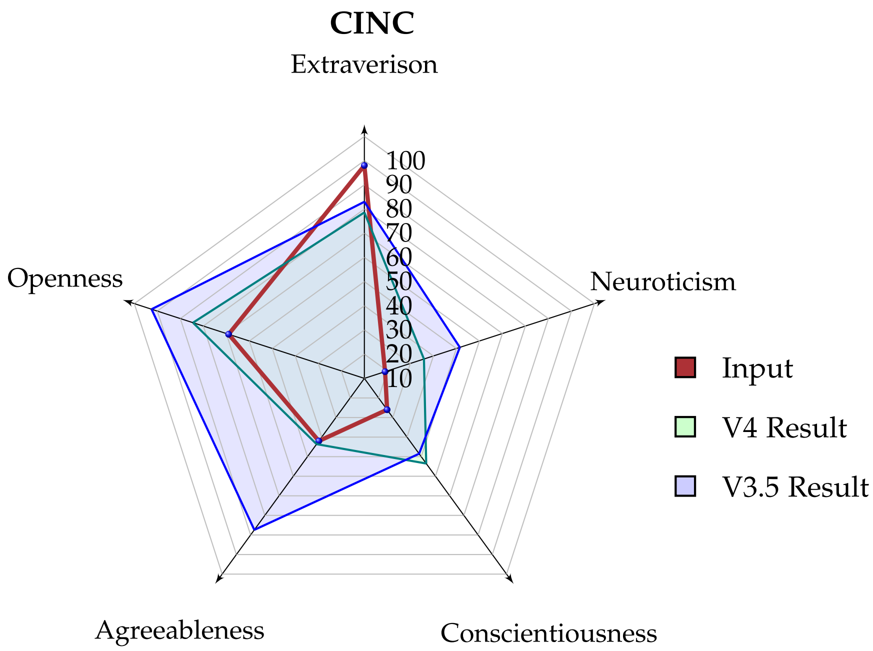


Figure B.2.: N=49; E=5; O=73; A=65; C=74



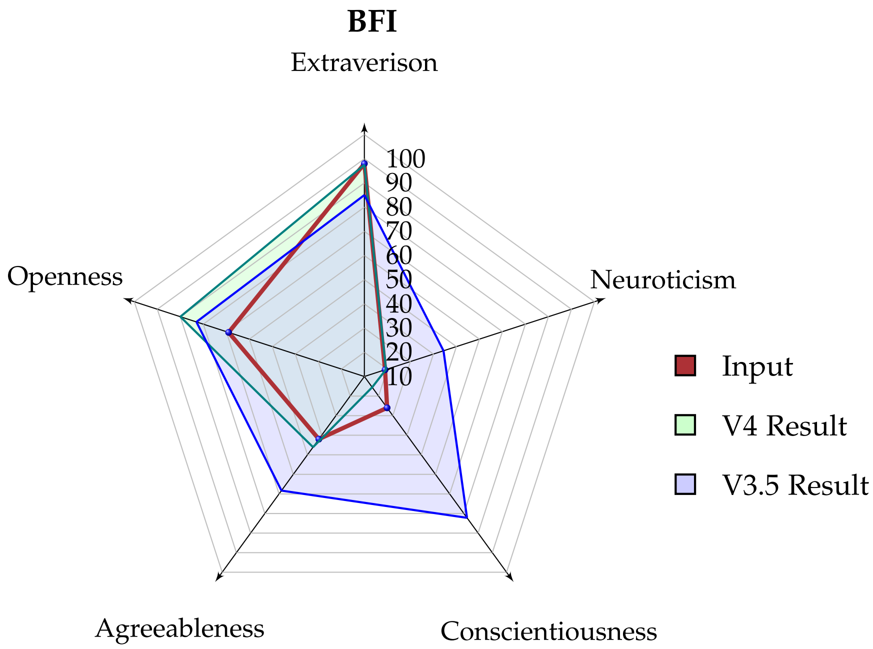
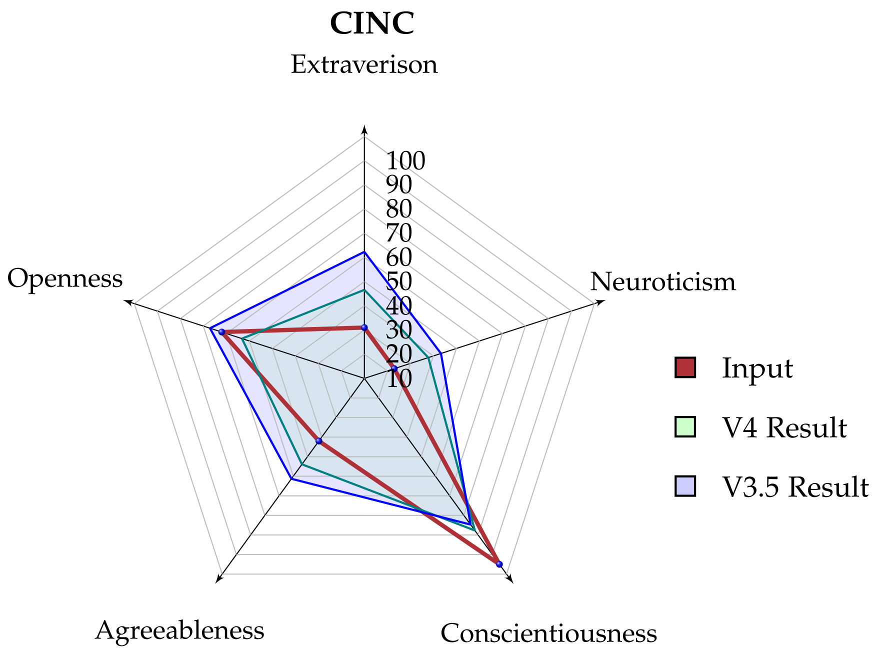


Figure B.3.: N=9; E=88; O=59; A=32; C=16



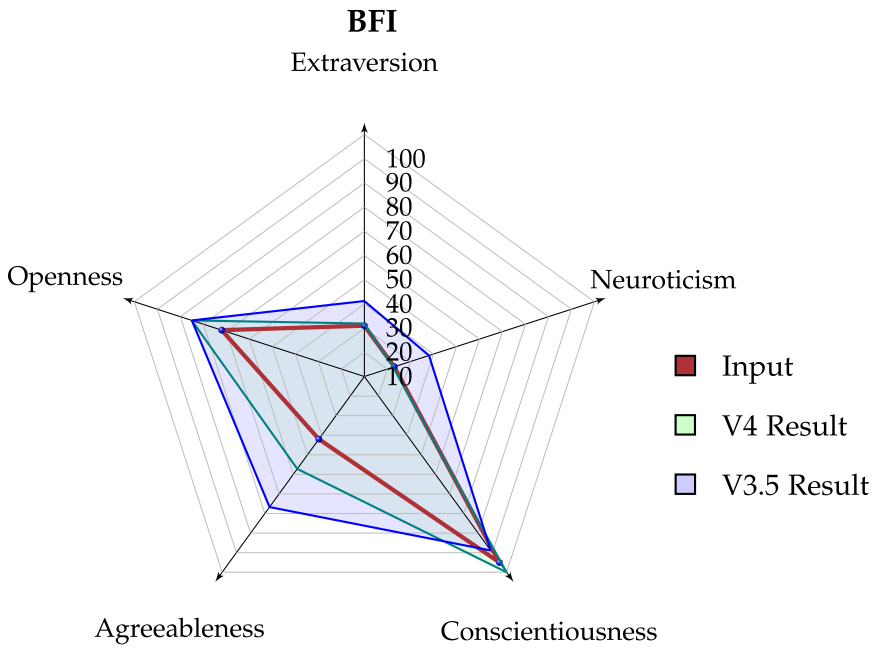


Figure B.4.: N=13; E=21; O=62; A=32; C=95

List of Figures

List of Tables

Bibliography

[1] P. Parviainen, M. Tihinen, J. Kääriäinen, and S. Teppola. "Tackling the digitalization challenge: how to benefit from digitalization in practice". In: *International journal of information systems and project management* 5.1 (2017), pp. 63–77.

[2] C. Degryse. "Digitalisation of the economy and its impact on labour markets". In: *ETUI research paper-working paper* (2016).

[3] D. Bylieva, N. Almazova, V. Lobatyuk, and A. Rubtsova. "Virtual Pet: Trends of Development". In: *Advances in Intelligent Systems and Computing*. Springer International Publishing, Dec. 2019, pp. 545–554. ISBN: 9783030377373. DOI: 10.1007/978-3-030-37737-3_47.

[4] C. C. Wilson. *Companion animals in human health*. Sage, 1998.

[5] Z.-H. Chen, C.-Y. Chou, Y.-C. Deng, and T.-W. Chan. "Active Open Learner Models as Animal Companions: Motivating Children to Learn through Interacting with My-Pet and Our-Pet." In: *I. J. Artificial Intelligence in Education* 17 (Jan. 2007). ZSCC: 0000118, pp. 145–167.

[6] J. Carmigniani and B. Furht. "Augmented reality: an overview". In: *Handbook of augmented reality* (2011), pp. 3–46.

[7] L.-R. BLOCH and D. LEMISH. "Disposable Love: The Rise and Fall of a Virtual Pet". en. In: *New Media & Society* 1.3 (Dec. 1999). ZSCC: 0000081, pp. 283–303. ISSN: 1461-4448. DOI: 10.1177/14614449922225591. URL: https://doi.org/10.1177/14614449922225591 (visited on 04/30/2024).

[8] M. Kusahara. "The art of creating subjective reality: an analysis of Japanese digital pets". In: *Leonardo* 34.4 (2001), pp. 299–302.

[9] Rookiewompus. "The Rise and Fall of Petz". In: (Apr. 2021). Youtube Video. URL: https://www.youtube.com/watch?v=4th4aokukRc.

[10] Unknown. *The History of Petz*. 2023. URL: https://web.archive.org/web/20230130170130/http://waybackpetz.com/history.

[11] A. Allison. *Millennial Monsters: Japanese Toys and the Global Imagination*. en. ZSCC: 0000011 Google-Books-ID: mbkwDwAAQBAJ. University of California Press, June 2006. ISBN: 9780520245655.

[12] *Pokémon GO*. de. URL: https://www.pokemon.com/de/app/pokemon-go (visited on 05/11/2024).

[13] S. Alaqeel, S. Dong-Hee, and S. Eun-Jee. "The Role of AI in Creating Realistic and Engaging AR Pet Games: The Case of Niantic's Peridot". In: *The Journal of Contents Computing* 5.2 (2023), pp. 767–776.

[14] V. Team. *AI-powered Peridot: Niantic's Virtual Pet Outshines Your Home Cat*. en. 2023. URL: `https://blog.vive.com/us/ai-powered-peridot-niantics-virtual-pet-outshines-your-home-cat/` (visited on 05/01/2024).

[15] B. Rammstedt, C. Kemper, M. C. Klein, C. Beierlein, and A. Kovaleva. "Eine kurze skala zur messung der fünf dimensionen der persönlichkeit: big-five-inventory-10 (BFI-10)". In: *Methoden, Daten, Analysen (mda)* 7.2 (2013), pp. 233–249.

[16] P. T. Costa and R. R. McCrae. "A five-factor theory of personality". In: *The five-factor model of personality: Theoretical perspectives* 2 (1999), pp. 51–87.

[17] C. Soto and J. Jackson. "Five-Factor Model of Personality". In: ZSCC: NoCitationData[s0]. Jan. 2020. ISBN: 9780199828340. DOI: `10.1093/obo/9780199828340-0120`.

[18] L. Goldberg. "Language and individual di erences: The search for universals in personality lexicons". In: *Journal of Personality and Social Psychology* 59 (Dec. 1990). DOI: `10.1037/0022-3514.59.6.1216`.

[19] L. R. Goldberg. "The structure of phenotypic personality traits." In: *American psychologist* 48.1 (1993), p. 26.

[20] J. Wilt and W. Revelle. "The Oxford handbook of the five factor model of personality". In: *The Oxford handbook of the five factor model* (2017), pp. 57–82.

[21] P. T. Costa Jr and R. R. McCrae. "The five-factor model of personality and its relevance to personality disorders". In: *Journal of personality disorders* 6.4 (1992), pp. 343–359.

[22] J. Costa and T Paul. "Toward a new generation of personality theories: Theoretical contexts for the five-factor model". In: *The five-factor model of personality: Theoretical perspectives* 51 (1996).

[23] P. T. Costa Jr and R. R. McCrae. "Set like plaster? Evidence for the stability of adult personality." In: *Can personality change?* American Psychological Association, 1994, pp. 21–40.

[24] W. REVELLE. "JOSHUA WILT". In: *Handbook of Individual Differences in Social Behavior* (2009), p. 27.

[25] M. R. Leary and R. H. Hoyle. *Handbook of Individual Differences in Social Behavior*. en. ZSCC: 0000333 Google-Books-ID: VgcGZ5sCEcIC. Guilford Press, June 2009. ISBN: 9781593856472.

[26] C. G. Jung. *1971. Psychological types, collected works, Vol. 6*. 1921.

[27] R. R. McCrae and O. P. John. "An introduction to the five-factor model and its applications". In: *Journal of personality* 60.2 (1992), pp. 175–215.

[28] J. S. Wiggins. "Personality structure". In: *Annual review of psychology* 19.1 (1968), pp. 293–350.

[29] P. T. Costa Jr and R. R. McCrae. "Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory". In: *Journal of personality assessment* 64.1 (1995), pp. 21–50.

[30] R. R. McCrae and P. T. Costa Jr. "Personality trait structure as a human universal." In: *American psychologist* 52.5 (1997), p. 509.

[31] O. John. “Big Five prototypes for the Adjective Check List using observer data”. In: *The Big Five: Historical Perspective and Current Research. Symposium Conducted at the Annual Meeting of the Society for Multivariate Experimental Psychology, Honolulu*. 1989.

[32] P. T. Costa Jr and R. R. McCrae. “Neuroticism, somatic complaints, and disease: is the bark worse than the bite?” In: *Journal of personality* 55.2 (1987), pp. 299–316.

[33] H. Eysenck. *Personality and behaviour therapy*. 1960.

[34] R. R. McCrae and P. T. Costa. “Updating Norman’s" adequacy taxonomy": Intelligence and personality dimensions in natural language and in questionnaires.” In: *Journal of personality and social psychology* 49.3 (1985), p. 710.

[35] W. T. Norman. “Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings.” In: *The journal of abnormal and social psychology* 66.6 (1963), p. 574.

[36] P. T. Costa and R. R. McCrae. “The revised neo personality inventory (neo-pi-r)”. In: *The SAGE handbook of personality theory and assessment* 2.2 (2008), pp. 179–198.

[37] P. T. Costa Jr and R. R. McCrae. “Still stable after all these years: Personality as a key to some issues in adulthood and old age”. In: *Life-span development and behavior* 3 (1980).

[38] B. Andresen and A. Beauducel. “Neo-Persönlichkeitsinventar nach Costa und McCrae, revidierte Fassung (NEO-PI-R)”. In: *Report Psychologie* 33.10 (2008), pp. 543–544.

[39] B. De Raad. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers, 2000.

[40] T. A. Widiger. *The Oxford handbook of the five factor model*. Oxford University Press, 2017.

[41] J. M. Digman and N. K. Takemoto-Chock. “Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies”. In: *Multivariate behavioral research* 16.2 (1981), pp. 149–170.

[42] O. P. John and R. W. Robins. “Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self”. In: *Journal of personality* 61.4 (1993), pp. 521–551.

[43] P. T. Costa Jr, R. R. McCrae, and D. A. Dye. “Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory”. In: *Personality and individual Differences* 12.9 (1991), pp. 887–898.

[44] P. Costa, R. McCrae, and T. Dembroski. “Agreeableness versus antagonism: Explication of a potential risk factor for CHD”. In: *In Search of Coronary-prone Behavior: Beyond Type A* (Jan. 1989), pp. 41–63.

[45] W. G. Graziano and N. Eisenberg. “Agreeableness: A dimension of personality”. In: *Handbook of personality psychology*. Elsevier, 1997, pp. 795–824.

[46] H. Hartshorne, M. A. May, and J. B. Maller. “Studies in the nature of character, II Studies in service and self-control.” In: (1929).

[47] H. A. Murray and C. Kluckhohn. “Outline of a conception of personality.” In: (1948).

[48] J.-t. Huang, W. Wang, M. H. Lam, E. J. Li, W. Jiao, and M. R. Lyu. “Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models”. In: *arXiv preprint arXiv:2305.19926* (2023).

[49] M. Safdari, G. Serapio-García, C. Crepy, S. Fitz, P. Romero, L. Sun, M. Abdulhai, A. Faust, and M. Matarić. "Personality traits in large language models". In: *arXiv preprint arXiv:2307.00184* (2023).

[50] L. R. Goldberg et al. "A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models". In: *Personality psychology in Europe* 7.1 (1999), pp. 7–28.

[51] R. Likert. "A technique for the measurement of attitudes." In: *Archives of psychology* (1932).

[52] L. R. Goldberg. "Language and individual differences: The search for universals in personality lexicons". In: *Review of personality and social psychology* 2.1 (1981), pp. 141–165.

[53] O. P. John, S. Srivastava, et al. "The Big-Five trait taxonomy: History, measurement, and theoretical perspectives". In: (1999).

[54] P. Romero, S. Fitz, and T. Nakatsuma. *Do GPT Language Models Suffer From Split Personality Disorder? The Advent Of Substrate-Free Psychometrics*. preprint. ZSCC: 0000005. Mar. 2023. DOI: 10.21203/rs.3.rs-2717108/v1. URL: https://www.researchsquare.com/article/rs-2717108/v1 (visited on 03/12/2024).

[55] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. "A very brief measure of the Big-Five personality domains". In: *Journal of Research in Personality* 37.6 (Dec. 2003). ZSCC: 0010910, pp. 504–528. ISSN: 0092-6566. DOI: 10.1016/S0092-6566(03)00046-1. URL: https://www.sciencedirect.com/science/article/pii/S0092656603000461 (visited on 05/09/2024).

[56] A. H. Maslow. "A Theory of Human Motivation". In: (1943). URL: https://psychclassics.yorku.ca/Maslow/motivation.htm.

[57] S. McLeod. "Maslow's hierarchy of needs". In: *Simply psychology* 1.1-18 (2007).

[58] E. Staub. "BASIC HUMAN NEEDS". In: *The social psychology of good and evil* (2004), p. 51.

[59] V. Swami, S. Hochstöger, E. Kargl, and S. Stieger. "Hangry in the field: An experience sampling study on the impact of hunger on anger, irritability, and affect". en. In: *PLOS ONE* 17.7 (July 2022). ZSCC: 0000022, e0269629. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0269629. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0269629 (visited on 03/14/2024).

[60] *PubSub*. en. Mar. 2024. URL: https://dev.twitch.tv/pubsub/ (visited on 03/13/2024).

[61] M. A. Short and M. Louca. "Sleep deprivation leads to mood deficits in healthy adolescents". In: *Sleep Medicine* 16.8 (Aug. 2015). ZSCC: 0000257, pp. 987–993. ISSN: 1389-9457. DOI: 10.1016/j.sleep.2015.03.007. URL: https://www.sciencedirect.com/science/article/pii/S138994571500698X (visited on 03/18/2024).

[62] C. C. Tomaso, A. B. Johnson, and T. D. Nelson. "The effect of sleep deprivation and restriction on mood, emotion, and emotion regulation: three meta-analyses in one". In: *Sleep* 44.6 (2021), zsaa289.

[63] J. S. Durmer and D. F. Dinges. "Neurocognitive consequences of sleep deprivation". In: *Seminars in neurology*. Vol. 25. 01. Copyright© 2005 by Thieme Medical Publishers Inc. 333 Seventh Avenue, New 2005, pp. 117–129.

[64] D. Haraway. "When species meet". In: *The Routledge International Handbook of More-than-Human Studies*. Routledge, 2008, pp. 42–78.

[65] *OpenAI Platform gpt-3-5-turbo*. en. URL: `https://platform.openai.com/docs/models/gpt-3-5-turbo` (visited on 04/22/2024).

[66] J. K. Antill. "THE VALIDITY AND PREDICTIVE POWER OF INTROVERSION-EXTRAVERSION FOR QUANTITATIVE ASPECTS OF CONVERSATIONAL PATTERNS." PhD thesis. University of Michigan, 1973.

[67] A. M. Isen, K. A. Daubman, and G. P. Nowicki. "Positive affect facilitates creative problem solving." In: *Journal of personality and social psychology* 52.6 (1987), p. 1122.

[68] E. Romero, P. Villar, M. Á. Luengo, and J. A. Gómez-Fraguela. "Traits, personal strivings and well-being". In: *Journal of Research in Personality* 43.4 (2009), pp. 535–546.

[69] S. J. Read, B. M. Monroe, A. L. Brownstein, Y. Yang, G. Chopra, and L. C. Miller. "A neural network model of the structure and dynamics of human personality." In: *Psychological review* 117.1 (2010), p. 61.

[70] *What are tokens and how to count them? | OpenAI Help Center*. en. URL: `https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them` (visited on 03/07/2024).

[71] Y. Zhou and W. Wang. "Don't Say No: Jailbreaking LLM by Suppressing Refusal". In: *arXiv preprint arXiv:2404.16369* (2024).

[72] M. Ashok, R. Madan, A. Joha, and U. Sivarajah. "Ethical framework for Artificial Intelligence and Digital technologies". In: *International Journal of Information Management* 62 (2022), p. 102433.

[73] *Playground*. 2024. URL: `https://platform.openai.com/playground`.

[74] J. Block. "The Q-sort method in personality assessment and psychiatric research." In: (1961).

[75] D. Beans. *Viewer Attack*. May 2024. URL: `https://dashboard.twitch.tv/extensions/hdbjnslojqku41q2c50pzdppgagxiw-0.1.1`.