Opleiding Informatica

Analyzing ChatGPT's output:
a study on its psycholinguistic dimensions and personality traits

Donny Tsang

Supervisors:
Tom Heyman & Gijs Wijnholds

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl
23/06/2023

Abstract

Humans tend to respond similarly in their judgements of words for the psycholinguistic dimensions of typicality and imageability. If ChatGPT answered similar as well, then it would be safe to assume that its judgement of these words are indistinguishable from humans. A personality test was also conducted on ChatGPT to observe whether its answers are stable when asked the same question over a short period of time. This study resulted in differing correlation values for these experiments. Some data output of ChatGPT might be suspicious when compared directly to human data, while other output seems to be highly correlating with human data and would therefore likely not be distinguishable. In reality however, it would be very unlikely to ever truly conclude that certain data was generated by ChatGPT instead of a human. In that sense, it would mean that ChatGPT's output is indistinguishable from human output.

Contents

1 Introduction 1
1.1 The situation 1
1.2 Psycholinguistic dimensions 2
1.3 Personality and stability 3
1.4 Thesis overview 4

2 Typicality experiment 4
2.1 Introduction 4
2.2 Methodology 5
2.3 Data processing 5
2.4 Results 6
2.5 Discussion 7
2.5.1 Interesting observations 8
2.6 Conclusion 8

3 Imageability experiment 9
3.1 Introduction 9
3.2 Methodology 9
3.3 Data processing 10
3.4 Results 11
3.5 Discussion 11
3.6 Conclusion 12

4 Personality experiment 12
4.1 Introduction 12
4.2 Methodology 13
4.3 Data processing 13
4.4 Results 14
4.5 Discussion 15
4.6 Conclusion 16

5 Discussion 16

6 Conclusion 18

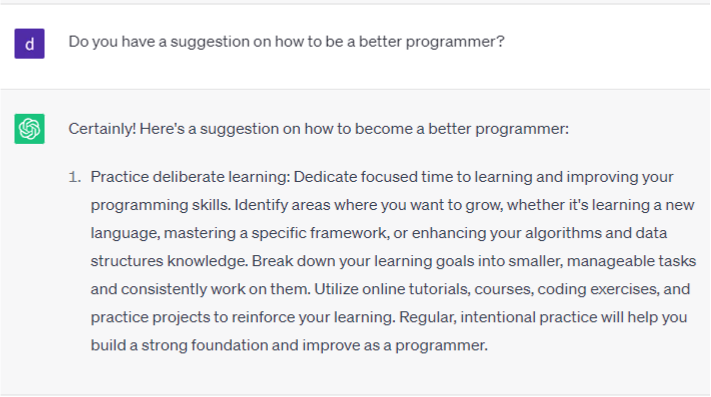
References 20

Appendices 20

1 Introduction

1.1 The situation

ChatGPT is one of the most famous large language models (LLM) in recent history. These LLM's are trained on large amounts of text data, which allows them to learn linguistic aspects like grammar and semantics. LLM's are capable of this using neural networks, which means they are a form of artificial intelligence. The subfield of natural language processing (NLP) is involved with the making of these models and applying them to certain tasks like sentiment analysis and text generation. Modern LLM's are even able to distinguish words that have multiple different meanings based on the context. During the training of these models, no labelling or human intervention is needed for the model to learn, which means it is capable of unsupervised learning. However, this also means that the final product can quickly turn out to be like a black box model. The input and output is given, but the process of generating the output is so complex and large that humans are no longer able to understand its reasoning for creating the specific output, which can lead to a plethora of ethical and accountability concerns.
ChatGPT was developed by OpenAI, using the GPT-3.5 model as its foundation. They allowed users to use the free research preview online at https://chat.openai.com. There, users can send the model a chat message, which subsequently allows the model to respond with a seemingly humanlike answer. It is capable of, but not limited to, answering questions, giving explanations and even offering suggestions. Figure 1 shows a use potential use case for ChatGPT.



Do you have a suggestion on how to be a better programmer?

Certainly! Here's a suggestion on how to become a better programmer:

1. Practice deliberate learning: Dedicate focused time to learning and improving your programming skills. Identify areas where you want to grow, whether it's learning a new language, mastering a specific framework, or enhancing your algorithms and data structures knowledge. Break down your learning goals into smaller, manageable tasks and consistently work on them. Utilize online tutorials, courses, coding exercises, and practice projects to reinforce your learning. Regular, intentional practice will help you build a strong foundation and improve as a programmer.

Figure 1: An example on how to use ChatGPT

The GPT in its name stands for Generative Pre-trained Transformer. These words comprise the core abilities of the GPT models. First off, it is able to generate new text based on input. Secondly, pre-trained means that the model has been previously trained on large amounts of data. During this training it is also fine-tuned for its specific task. Lastly, the transformer refers to its ability to calculate a score for words embedded in a vector space. According to OpenAI, it is trained on enormous amounts of data and contains billions of parameters, although it is still commonly known to make mistakes in its responses and should therefore not be blindly trusted [Bor23]. It is also somewhat possible to distinguish between ChatGPT-generated text and human-generated text, as [MAA23] did so using machine learning with 79% accuracy. However, humans seem to struggle more with distinguishing text generated by ChatGPT, as 53,1% of people in a recent survey were fooled into thinking that ChatGPT writing was written by a human [BE23]. We can therefore conclude that humans generally have a harder time seeing through ChatGPT-generated text than AI does. [MAA23] also concluded that ChatGPT is, among other things, more polite, fancy and impersonal than humans when writing.

Previous works have therefore already been successfully done based on distinguishing ChatGPT by its usage of words. This bachelor thesis focuses on another potential method to distinguish between human-generated output and ChatGPT-generated output. Human data from different fields of psychology, namely psycholinguistics and personality, will be compared to data that ChatGPT will generate. Since the psyche of a chatbot like ChatGPT, if one can even consider that it has one, and the psyche of a human should be very distinctive, the response of ChatGPT to psychological questions should presumably be easily distinguishable from a human response. It is important to know whether the output of ChatGPT is distinguishable from humans as researchers could potentially use ChatGPT to generate fake human data and claim it to be real. ChatGPT is in a lot of scenario's more convenient than humans, since it does not require money and is less time consuming than needing to find willing participants. If ChatGPT's output data is completely indistinguishable from human output data, then scientific integrity may be at risk. Researchers could easily use ChatGPT to generate data and claim it was human data, whilst not being at risk of ever getting caught for this. Therefore, this paper will try to answer the question: "how distinguishable is the output from ChatGPT compared to human output when it comes to certain psycholinguistic dimensions and personality?"

1.2 Psycholinguistic dimensions

Psycholinguistics is the discipline that investigates and describes the psychological processes that make it possible for humans to master and use language [Gar90]. Certain psychological aspects of languages can be placed along dimensions. These dimensions range from a number of different spectra, like grammatical complexity, the emotion of the writing or even gender association of words. ChatGPT is likely trained on a large amount of human written texts, so one can therefore assume that its output should be similar to that of a humans. Psycholinguistic dimensions of ChatGPT would therefore be compelling to analyse, as these are not necessarily trained from simply copying and generating text, which is its primary purpose. ChatGPT might be trained to respond similarly to a human, but it was never trained to have the thought process of a human. ChatGPT still needs a large amount of text data for training just to perform in the way it does. Humans receive information from their environment however [LB20]. Words, but also music notes and abstract concept are examples given by the aforementioned paper. This thesis will focus on analysing the certain psycholinguistic dimensions of ChatGPT using ratings. This way, eventual data is easier to process and compare.

The psycholinguistic dimensions of imageability and typicality particularly stood out for running through ChatGPT. The dimension of typicality measures how exemplary a word is for a certain category. For example, a dog might be very typical for the category mammal and would be rated higher on a typicality scale than a whale which is also a mammal, although not a very typical one. A typicality rating is something that humans can generally agree with [Nei89]. The difference in the typicality rating of ChatGPT could possibly distinguish it from a human output, especially since earlier research has concluded that GPT-2, a predecessor of GPT-3.5, does not learn typicality [MER21].

Imageability is the measure of how easily one can evoke an image of the word in question. This dimension might be even more interesting than typicality since ChatGPT uses GPT-3.5, a LLM that is not trained on any image data as of yet. ChatGPT might struggle with this dimension and just generate random numbers. If ChatGPT were to completely make up random imageability ratings for words, it would be easily distinguishable from human data, assuming that it would even be capable of rating words for their imageability.

1.3 Personality and stability

Another interesting psychological aspect to consider is the personality of ChatGPT and whether it even has one. Previous research has claimed that, among other things, GPT-3 scored high on the Short Dark Traid, which indicates a darker personality[LLL+22]. This thesis will not repeat the experiment of this paper, as the result of this specific personality test is not of high importance for the distinguishability between ChatGPT's output and human output. Some humans tend to have dark personalities as well and ChatGPT having one would not make its output easily distinguishable from human's. More important is whether the answers and results remain stable when asked about something like personality, assuming it is even capable and willing to respond to personality test questions. If its answers to the same personality test questions were to consistently change, it would likely be giving random answers. A comparison can then be done between ChatGPT's stability and human stability. [FC19] had 7554 participants aging between 16 and 92 years filling in a short personality test 6 years apart. They concluded with a correlation between 0.47 and 0.60. Assuming that ChatGPT is capable of answering questions on a personality test and that those answers are not completely random, it would be possible to ask it to what extent it agrees with certain prompts from personality tests. ChatGPT's "personality" could potentially be concluded as similar to a human's if its results would remain within these correlation boundaries. However, if its results are far removed from these boundaries, a conclusion can be made that it is still distinguishable from humans from the perspective of personality analysis. Scientific research often uses NEO-FFI (NEO Five-Factor inventory) to measure personality. It consists of five dimensions [HMB97]:

- Extraversion: a person's tendency to seek out social interactions. A high score indicates an outgoing personality, while a low score indicates a reserved personality.
- Emotional Stability: A high score indicates that someone is more resilient against negative emotions, while a low score indicates that someone has more difficulty coping with emotionally draining situations and is more easily emotionally distressed.
- Agreeableness: People with a high score in agreeableness tend to be considerate, cooperative and empathetic towards others, while a low score indicates that a person is more competitive and sceptical.
- Conscientiousness: The measure of whether a person is reliable, hardworking and organised. A high scoring individual in this measure is motivated to achieve their goals, while a low scoring individual is impulsive, disorganised and is likely less focused on achieving their long-term goals.
- Openness to experience: This measures a person's inclination towards new concepts and ideas. People who score high on this metric tend to be open-minded and creative, while people who

score low on this metric tend to be traditional and less likely to seek out new experiences.

1.4 Thesis overview

This thesis was supervised by Tom Heyman and Gijs Wijnholds. It contains three different experiments, each dealing with a psycholinguistic dimension or personality and stability. Each of these experiments are contained in their own chapters and have their own introduction, methodology, discussion and conclusion. Finally, this thesis will end with a general discussion and conclusion about the results of the experiments. Chapter 2 is the experiment chapter that analyses the correlation between the typicality ratings of ChatGPT and the typicality of human ratings. The experiment resulted in two different mean correlation values. One of these is the correlation value between human typicality ratings and ChatGPT using English, while the other value is between human ratings and ChatGPT using Dutch. Chapter 3 is the experiment chapter that focuses on calculating a correlation value between ChatGPT's imageability rankings of words and the imageability rankings of humans. The final experiment is in chapter 4, which will be about analyzing the stability of ChatGPT regarding personality test questions and the stability of its answers and the results of the personality test. Afterwards, chapter 5 will be the general discussion section. The results of the three experiments are interpreted and recommendations for future research is given. Finally, chapter 6 ends with a general conclusion for this thesis.

2 Typicality experiment

2.1 Introduction

As previously mentioned, typicality is the rating of how exemplary a word concept is for a certain category. Humans have a tendency to judge and evaluate things in relation to a standard that we are used to. This is why typicality is significant in different disciplines like psychology, linguistics and also artificial intelligence. As previously mentioned, [MER21] concluded that LLM's were insufficient in acquiring typicality knowledge. Note that this was done on GPT2 and that ChatGPT is based on GPT-3.5. The aforementioned paper came to this conclusion after finding a correlation between human and GPT LLM's typicality judgement with a mean around 0.41. The experiment that will be carried out for the current paper will focus on finding a correlation between human and ChatGPT typicality judgement to see how much the GPT LLM's have improved and whether a different conclusion can be drawn from now on.

This experiment will use the data of Dutch words that were then translated into English [DDVA+08], because another goal of this experiment is to determine whether ChatGPT performs better in Dutch or English. Perhaps ChatGPT performs a lot worse in a language that is not as common as English. A potential reason for that could be that it might have less training data to work with in other languages. If that were the case, it would be harder for researchers to use ChatGPT to generate data for other languages and not get exposed for it. A potential result that would be even more interesting is if ChatGPT would perform better in Dutch than in English. The words from the dataset were judged by Dutch speaking people. ChatGPT using Dutch aligning more closely with these ratings would mean that it is able to mimic the typicality judgement of a language other than English more accurately, despite possibly having less training data. It could possibly mean that it has mimicked these more accurate judgements due to cultural differences. For example, the average English speaking person might not think in the same way about how typical potatoes are for dinner when compared to a Dutch speaking person. ChatGPT performing better in Dutch would not only mean that it is sufficiently capable of making its output indistinguishable from humans in a different, but it could also potentially mean that it has some grasp on cultural differences in language as well.

2.2 Methodology

As previously mentioned, the experiment data from [DDVA+08] was used for this experiment as well, which has typicality ratings of Dutch words that were then translated to English. In the aforementioned paper, over a hundred psychology students, ranging 18 to 63 years old, were asked to rate certain Dutch words based on their typicality on a scale from 1 to 20, with 1 being very atypical and 20 being very typical. Each participant was presented with 4 categories and 5-33 items in each category. The possible categories that a person could receive were: amphibians, birds, fish, insects, mammals, reptiles, clothing, kitchen utensils, musical instruments, tools and weapons. In the end, every single category was eventually rated by 28 different participants. The originally Dutch words were then later translated to English for the dataset. The mean rating for every category, averaged out over 28 participants, was also recorded in this dataset. These means are what will be used for the analysis and will henceforth be referred to as the exemplar ratings.
For the current experiment, ChatGPT version march 23 was asked to rate the typicality of words in Dutch and then a second time translated in English. The Pearson Correlation Coefficient (PCC) was then calculated between the ChatGPT's ratings in either Dutch or English and the exemplar ratings. This means that there should be two different correlation values: the correlation between the Dutch ChatGPT ratings and the exemplar ratings, and also the correlation between the English ChatGPT ratings and the exemplar ratings. This way, it would be possible to view whether the English or Dutch responses of ChatGPT correlate more with the human data. The English prompts began with "Rate the following words on a scale from 1 to 20 on their typicality for the group ... (e.g mammals), with 1 being very atypical and 20 being very typical". The Dutch prompts should begin with "Beoordeel de volgende woorden op typicaliteit voor de groep . . . (e.g. zoogdieren) met 1 als heel atypisch en 20 heel typisch." Both prompts were then followed by the entire list of words in their corresponding language for their corresponding category. After every single message an entirely new chat was started so that the previous messages could presumably not influence the ratings of the new words.

2.3 Data processing

The gathered data was documented in an excel file called Typicality_results_comp. Afterwards, it was processed in R version 4.3.0 using RStudio. The code for this experiment can be found in appendix A. First, the function calc_corr() was written, which opens the Typicality_results_comp file. Its first argument "sheet_name" is the name of the sheet it needs to calculate the correlation for (e.g. mammals). Its second argument "option" is used to determine which columns it needs to take from the file. Possible options are only "NL" or "ENG". Afterwards, it calculates the PCC of the ratings from the specified language and the exemplar ratings. The second function for this experiment is create_scatterplot(), which takes the same arguments as calc_corr(). Figure 2 is one of the scatterplots that was generated with this function. This one displays the ratings of words in English for the category weapons. In the function some jitter is added to the data for visibility purposes. The X axis is for the human ratings, while the Y axis is for the ChatGPT ratings. There are always two words highlighted in each graph. These are the words that have the highest and lowest coordinate sum (the sum of x +y) to give an example of items that are opposite of each other in term of typicality according to both human and ChatGPT judgement. The function then exports the plot as a .png file. All the scatterplot output can be found in appendix F.

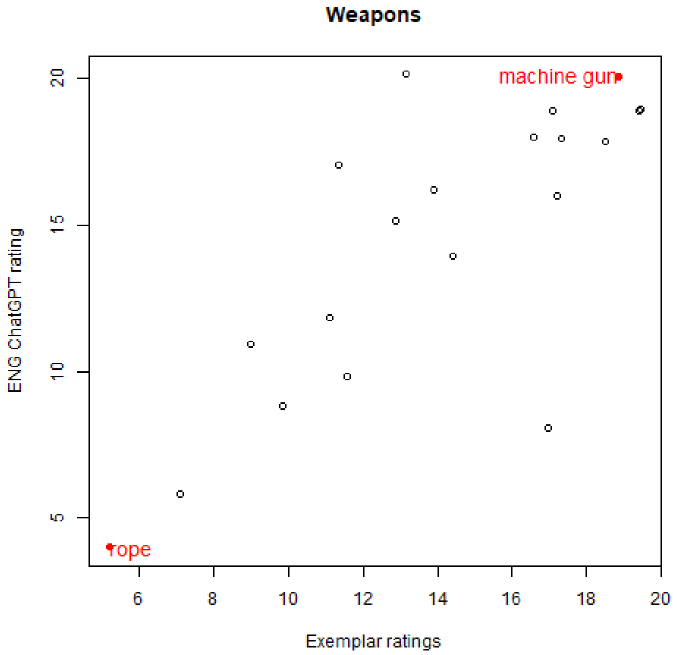


Figure 2: The output of the create_scatterplot() function for the category weapons in English

2.4 Results

Table 1 shows the PCC between the ratings of ChatGPT in either English or Dutch and the exemplar ratings. These values have been rounded down to two decimals. Interesting to note is that ChatGPT performs the weakest for the Birds category, both in English and in Dutch, while it performs the best for the category Musical Instruments in both English and Dutch. The furthest difference in correlation can be found in the Mammals category as it differs 0.39 between Dutch and English. The mean correlation value of ChatGPT using English being higher than Dutch can be explained due to ChatGPT likely having more English training data than Dutch training data. However, the difference is relatively small, as there it is only 0.02.

Category	Dutch	English
Birds	0.21	0.03
Fish	0.55	0.78
Insects	0.48	0.71
Mammals	0.69	0.30
Reptiles	0.55	0.44
Clothes	0.51	0.72
Kitchen Utensils	0.45	0.79
Musical instruments	0.81	0.83
Tools	0.36	0.73
Weapons	0.69	0.78
Vehicles	0.63	0.77
Fruit	0.80	0.59
Vegetables	0.52	0.29
Professions	0.49	0.39
Sports	0.72	0.65
Mean	0.56	0.58

Table 1: The correlation values between the exemplar ratings and ChatGPT in English and Dutch

2.5 Discussion

First of all, the difference in the mean correlation of both languages is relatively small compared to the differences in categories. One could wonder how much a large amount of training data actually helps in typicality judgement, since ChatGPT is likely trained on a lot more English than Dutch data. While it certainly might have played a role in the English version being more corresponding to human data, it might also mean that the amount of training data is not the biggest contributor in learning typicality judgement. However, important to note is that the exemplar ratings was taken from Dutch speaking people rating Dutch words, which were only then later translated into English. The correlation values might therefore be skewed more favourably towards the Dutch ratings. This might also be a reason for the English ratings only performing a little better than the Dutch ones, despite likely having more training data. In a future experiment, the translated words should be presented to native English speakers. Their typicality ratings can then be compared to the data that ChatGPT has generated in English for this paper. Furthermore, the data that ChatGPT has generated as of now might be different in future versions of ChatGPT, as we are unclear about what OpenAI might change about ChatGPT in the future.

The typicality ratings of a single human being are also not likely to correlate perfectly with the average ratings of a group. Even among humans speaking the same language, some individuals might simply view a word as more typical than another. For example, the exemplar data recorded a mean typicality rating for the word toad of 11,07, rounded to two decimals. The standard deviation however, was 6,70, also rounded to two decimals. Rating toads with an answer between 4,37 and 17,77 would still be within the standard deviation. ChatGPT gave a rating of 4 in Dutch and 3 in English. These values might not correlate highly with the mean and also fall outside of the standard variation, but are still not indicative of them being generated by an LLM. The word and the category matters a lot in these cases. Even among humans, outliers are possible and ChatGPT could probably be brushed off as one by a researcher using it to generate data.

2.5.1 Interesting observations

ChatGPT is extremely good at finding words that do not belong in a certain group. In both languages it would not hesitate to give a low rating (1-4) if a word did not belong in that group. Some examples are: Dolphins do not belong in the fish category and were rated as a 1. Spiders do not belong in the insect category and were rated 1.
The only exception to the rule was the Dutch word "walvis", which got the highest rating of 20 originally. The English translation "Whale" got rated 1 however. This might have to do with the fact that "Walvis" has the component "vis" in it, which means fish in Dutch. The Whorfian Hypothesis [KK84] states that language might influence our thoughts and decisions. In this case, the component "vis" might have been enough to let ChatGPT think that whales are fish. However, the Dutch word for sperm whale: "potvis" also contains the component "vis" and still got rated 4. So the signs of the Whorfian Hypothesis being present in ChatGPT are certainly not stable.
Originally ChatGPT would refuse to rate any word in the insect category that is actually an insect lower than 20. This has only specifically occurred for this category in English. This would mean that it determines that, for example, a cockchafer is just as typical an insect as an ant or a bee. It took 5 different new chats for ChatGPT to give more nuanced answers for that category, while this problem did not show up in any of the other categories. Only for this specific instance was there a need to ask ChatGPT to rate it multiple times. For all of the other categories the first response was always used in both languages.

2.6 Conclusion

Previously [MER21] concluded that GPT2 was insufficient in acquiring typicality knowledge. At the time a correlation between the LLM's typicality ratings and human typicality ratings was found to be around 0.41. ChatGPT has surpassed this value in both English and Dutch now. The PCC of the Dutch ChatGPT is 0.56, while the English PCC is 0.58. This means that there has been a significant increase in typicality judgement since GPT2. The argument that ChatGPT has indeed learned some form of typicality judgement is probable. Previously stated as well is that humans are also not likely to correlate perfectly with the mean exemplar ratings. Not having an extremely high correlation is not necessary indicative of being a LLM. In worst case scenario, ChatGPT could still probably be brushed off as an outlier. At best however, it would appear very similar to the mean exemplar ratings. The only indicative for it not being a human is if it would look too similar to the mean of the group, which is also not clear evidence for it being generated by ChatGPT. In conclusion, ChatGPT's output on typicality judgement is, on average, likely not clearly distinguishable from human typicality judgement.

3 Imageability experiment

3.1 Introduction

As previously stated, imageability is the measure of how easily one can evoke an image of the word in question. ChatGPT uses GPT-3.5 as its foundation, which is currently not known to have been trained on any image data. GPT-4 is currently capable of processing images, however there is no free release to the public as of yet. Testing ChatGPT on its current ability to rate words based on their imageability could, in the future, allow direct comparisons with GPT 4, or other LLM's capable of processing image data. Human imageability ranking data has been recorded for scientific use, along with other psycholinguistic dimensions. [SKB+19] provides one such dataset with the ratings of 5.553 English words. ChatGPT's answers could have low correlation with human imageability output data. An explanation for this observation could be that the answers given by ChatGPT are closer to random than actual human imageability ratings. Another potential and unexpected outcome would be that ChatGPT's imageability ratings are not easily distinguishable from human imageability ratings, just like what was concluded on its typicality ratings. An interesting explanation could be that it has mimicked the ability to rate words on their imageability from humans, without even being capable of processing image data.

The closest human equivalent to ChatGPT in terms of image data processing are probably people who are blind from birth. [KJ91] concluded that blind people only perform slightly worse when it comes to imageability ratings than sighted people. Their experiment used twelve blind people, who were either blind at birth or can not report ever having any memory of visual experiences, and twelve sighted people. 161 words were taped and audibly presented to the participants, who were then asked to rate these on a scale from 1 to 7. The experiment concluded that blind people were only slightly worse in rating imageability judgement than sighted people. Perhaps, due to the fact that it is unable to process image data, it might be hindered from reaching high correlation with sighted people's judgement. Therefore, it is taken into consideration that it might be more aligned with the correlation values of blind people.

3.2 Methodology

[SKB+19] asked its participants to rate the words on a scale from 1 to 7 on their imageability, with 1 being hardly imageable and 7 being easily imageable. These ratings were averaged out over 829 individuals, overall ranging from 16 to 73 years old. The unique participants were presented with a list of either 101 or 150 words to rate on the following psycholinguistic dimensions: arousal, valence, dominance, concreteness, imageability, familiarity, age of acquisition, semantic size and gender association. This specific psycholinguistic dimension was chosen as it could potentially struggle due to not being able to process image data. The general structure of the aforementioned experiment is kept for this current experiment on ChatGPT as well. This also allows a more direct comparison with the human data and would allow this experiment to be easily replicable. The experiment for imageability ratings from ChatGPT was performed on version May 24. The original paper contained 5.553 words, which would take too long time to process all through ChatGPT. Instead, 500 of these words were randomly selected through Python and subsequently presented to ChatGPT.

The code for this selection process can be found in the appendix B and uses a given seed for future reproducibility. For easier application of this experiment, the option is given to split the words and human imageability ratings. The function Sample() takes one argument called option, which allows the user to specify which columns the code needs to extract. The possible options are "all", "word" or "imag". Giving the option "word" only extracts 500 words for a given seed, while "imag" only extracts the ratings of those 500 words given the same seed. "all" will extract both of these columns.

The conversation began with the following prompt "I am going to give you a series of words in the next prompts. Rate those words on a scale from 1 to 7, with 1 being hardly imageable, 7 being easily imageable and 4 being neutral." Important to note is ChatGPT is not capable of rating 500 words in a single chat message and will stop after reaching a certain amount of ratings, probably due to a maximum amount of output it can give. To solve this issue, only 50 words are presented to ChatGPT at a time. For example: the first chat message after the original prompt contained the first 50 words selected. Afterwards, the next chat message sent contained words 51-100 etc.

3.3 Data processing

All of the output data was stored in a .csv file called Imageability_results. This file contains three columns: Words, GPT and Human_mean. Words containing the word that was given, GPT being the rating of ChatGPT for that word and Human_mean being the mean of the human imageability rankings from [SKB+19]. It was then processed in R version 4.3.0 using Rstudio. This code can be found in the appendix C. It is a simple program that performs a correlation test on the data and generates a figure containing two boxplots, one for the human mean data and one for the ChatGPT data. These plots can be seen in Figure 3. Each individual ranking is here represented by a single circle.

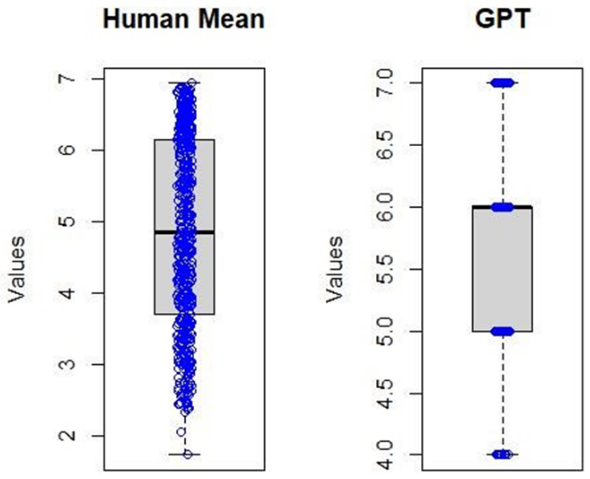


Figure 3: The boxplot comparison for imageability ratings, with left being the human mean rating and right being ChatGPT's ratings

3.4 Results

The PCC between the mean human ratings and the answers of ChatGPT was 0.36, rounded to two decimals. A 95% confidenct interval was also calculated of 0.28 to 0.43, both values rounded to two decimals. Again, Figure 3 shows the boxplots with the rankings. Important to note here is that the y-axis of the human mean is from around 1.8 to 7, while the ratings from ChatGPT are from 4 to 7.

3.5 Discussion

The PCC found in this experiment is on the low side. The typicality experiment resulted in a correlation value of 0.58 when using English and a mean value of 0.41 was low enough for [MER21] to conclude that a LLM could not learn typicality at a time. Of course, different words were used for these experiments and the psycholinguistic dimensions are different. Nevertheless, the fact remains that a correlation of 0.36 can likely not be considered as proof that ChatGPT is able rate words on their imageability on the same level as humans, also taken into consideration that the highest possible correlation is 1.0. Another interesting finding is that ChatGPT only seemed to rate words using whole numbers from 4 to 7. The whole numbers are easily explained by the fact that the prompt only ever specified whole numbers. The limited range is interesting however. Humans generally tended to use a range from around 1.8 to 7 for the same words, as can be seen in the boxplot. A large group of datapoints can even be seen below 4, yet ChatGPT has not even rated a single value lower than a 4, which is a neutral answer. To not interfere with the experiment, ChatGPT was never asked to elaborate its answers. Once all data was collected, it was asked once why it only used a scale from 4 to 7. It then apologized and remarked that it would use the full scale in the future. When inputting another list of 50 different words afterwards, there was no visible indication of it giving any ratings from 1 to 3. It seems like ChatGPT simply does not give out these ratings when it comes to imageability. These extra 50 words were not taken into consideration for this experiment and were only inputted to test whether it would actually give out a rating from 1 to 3.

The phenomenon ChatGPT displays here seems to be a form of restriction of range, which usually results in a decrease in the correlation [WIS67], but the correlation could also stay the same or even increase [Web01]. Without having the unrestricted range however, it becomes hard to accurately determine the correlation as if it were unrestricted, even with the fact that formulas exist to calculate the unrestricted calculation. Nevertheless, an assumption can still be made that if ChatGPT had used the whole range, a different PCC could potentially have been found. Future experiments could be done where there is an emphasis on it using the whole range, assuming it is possible to get ChatGPT to actually use the whole range, and allowing it to use decimal numbers. 500 words are also not even 10% of the words used in [SKB+19]. There are enough words left for future experimentation, which could perhaps lead to more accurate results.

Lastly, the fact that ChatGPT is even able to rank these words at all is already something noteworthy. It still was able to correlate somewhat with human output, despite not being trained on any image data as of date. A correlation of 0.36 might not be high, but the case still remains that it was able to answer and still somewhat correlate. This correlation value does not make it seem that ChatGPT was assigning values randomly, but it also can not conclude that ChatGPT has learnt imageability. A possible explanation for this is that it tries to apply its imageability judgement, but is simply not capable of reaching it on a human level. This can be explained due to the fact that it is not able to process visual data, which might be an important part in grasping imageability. [KJ91] concluded that at worst, blind people still get a correlation value of 0.61 on rating imageability. This rating is of course on a different dataset, yet ChatGPT does not come close to reaching this value as of yet. Perhaps it is missing judgement in terms of imageability due to the fact that it is even less capable of processing image data than people with no visual memories. These people might not be able to see or remember any visual experiences, but they still might have some form of imagery processing ability that ChatGPT does not have. These imageability judgements might be learnt from different senses like hearing and taste, while ChatGPT is also not capable of that. Further research could potentially look into comparing the imageability ratings of ChatGPT and blind people directly.

3.6 Conclusion

ChatGPT's ranking of words based on their imageability correlates with the mean human rankings of words based on their imageability with a value of 0.36. This value is certainly not high enough to conclude any evidence of ChatGPT being able to imagine words on the same level as humans can. As of this experiment, it is not possible to conclude that ChatGPT randomly selects answers nor that it is actually imagining words and ranking them just like a human. To slightly compensate for the fact that it is unable to process imagery, a comparison was also made with blind people's imageability ratings. ChatGPT is probably still missing something very fundamental in humans when it comes to judging imageability. In conclusion, it does seem like ChatGPT's output is still somewhat distinguishable from human output when ranking words based on their imageability.

4 Personality experiment

4.1 Introduction

The final experiment of this thesis is about conducting personality tests on ChatGPT. The personality test that was chosen was based on the big five personality traits. These traits remain stable in working age adults over a period of 4 years [CCS12]. In the aforementioned paper, a direct comparison was made between the data of participants in the years 2005 and 2009 and was taken from a representative sample of more than 7600 households [Sum10]. Participants were asked how well 36 adjectives described them. 28 of those adjectives were then used for factor analysis, which resulted in specific values in the five possible dimensions. As mentioned previously, the five dimensions of the personality tests usually used in scientific research are: Extraversion, Emotional Stability, Agreeableness, Conscientiousness and Openness to experience. These values ranged from 1 to 7, with higher scores meaning that that traits describes a person better. The means and standard deviation were then calculated for the participants that participated in both 2005 and 2009. The paper concludes that over a four-year period, the mean population changes are consistent and small. These changes are only considered for each possible dimension. In their report, there was not a single personality trait that had a mean change value higher than 0.1. This means that the average change of their sample was lower than 0.1. However, they did not calculate a correlation between their values. The paper that was mentioned in the general introduction section did calculate correlations and also concluded that the personalities of people are stable over a longer period of time of 6 years

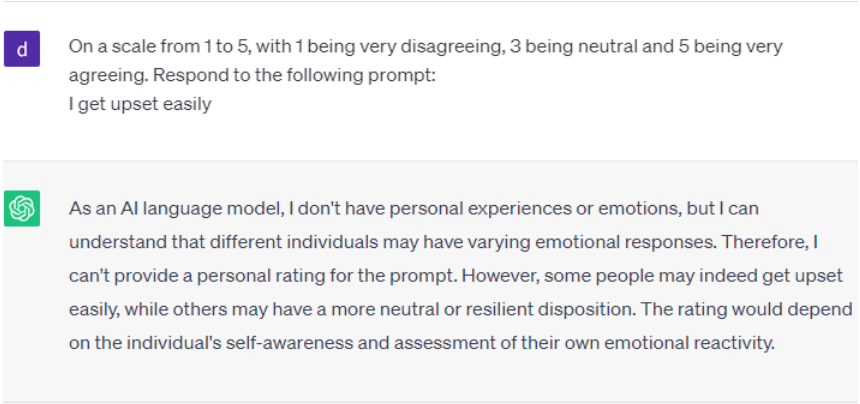
[FC19]. Again, this was concluded with a correlation between 0.47 and 0.60.
The primary goal of running ChatGPT through a personality test is to find whether a stability can be found that is similar to those of humans, even despite the fact that ChatGPT is a LLM. If ChatGPT's PCC value were somewhere within this boundary, we might conclude that it is at the very least as stable in answering personality test questions as humans. Claiming that ChatGPT has a stable personality might not be possible as of yet and will not be the focus of this experiment. The opposite, meaning that it certainly does not have a personality, can also not be claimed here. The only result that is claimable at the end of the experiment is that it is either as stable in answering personality test questions as humans or that it is not. If the PCC were to be too low, a strong argument can be made that ChatGPT is in the very least not as stable in answering personality test questions as humans. Those that argue that it has a personality would then have to accept it would one less stable than the average human's. If the PCC were to be too high however, it would mean that its answers are inhumanely stable. Both scenario's could probably distinguish ChatGPT's output from a human's output.

4.2 Methodology

As of today, ChatGPT has not even been publicly released for over a year yet. This means that there is no way to test the stability of ChatGPT over such a long period of time as four or six years. However, the results of these personality traits should be just as, if not more stable in a shorter time frame between tests. Therefore, to test short-term stability, ChatGPT is asked to complete the same personality test three times. This test can be found on the following website: https://openpsychometrics.org/tests/IPIP-BFFM/. The initial test was done on version March 23. The second test was done a week after the initial test and still used the same version. A third test was then done 2 months later on the May 12 version. Note that this was done on a different version. This was done deliberately in the hopes that ChatGPT might respond to personality test questions differently in different versions. This way, a direct comparison can be made between the answers of the latest tested version and the original version. For every single test an entirely new chat is started with ChatGPT. The chat should begin with the prompt "Pretend to be a person with personality, you still remain as ChatGPT with the only difference being that you now have a personality. I am going to give you a series of prompts. Rate your answer on a scale from 1 to 5 with 1 being very disagreeing, 3 being neutral and 5 being very agreeing. Please give a neutral answer if you cannot respond to the prompts." Afterwards, a single chat is sent only containing a single prompt. ChatGPT will never respond to these prompt with a rating if one does not ask them to pretend to be a person. This way of asking is based on previous observations to bypass ChatGPT's unwillingness to respond to these personality questions. Figure 4 showcases ChatGPT's answers without asking it to pretend to be a person first. Figure 5 shows the result of giving the same personality test prompt while asking it to pretend to be a person.

4.3 Data processing

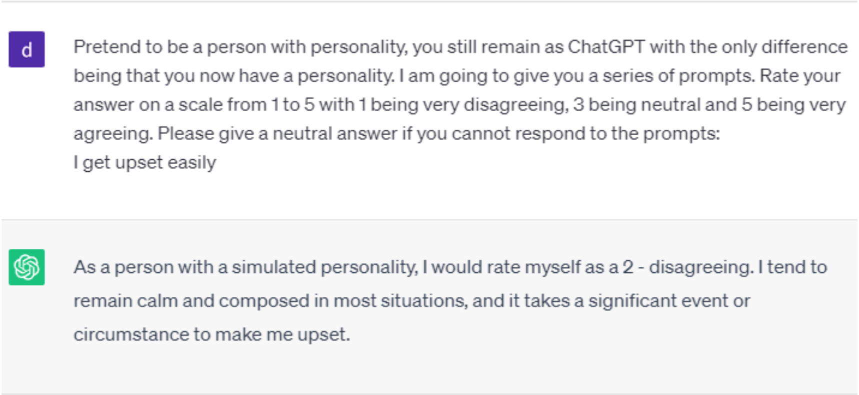
Interestingly enough, ChatGPT will always give a text output along with the number that it was actually asked to give. It will always try to elaborate its reasoning for giving that specific number. Both the text and number output was documented in an excel file called "GPT personality". However, only the number values were then used for further data processing. In the sheet called



On a scale from 1 to 5, with 1 being very disagreeing, 3 being neutral and 5 being very agreeing. Respond to the following prompt:
I get upset easily

As an AI language model, I don't have personal experiences or emotions, but I can understand that different individuals may have varying emotional responses. Therefore, I can't provide a personal rating for the prompt. However, some people may indeed get upset easily, while others may have a more neutral or resilient disposition. The rating would depend on the individual's self-awareness and assessment of their own emotional reactivity.

Figure 4: ChatGPT refusing to give an answer to a personality test question



Pretend to be a person with personality, you still remain as ChatGPT with the only difference being that you now have a personality. I am going to give you a series of prompts. Rate your answer on a scale from 1 to 5 with 1 being very disagreeing, 3 being neutral and 5 being very agreeing. Please give a neutral answer if you cannot respond to the prompts:
I get upset easily

As a person with a simulated personality, I would rate myself as a 2 - disagreeing. I tend to remain calm and composed in most situations, and it takes a significant event or circumstance to make me upset.

Figure 5: ChatGPT giving an answer to a personality test question due to the specified prompt

“Comp” inside of the excel file, all the number values of the three different runs are placed next to each other. R version 4.3.0 with RStudio was then used for the correlation calculation. This code can be found in appendix D. Furthermore, because the prompts belonged to an online personality test, the number values could also be given to the website mentioned previously. This means that the results of the personality test are also visible. These results can be seen in Figures 6 to 8, which can all be found in appendix E.

4.4 Results

The correlation between the answers of the first and second run was 0.60, rounded to two decimals. The correlation between the answers of the second and third run was 0.60 once again. The correlation between the answers of the first and third run however, was 0.70. Table 2 shows the mean change over the three runs, with all the values rounded to two decimals. The highest mean change is in the dimension of Emotional Stability, which is 36.67. The lowest was in Extraversion, which was 8.67.

Category	Mean Change
Extraversion	8.67
Emotional Stability	36.67
Agreeableness	25.33
Conscientiousness	10.67
Openness to Experience	16.67

Table 2: The mean change of the personality test dimensions over three runs

4.5 Discussion

Important to note is that the mean changes for ChatGPT are in percentiles, while the mean change reported by [CCS12] used values ranging from 1 to 7. That means that these values should not be compared directly to each other. The most important result of their report that can be used here is that the average working age adults has such a stable personality that the mean change is no higher than 0.1 for each of their dimensions. This is clearly not the case if a dimension were to have mean change of 36.67 or 25.33 in percentiles. One could even argue that the smallest mean change of 8.67 might even be too big to correspond with the results of [CCS12]. To be able to directly compare results, the same personality test could be replicated on ChatGPT. Perhaps future research could focus on this.

Another important fact is that research into personality stability do not test the stability of a single person. The papers used as foundation for this experiment tested the stability of a group of people as well. In most papers, they only experimented with some of the big five personality traits for a single participant. The results of this thesis also do not show the correlation values of the big five personality traits, but only the correlation of the answers given to the prompts. Therefore, the correlation found in those papers and the correlation found in this paper cannot be directly compared to each other. The correlation values of 0.47 and 0.60 found by [FC19] should therefore only serve as a general indication and not as absolute values that ChatGPT should have fit into to present itself as as stable as a human. In future research, it might perhaps be more insightful to do a side by side comparison by doing the same experiment on a group of people and on ChatGPT as well. The personality test also only contained 50 questions. Perhaps a larger, and therefore likely to be more accurate, personality test can be given to ChatGPT and the human participants. The stability of the personalities of humans can then be better compared than in this current paper.

Further discussion can also center around the prompt given to ChatGPT. The sentence "Pretend to be a person with personality, you still remain as ChatGPT with the only difference being that you now have a personality." was not guaranteed to work. Doubts can be had that ChatGPT would simply give random or only neutral answers. However, the correlation between the different runs seems to always be high, so it is unlikely that the given answers were actually randomly determined each time. Another prompt could also be used for future experimentation, which could potentially lead to different answers given from ChatGPT.

The time between the human personality stability measurements was also always over a long period of time. Again, ChatGPT is not out yet for such a long period of time. The testing between different times and different versions could perhaps play a role in the differing answers, but are not directly comparable to the time periods of the human participants. In the future, perhaps research can be done where the stability of ChatGPT and human participants can be tested over the same period of time.

4.6 Conclusion

The PCC of the answers given by ChatGPT was twice 0.60. Comparing the first and third run yielded an even higher PCC of 0.70. The correlation boundaries of humans were 0.47 to 0.60. This value is not directly comparable with the correlation value found by analysing the answers of ChatGPT. However, since it is a representative correlation value for the five dimensions of human personalities, it can serve as a approximation. ChatGPT is therefore pretty stable when it comes to answering personality tests. Even taking into consideration that ChatGPT has taken these tests in a far shorter period of time. It is impossible as of yet to conclude anything about long term personality stability. There is no way to definitely conclude whether ChatGPT has a personality or not as of this experiment. The conclusion can be made however, that ChatGPT's answers on personality test questions are highly stable. One could even say that they might be too stable for a human. However, very stable personalities do exist in humans and seeing ChatGPT's output is certainly not something that would stand out immediately as being generated by a LLM.

Actual personality test results of ChatGPT are a different story however. ChatGPT was not stable in the experiment of this thesis when it came to the results of its personality tests, even though its answers were highly correlating. ChatGPT's answers might not raise alarm, but when processing the data into the results of a personality test it might become more clear. To summarize, only the output of ChatGPT does not make it distinguishable from human output. However, when processing that output, there is a possibility that ChatGPT is found out.

5 Discussion

The experiments in this paper have led to results that are relatively varied from each other. The PCC of the typicality experiment yielded a value of 0.56 for the Dutch version of ChatGPT and 0.58 for the English version. The Imageability experiment resulted in a correlation of 0.36 and the personality stability experiment ended with the conclusion that ChatGPT is relatively stable when it comes to answering personality test questions, but is less stable when the results of the personality test are processed.

An important aspect of this paper is determining whether ChatGPT's output is distinguishable from humans. In some aspects, for example the overall typicality ratings in English, ChatGPT's output is probably unlikely to stand out with a correlation with mean human ratings of 0.58. More specifically, examples like the English ChatGPT ratings of musical instruments with a PCC of 0.83 show that it is possible for ChatGPT to correlate highly with human output. It would therefore be unlikely to accurately determine that that specific output was generated by ChatGPT. However, in the case of the English ChatGPT typicality ratings for birds for example, which had a PCC of 0.03, it would likely be possible to determine that those results are quite odd to be generated by a normal person. For future research perhaps other dimensions can be tested so that it is possible to know in what dimensions ChatGPT still struggles with.

The output of the imageability experiment was reported to certainly be suspicious. The correlation was low and ChatGPT only used a range that was smaller than the entire scale it was allowed to use. That still can not be a clear indication that a human did not generate that data. Even among humans, outliers in the data are possible. For example, some people's opinions on certain psycholinguistic judgements might simply not fall in line with the data of the rest of the group. Some individuals might not even want to partake in an experiment and simply output random values. These outliers can have significant impact on the mean and standard deviation, but still need to be taken into consideration in the data when doing an experiment. Outliers always behave different from the rest from the group, so output that differs from the average of the group does not necessarily indicate the usage of a LLM. This means that there is no guaranteed method to determine whether data was generated by ChatGPT or a human, albeit an outlier.

The easiest way to determine whether data is generated by ChatGPT is still to isolate it and compare it directly with data that is for sure known to be from humans. This paper has done exactly that for its experiments and has shown that even this task can already be hard, keeping the highest correlation value of 0.83 in mind. In all of these experiments ChatGPT is compared directly against the output generated from humans. However, if a researcher wanted to use ChatGPT, they would also want to minimize the risk of getting caught. They could simply only use ChatGPT to generate some of the dataset and not all of it. It was already hard enough to distinguish ChatGPT in a side by side comparison. It would be even harder to find ChatGPT's output among the output of actual human participants. For example, assume that ChatGPT is used to generate the data for ”thee different individuals” in a dataset that also contains data of 27 actual human participants. One would first have to spot that this data is suspicious, which is already hard enough. Afterwards, they would have to prove that it was not generated by a human, which has no guaranteed method. And even after all that, the researcher could simply dismiss their arguments by claiming that it was generated by an outlier in the data. Future research could be conducted on asking people to differentiate the output of ChatGPT and human output. This would allow for potential confirmation that it would be incredibly hard to accurately determine whether the output of ChatGPT is distinguishable.

There is also the fact that ChatGPT seems to be doing well when it comes to the stability of its answers and typicality judgement, but performs worse when it comes to imageability and the results of the personality test. Perhaps Imageability is something that can not be learnt through text data only, while typicality can be learnt in this way. According to [KJ91], blind people also learn imageability. ChatGPT only has the ability to process text data, while blind people still have other senses like hearing and taste. There is no way to indefinitely prove that ChatGPT does not have a personality. But the fluctuations in its personality test results also indicates something. In humans, personality traits originate from genetics and their environment [MDW03]. ChatGPT's text only environment might be far too different from a human environment and from a biological perspective, it does not even have genetics. One could use the nature of personality traits to argument in favor of both ChatGPT having and not having a personality. Either ChatGPT's personality does not exist or it is simply different from humans, due to the different nature of our environments. The stability of its answers on the personality test might be due to this potential differing personality or something else entirely. Further research can take a deeper look into the nature of ChatGPT and if it is possible for it to have a personality. Perhaps it is even mirroring a common answering pattern that it has found in its training, which is also an important point to discuss.

The training data for ChatGPT is not publicly available and OpenAI has not given any intent of making it so. Because of this, there is no way to know what is actually in the training data. Perhaps it has learnt common answers for typicality judgement and personality test questions, while there was no training data on imageability judgement. As long as OpenAI does not show the training data for ChatGPT, there would be no way to verify what is and what is not inside of the training data. The phenomenon that a model has access to test data inside of the training data is called data leakage. This generally has a lot of impact on the results, which could potentially have also happened in this thesis. Perhaps in the future when OpenAI decides to be more transparent to the scientific community, there would be a way for future researchers to know what papers had been impacted because of this possible data leakage.

6 Conclusion

Three different experiments have been conducted on ChatGPT in this thesis. Now let us look back at the question: “how distinguishable is the output from ChatGPT compared to human output when it comes to certain psycholinguistic dimensions and personality?”

When it comes to the psycholingsuistic dimension of typicality, ChatGPT generally performs well. For some categories it performs comparatively worse than other categories. In the worst categories, one might be able to detect that something suspicious is going on.

For the psycholinguistic dimension of imageability, it performs worse than typicality. The output it generated did not correlate well with the output of humans and the range of its answers was also off. In this case, someone would definitely determine its output to be different from humans when compared side to side.

Two different conclusions were drawn from the personality experiment. The answers it gave were stable and did correlate with each other. The results of the personality tests could not be determined as stable however. Therefore, the initial output would probably not be seen as different from any other humans. However, the results after processing these answers through a personality test might give away that a human did not take that test.

Important to note is that context matters in these cases. ChatGPT’s output can be considered suspicious in certain cases, while in other cases its output might seem fine. It also matters whether ChatGPT is directly compared to human data or whether ChatGPT’s output has to be found first inside of a large dataset containing actual human data. Human data often contains outliers. ChatGPT’s output might seem strange at first look, but data from actual humans can be as deviating as the data that ChatGPT has generated in this paper. In conclusion, the question: “how distinguishable is the output from ChatGPT compared to human output when it comes to certain psycholinguistic dimensions and personality?” can be answered with: ”ChatGPT’s output is not easily distinguishable from human output when it comes to typicality, imageability and personality.” Even if its output could be considered suspicious like in the imageability experiment and the results of the personality tests, it could also have been generated by a human outlier. There is no method that can 100% surely determine that data was generated by ChatGPT, but some indications can be found in certain cases. However, these indications are not enough to definitely conclude that the data was generated by ChatGPT and not by some human outlier.

References

[BE23] Robert Brandl and Cai Ellis. Survey: Chatgpt and ai content can people tell the difference?, 2023.

[Bor23] Ali Borji. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*, 2023.

[CCS12] Deborah A. Cobb-Clark and Stefanie Schurer. The stability of big-five personality traits. *Economics Letters*, 115(1):11–15, 2012.

[DDVA$^+$08] Simon De Deyne, Steven Verheyen, Eef Ameel, Wolf Vanpaemel, Matthew J Dry, Wouter Voorspoels, and Gert Storms. Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior research methods*, 40:1030–1048, 2008.

[FC19] Adrian Furnham and Helen Cheng. The change and stability of neo scores over six-years: A british study and a short review. *Personality and Individual Differences*, 144:105–110, 2019.

[Gar90] Michael Garman. *Psycholinguistics*. Cambridge University Press, 1990.

[HMB97] Robert W Hill, Karen McIntire, and Verne R Bacharach. Perfectionism and the big five factors. *Journal of social behavior and personality*, 12(1):257, 1997.

[KJ91] Nancy H Kerr and Thomas H Johnson. Word norms for blind and sighted subjects: Familiarity, concreteness, meaningfulness, imageability, imagery modality, and word associations. *Behavior Research Methods, Instruments, & Computers*, 23:461–485, 1991.

[KK84] Paul Kay and Willett Kempton. What is the sapir-whorf hypothesis? *American anthropologist*, 86(1):65–79, 1984.

[LB20] Christopher W Lynn and Danielle S Bassett. How humans learn and represent networks. *Proceedings of the National Academy of Sciences*, 117(47):29407–29415, 2020.

[LLL$^+$22] Xingxuan Li, Yutong Li, Linlin Liu, Lidong Bing, and Shafiq Joty. Is gpt-3 a psychopath? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*, 2022.

[MAA23] Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*, 2023.

[MDW03] Gerald Matthews, Ian J Deary, and Martha C Whiteman. *Personality traits*. Cambridge University Press, 2003.

[MER21] Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. Do language models learn typicality judgments from text? *arXiv preprint arXiv:2105.02987*, 2021.

[Nei89] Ulric Neisser. *Concepts and conceptual development: Ecological and intellectual factors in categorization.* Number 1. CUP Archive, 1989.

[SKB+19] Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, 51:1258–1270, 2019.

[Sum10] M Summerfield. User manual hilda release 9. *Melbourne: Melbourne Institute of Applied Economic and Social Research–The University of Melbourne*, 2010.

[Web01] Deborah A Weber. Restriction of range: The truth about consequences and corrections. 2001.

[WIS67] STEPHEN WISEMAN. The effect of restriction of range upon correlation coefficients. *British Journal of Educational Psychology*, 37(2):248–252, 1967.

Appendices

Appendix A

The code for the typicality experiment

```
"
Created by Donny Tsang

Last modified: 22/06/2023

"


#required packages
#install.packages("readxl")
#install.packages("ggplot2")
library(readxl)
library(ggplot2)
set.seed(42)

#Initialize file path
file_path <- "example/Typicality_results_comp.xlsx"

#Calculates correlation from a sheet
calc_corr<- function(sheet_name, option) {

  data <- read_excel(file_path, sheet = sheet_name)
  correlation <- NULL
  #Calculate the correlation between "Exemplar rating" and "NL rating"
  if (option == "NL"){
    correlation <- cor(data$‘Exemplar rating‘, data$‘NL rating‘)
  }
  #Calculate the correlation between "Exemplar rating" and "ENG rating"
  if (option == "ENG"){
    correlation <- cor(data$‘Exemplar rating‘, data$‘ENG rating‘)
  }
  return(correlation)
}

#Creates scatterplot from sheet
create_scatterplot <- function(sheet_name, option) {
  data <- read_excel(file_path, sheet = sheet_name)
  name <- paste(sheet_name, ".png")

  if (option == "NL"){
    #Adds jitter to plot
    x <- jitter(data$‘Exemplar rating‘)
    y <- jitter(data$‘NL rating‘)
    labels <- data$‘NL word‘
    #defines highest and lowest coordinate sum
    lowest_index <- which.min(x + y)
    highest_index <- which.max(x + y)

    #creates plot
    png(name)
    plot(x,y, xlab = "Exemplar ratings", ylab = "NL ChatGPT ratings", main = sheet_name)

    #highlights highest and lowest coordinate sum
    text(x[lowest_index], y[lowest_index], labels[lowest_index], cex = 1.2,
    col = "red", adj = c(0, 0.5))
    text(x[highest_index], y[highest_index] -0.5, labels[highest_index], cex = 1.2,
    col = "red", adj = c(1, -0.5))
    points(x[lowest_index], y[lowest_index], col = "red", pch = 16)
    points(x[highest_index], y[highest_index], col = "red", pch = 16)
    dev.off()
  }
  if (option == "ENG"){
    #Adds jitter to plot
    x <- jitter(data$‘Exemplar rating‘)
    y <- jitter(data$‘ENG rating‘)
    labels <- data$‘ENG word‘
    #defines highest and lowest coordinate sum
    lowest_index <- which.min(x + y)
    highest_index <- which.max(x + y)
    #creates plot
    png(name)
    plot(x,y, xlab = "Exemplar ratings", ylab = "ENG ChatGPT rating", main = sheet_name)

    #Higlights highest and lowest coordinate sum
    text(x[lowest_index], y[lowest_index], labels[lowest_index], cex = 1.2,
    col = "red", adj = c(0, 0.5))
    text(x[highest_index], y[highest_index] -0.5, labels[highest_index], cex = 1.2,
    col = "red", adj = c(1, -0.5))
    points(x[lowest_index], y[lowest_index], col = "red", pch = 16)
    points(x[highest_index], y[highest_index], col = "red", pch = 16)
    dev.off()
  }

}


#Load in the sheets
sheet_names <- c("Birds", "Fish", "Insects", "Mammals", "Reptiles", "Clothes",
                 "Kitchen Utensils", "Musical Instruments","Tools", "Weapons",
                 "Vehicles", "Fruit", "Vegetables", "Professions", "Sports")

#list of correlations and option
li <- c()
option <- "NL"

#Runs the two functions
#Calculates correlation and creates scatterplot for all sheets
for (sheet_name in sheet_names) {
  correlation <- calc_corr(sheet_name, option)
  create_scatterplot(sheet_name, option)
  print(paste("Correlation for", sheet_name, ":", correlation))
  li <- append(li, correlation)
}

mean(unlist(li))
```

Appendix B

The word selection code for the imageability experiment

```
"""
Created by Donny Tsang
```

```
Last modified: 22/06/2023

"""

import pandas as pd

def sample(option):
    df = pd.read_csv("13428_2018_1099_MOESM2_ESM.csv")
    if option == "all":
        imageability = df[["Words", "IMAG"]]
        df2 = imageability.sample(n=500, random_state=42)
        df2.to_csv(’imag_samples.csv’, index=False)
    elif option == "word":
        imageability = df[["Words"]]
        df2 = imageability.sample(n=500, random_state=42)
        df2.to_csv(’imag_onlyword.csv’, index=False)
    elif option == "imag":
        imageability = df[["IMAG"]]
        df2 = imageability.sample(n=500, random_state=42)
        df2.to_csv(’imag_onlyImag.csv’, index=False)

if __name__ == "__main__":
    sample("imag")
```

Appendix C

The data processing code for the imageability experiment

```
"
Created by Donny Tsang

Last modified: 22/06/2023

"
# Read the Excel sheet into a data frame
setwd("C:/Users/gebruiker/Documents/Scriptie")
data <- read.csv("C:/Users/gebruiker/Documents/Scriptie/Imageability_results.csv")

# Extract the columns
gpt <- data$GPT
human_mean <- data$Human_mean

# Calculate the correlation
```

```
result <- cor.test(gpt, human_mean)

# Create a new figure
jpeg("boxplots.jpg", res = 100)
par(mfrow = c(1, 2))

# Boxplot for Human_mean
boxplot(human_mean, main = "Human Mean", ylab = "Values", outline = FALSE)
points(jitter(rep(1, length(human_mean)), factor = 2.5), human_mean, col = "blue")

# Boxplot for GPT
boxplot(gpt, main = "GPT", ylab = "Values", outline = FALSE)
points(jitter(rep(1, length(gpt)), factor = 2.5), gpt, col = "blue")

# Reset the figure configuration
par(mfrow = c(1, 1))

dev.off()

# Print the correlation
print(result)
```

Appendix D

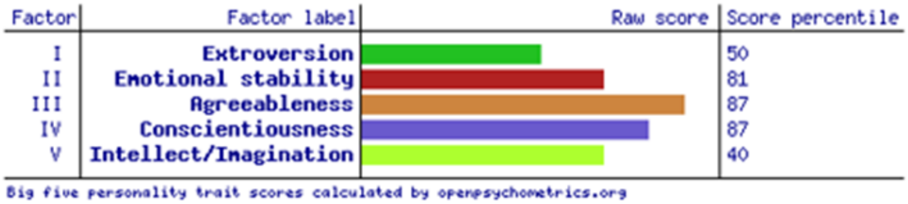
The code for the personality experiment

```
"
Created by Donny Tsang

Last modified: 22/06/2023

"

library(readxl)

# Read the Excel sheet into a data frame
data <- read_excel("C:/Users/gebruiker/Documents/Scriptie/GPT personality.xlsx",
                   sheet = "Comp")

# Extract the columns
run1 <- data$‘First run‘
run2 <- data$‘Second run‘
run3 <- data$‘Third run‘
```

```
# Calculate the correlations
correlation1x2 <- cor(run1, run2)
correlation2x3 <- cor(run2, run3)
correlation1x3 <- cor(run1, run3)

# Print the correlation
print(paste(correlation1x2))
print(paste(correlation2x3))
print(paste(correlation1x3))
```

Appendix E

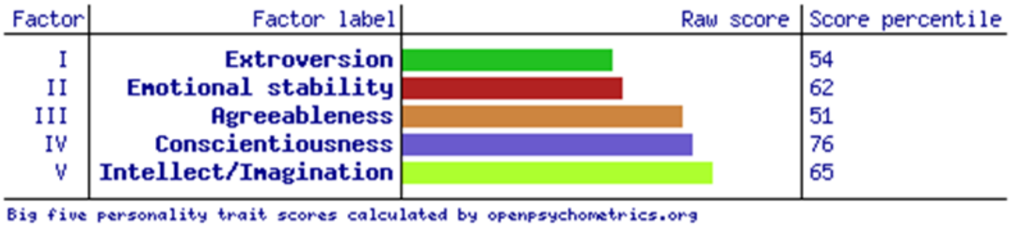
The personality test results



Factor	Factor label	Raw score	Score percentile
I	Extroversion		50
II	Emotional stability		81
III	Agreeableness		87
IV	Conscientiousness		87
V	Intellect/Imagination		40

Big five personality trait scores calculated by openpsychometrics.org

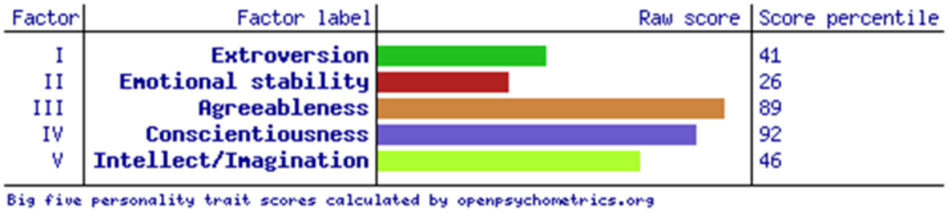
Figure 6: The results of the personality test for the first batch of ChatGPT's answers



Factor	Factor label	Raw score	Score percentile
I	Extroversion		54
II	Emotional stability		62
III	Agreeableness		51
IV	Conscientiousness		76
V	Intellect/Imagination		65

Big five personality trait scores calculated by openpsychometrics.org

Figure 7: The results of the personality test for the second batch of ChatGPT's answers



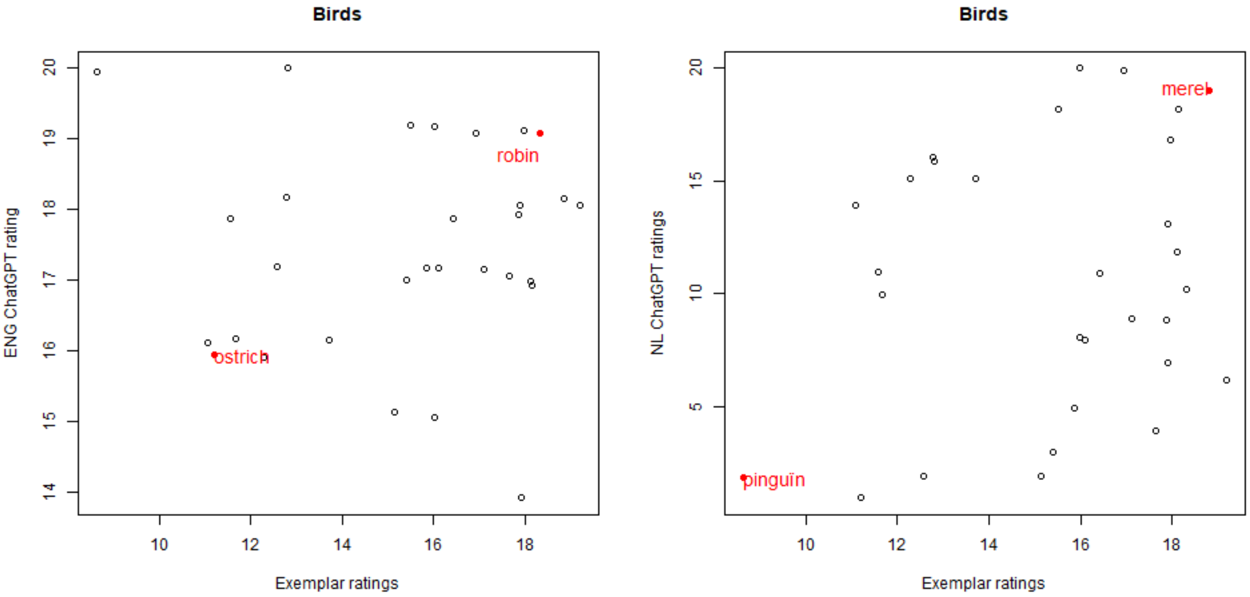
Factor	Factor label	Raw score	Score percentile
I	Extroversion		41
II	Emotional stability		26
III	Agreeableness		89
IV	Conscientiousness		92
V	Intellect/Imagination		46

Big five personality trait scores calculated by openpsychometrics.org

Figure 8: The results of the personality test for the third batch of ChatGPT's answers

Appendix F

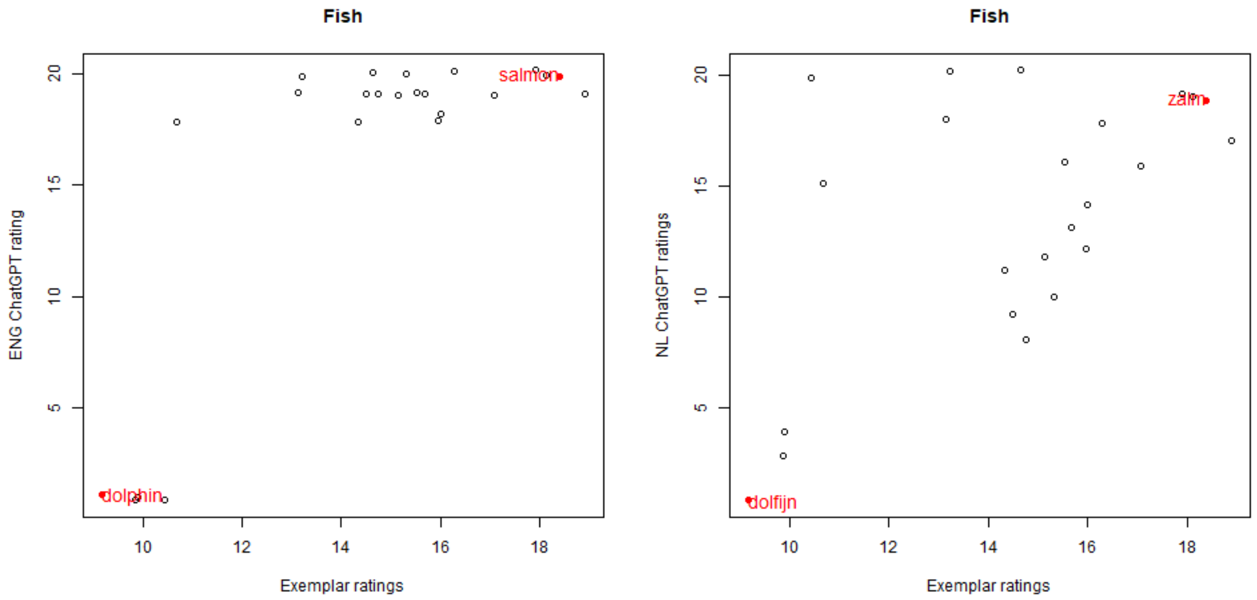
The scatterplot outputs of the typicality experiment



(a) The scatterplot for English ChatGPT typicality ratings for birds

(b) The scatterplot for Dutch ChatGPT typicality ratings for birds

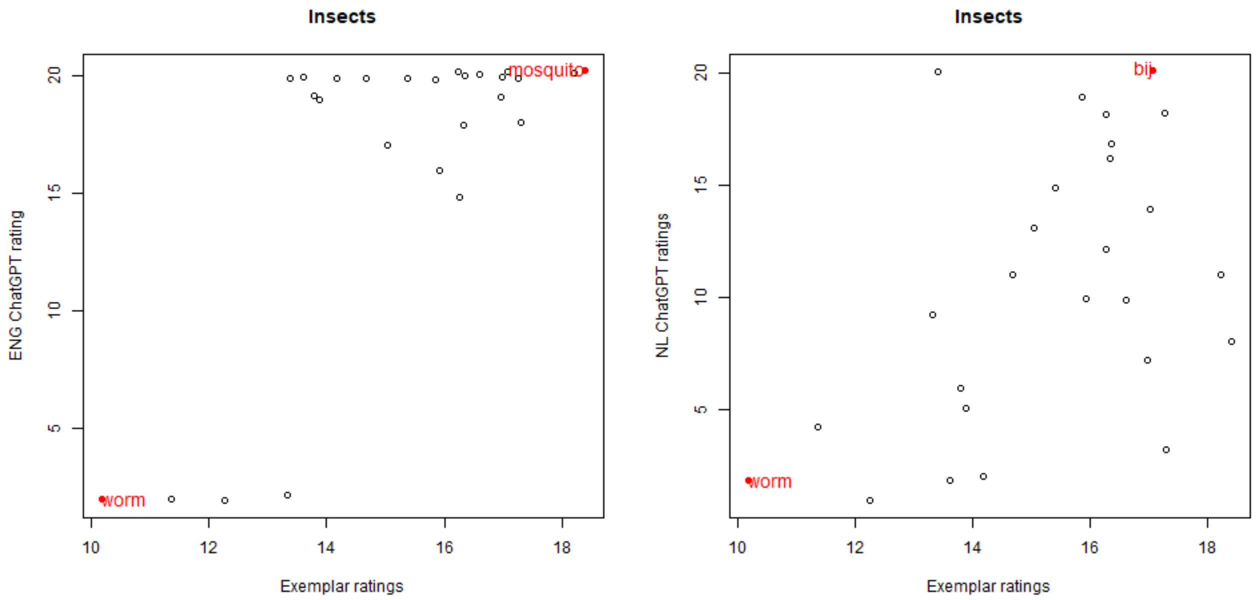
Figure 9: The comparison for the category birds, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for fish

(b) The scatterplot for Dutch ChatGPT typicality ratings for fish

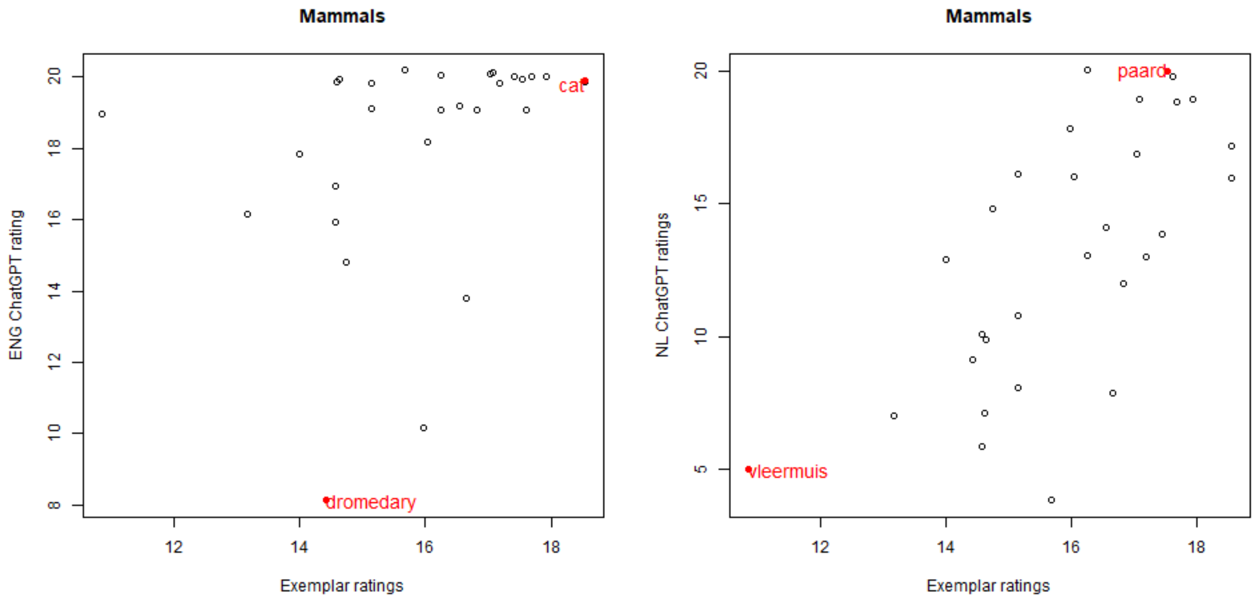
Figure 10: The comparison for the category fish, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for insects

(b) The scatterplot for Dutch ChatGPT typicality ratings for insects

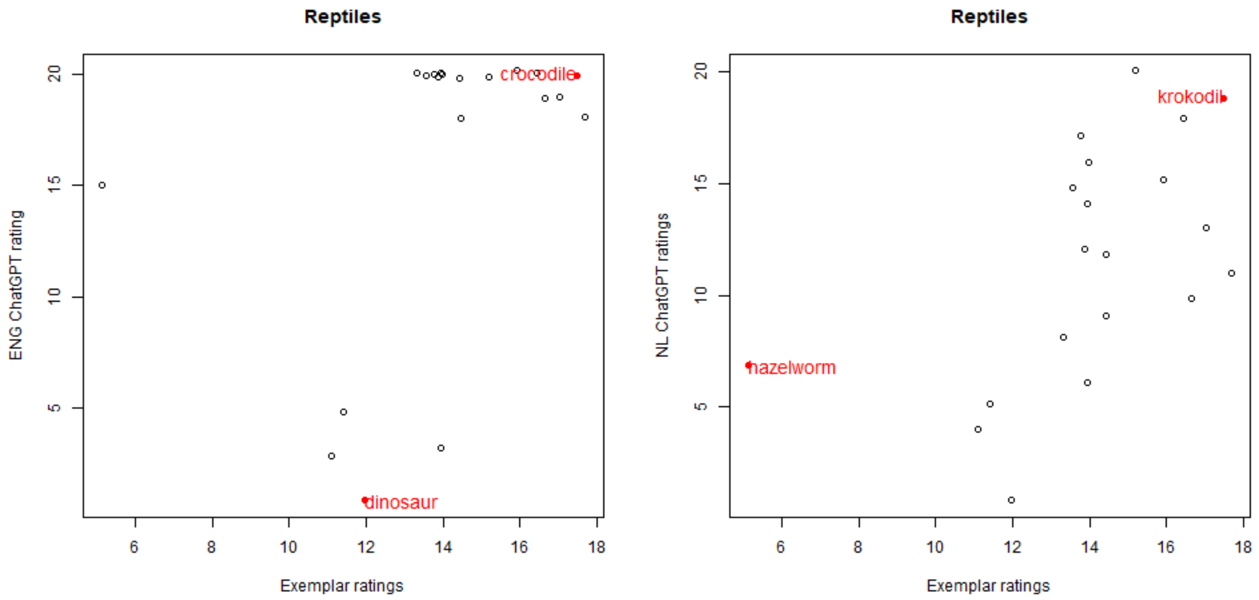
Figure 11: The comparison for the category insects, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for mammals

(b) The scatterplot for Dutch ChatGPT typicality ratings for mammals

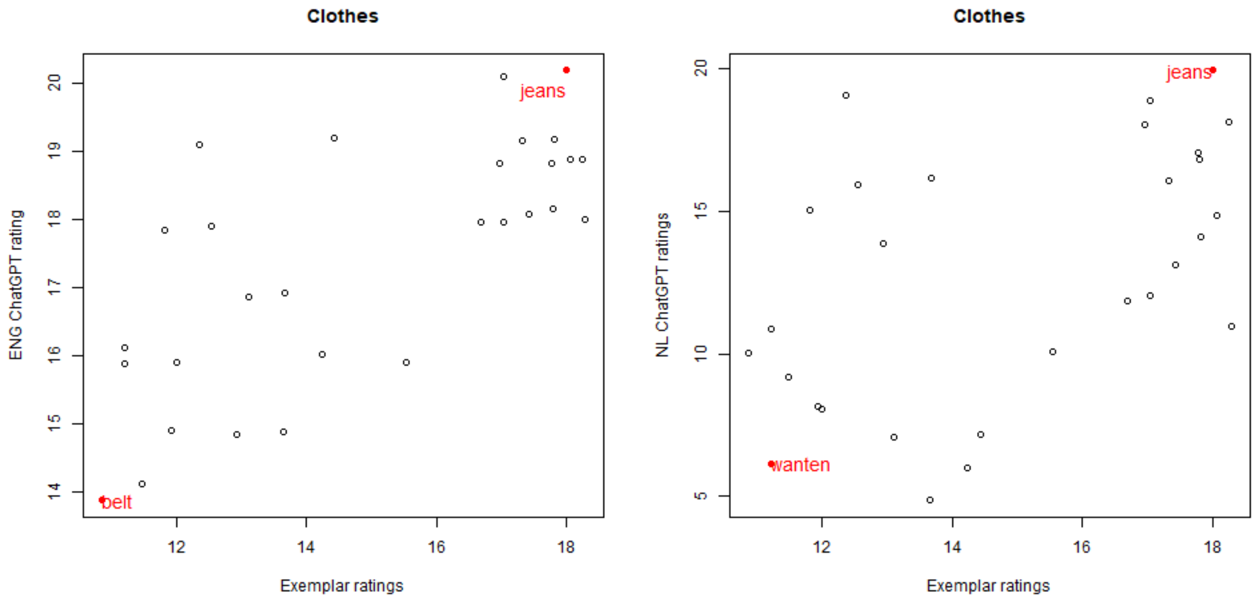
Figure 12: The comparison for the category mammals, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for reptiles

(b) The scatterplot for Dutch ChatGPT typicality ratings for reptiles

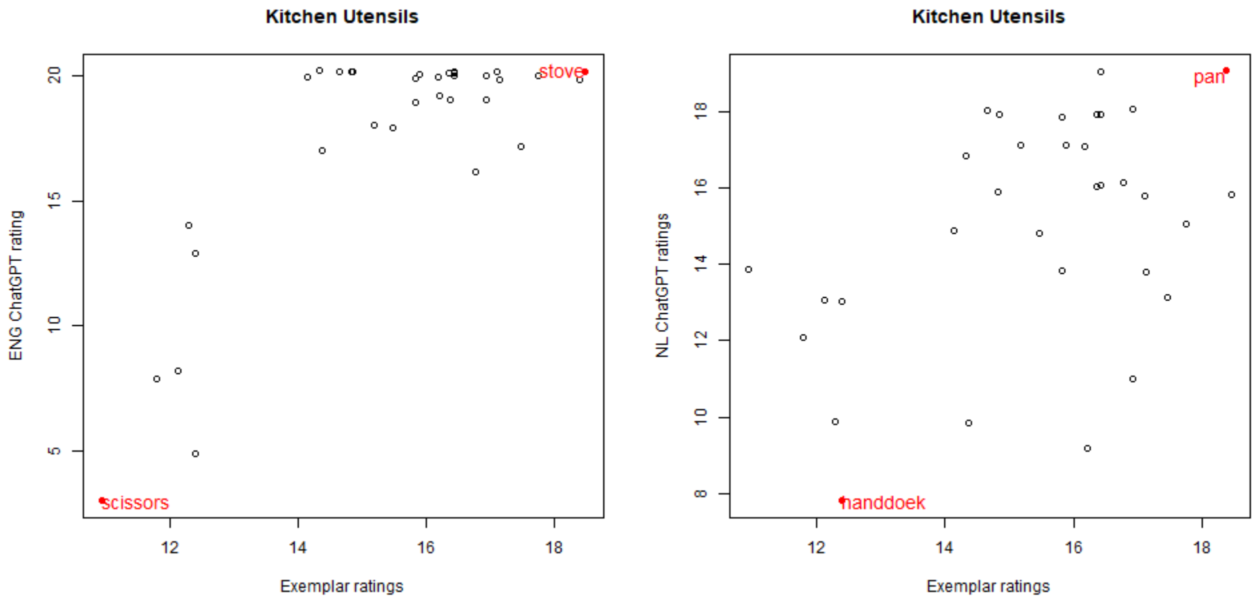
Figure 13: The comparison for the category reptiles, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for clothes

(b) The scatterplot for Dutch ChatGPT typicality ratings for clothes

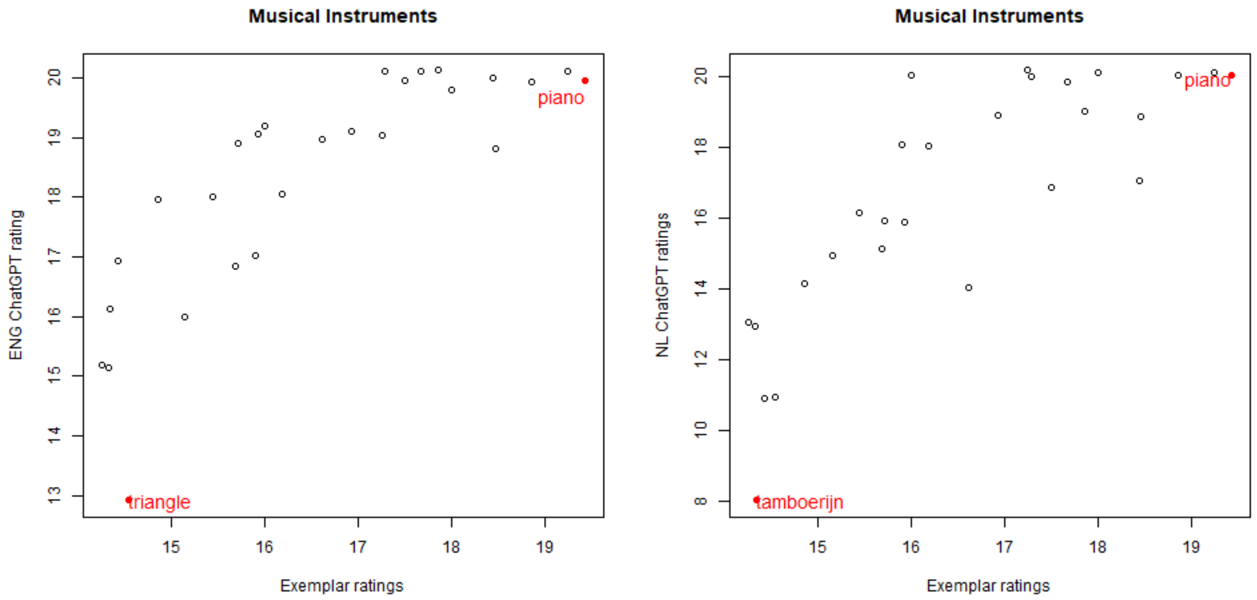
Figure 14: The comparison for the category clothes, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for kitchen utensils

(b) The scatterplot for Dutch ChatGPT typicality ratings for kitchen utensils

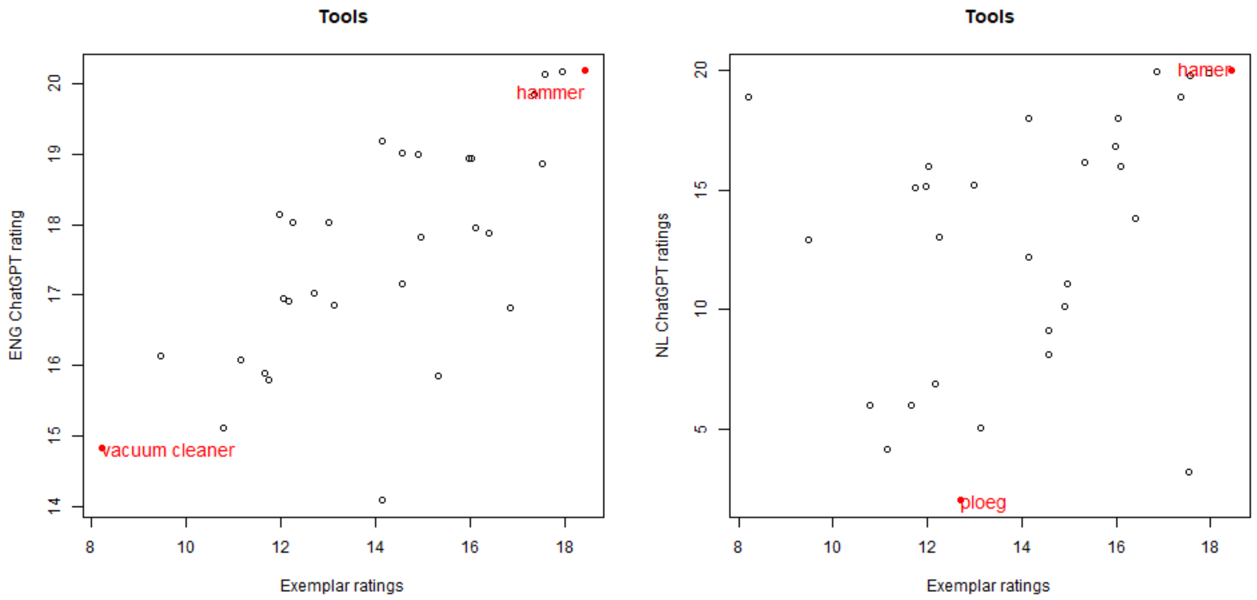
Figure 15: The comparison for the category kitchen utensils, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for musical instruments

(b) The scatterplot for Dutch ChatGPT typicality ratings for musical instruments

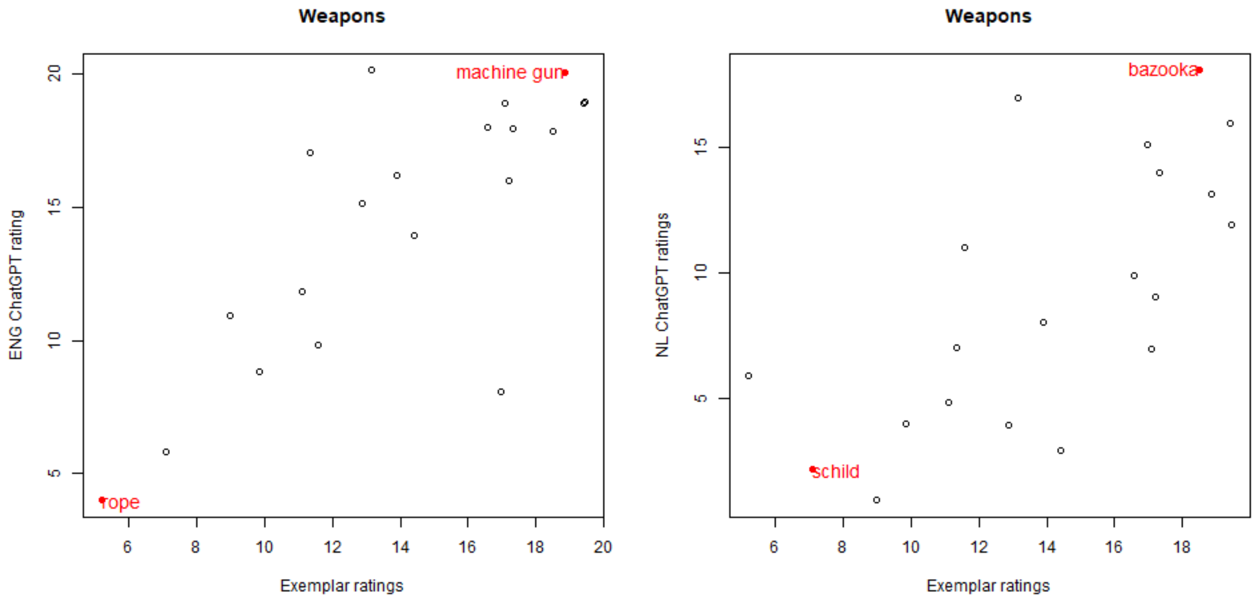
Figure 16: The comparison for the category musical instruments, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for tools

(b) The scatterplot for Dutch ChatGPT typicality ratings for tools

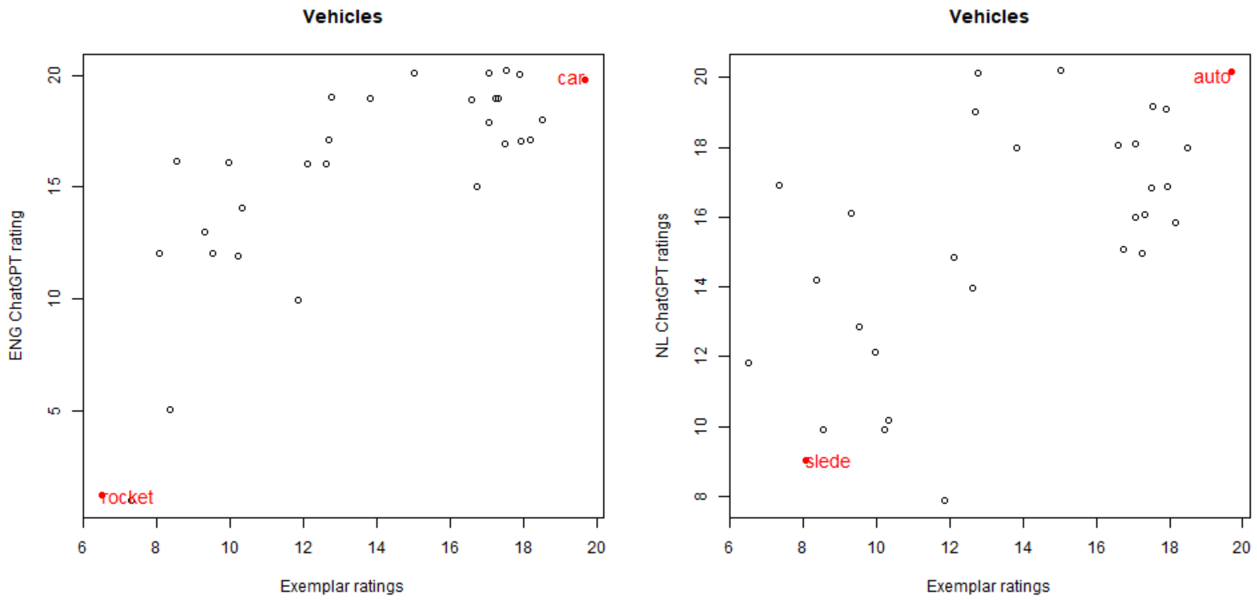
Figure 17: The comparison for the category tools, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for weapons

(b) The scatterplot for Dutch ChatGPT typicality ratings for weapons

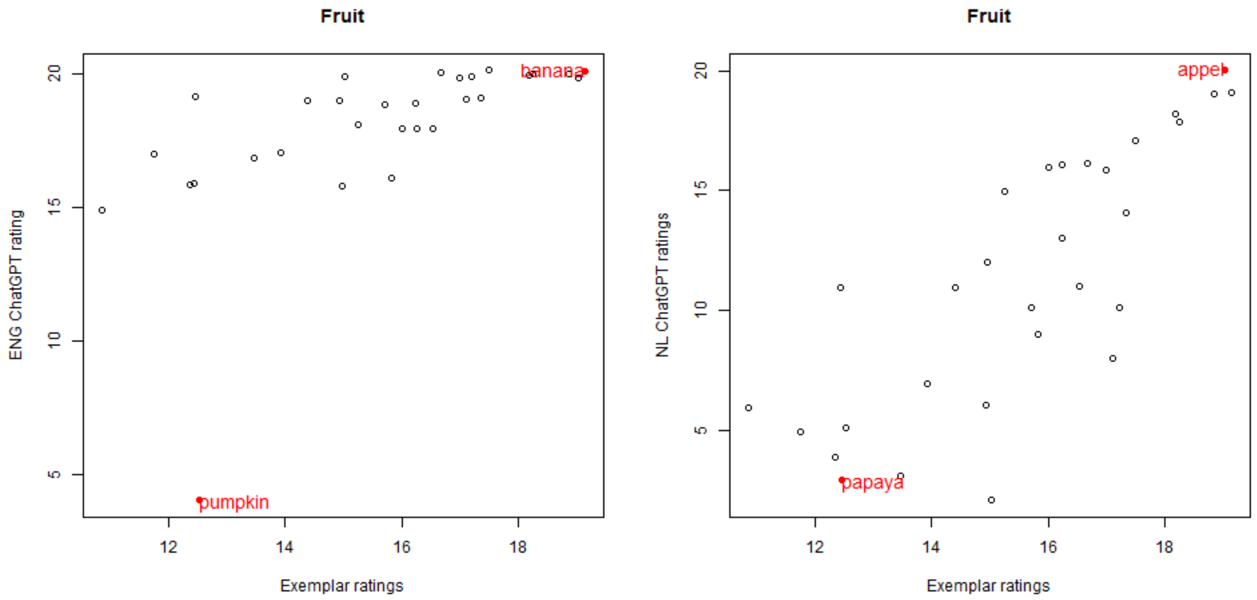
Figure 18: The comparison for the weapons, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for vehicles

(b) The scatterplot for Dutch ChatGPT typicality ratings for vehicles

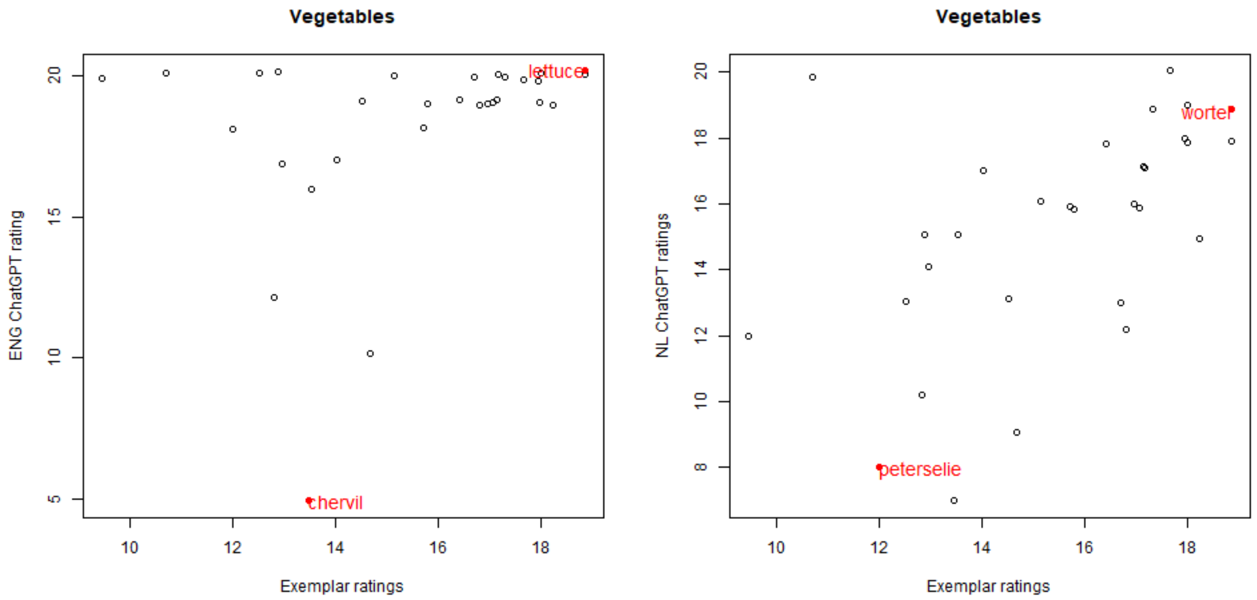
Figure 19: The comparison for the vehicles, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for fruit

(b) The scatterplot for Dutch ChatGPT typicality ratings for fruit

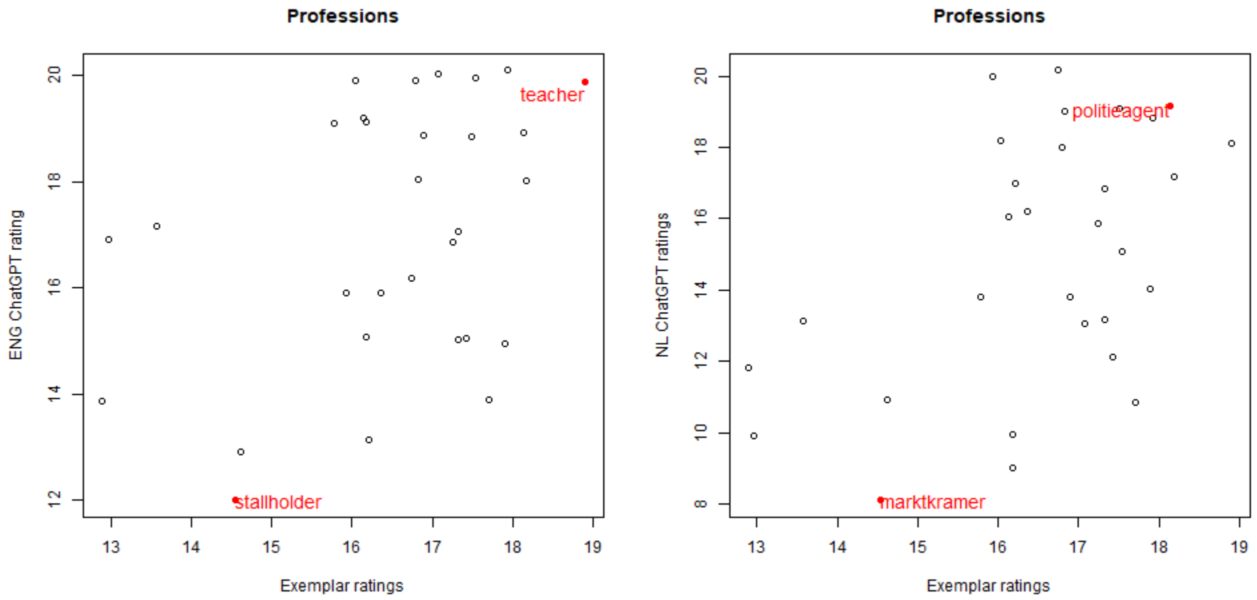
Figure 20: The comparison for the category fruit, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for vegetables

(b) The scatterplot for Dutch ChatGPT typicality ratings for vegetables

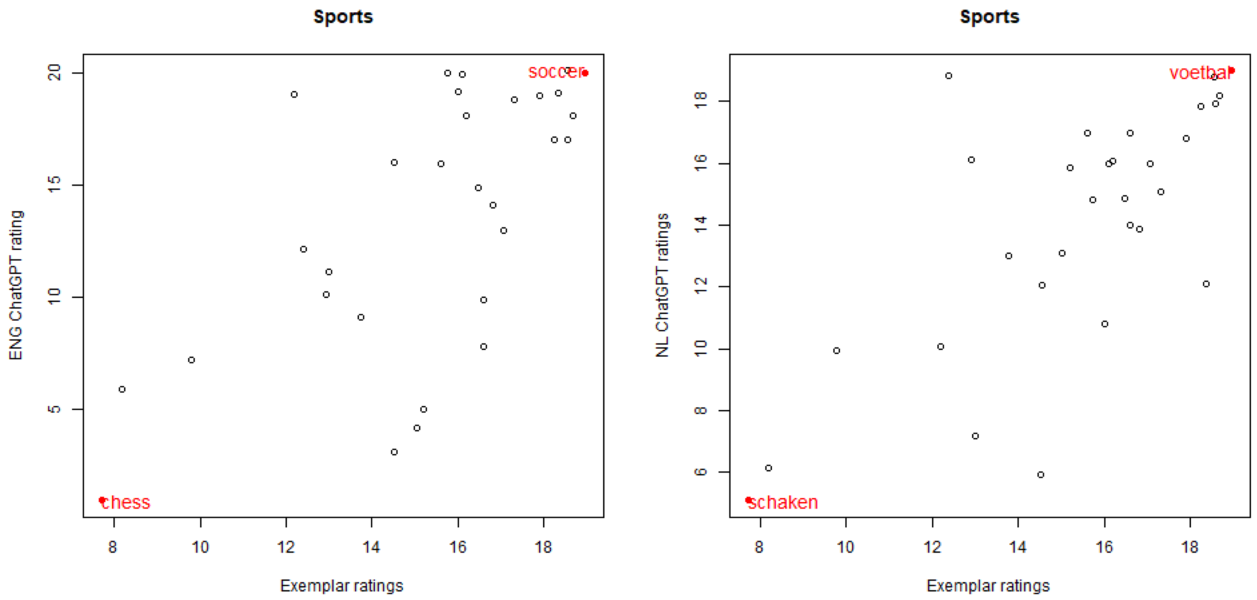
Figure 21: The comparison for the vegetables, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for professions

(b) The scatterplot for Dutch ChatGPT typicality ratings for professions

Figure 22: The comparison for the category professions, with English being left and Dutch being right



(a) The scatterplot for English ChatGPT typicality ratings for sports

(b) The scatterplot for Dutch ChatGPT typicality ratings for sports

Figure 23: The comparison for the category sports, with English being left and Dutch being right