An NLP Analysis of ChatGPT's Personality Simulation Capabilities and Implications for Human-centric Explainable AI Interfaces

Thorsten Zylowski and **Matthias Wölfel**
University of Hohenheim, Schloss Hohenheim 1, 70599 Stuttgart, Germany
Karlsruhe University of Applied Sciences, Moltkestraße 30, 76133 Karlsruhe, Germany
thorsten.zylowski@uni-hohenheim.de

Abstract

This paper explores the potential of ChatGPT in simulating diverse personalities for application in adaptive human-centric eXplainable Artificial Intelligence (XAI) interfaces. A dataset of 4329 text datapoints across 13 simulated personalities from ChatGPT were collected. Extensive linguistic analyses were conducted using metrics from Natural Language Processing (NLP) for basic linguistic features, readability, lexical richness, and sentiment. Additionally, a personality classifier was trained with a F1-score of 0.79 to understand which personalities are unique in wording and style. This was further substantiated through the application of the SHAP (SHapley Additive exPlanations) framework, which unveiled important words in personality classification. It was found that ChatGPT is capable of simulating several levels of professionalism as well as more emotional personalities that adapt human characteristics and can be used in human-centric XAI interfaces, although specific user testing is still pending.

1 Introduction

With ChatGPT, large language models (LLM) and generative artificial intelligence (generative AI) are entering daily life at a speed never before seen with any other technology. ChatGPT contributes to empowering people and assists with many everyday and professional tasks. One reason for this high adoption rate is the chosen interface in form of a conversation, which is a natural way of interaction. When a question is asked, ChatGPT tries to provide accompanying descriptions and explanations. However, it is also known that ChatGPT tends to provide false information and express it confidently and also make up facts and sources. This effect is referred to as hallucination. (Ji et al., 2023) ChatGPT is known for generating high-quality texts for various situations, but it is also known to differ in its wording from people. (Mitrović et al., 2023)

Due to its ability to produce well readable and high quality text, ChatGPT has great potential to assist in the development of human-centered explainable artificial intelligence (XAI) interfaces, which help to increase trust in artificial intelligence (AI) systems by making AI decisions more transparent. Many machine learning (ML) models still have the problem that they are opaque and cannot be explained. The reason for this is the black-box nature of these models, which makes it impossible for a human to understand the decision paths of the model. One task of XAI is to extract information from the model that can be provided to a human so that the model's decisions can be understood. In addition, it is of high importance *how* the explanation is given.

Miller (2019) describes that a good explanation is *social*, referring to work by Hilton (1990), according to whom a *"causal explanation is first and foremost a form of social interaction."* Consequently, it is important for a good explanation *who* the explainer is, who the receiver of the explanation is (explainee) and what the context is. For example, it makes a difference whether a professor is explaining something to another professor within the same research field, or whether he/she is explaining something to a student. The way the explanation is given differs in both situations. In contrast, ChatGPT responds to every request in the same way, unless prompted otherwise. It has no information about who it is having the conversation with unless it is made aware of it. Miller (2019) further elaborates on the work of Hilton (1990), who describes that a causal explanation is always a *conversation*. Accordingly, it would be desirable to have XAI interfaces capable of generating explanations in natural language and adapted to the situation and to the human in form of a conversation. This is further reinforced by the fact that people demand that a good explanation can adapt to their needs. (Zylowski, 2022) This includes,

among other things, the ability to get explanations on demand and in different formats and granularities. Before the developments in the field of LLM, developing a conversational human-centric XAI interface that can be adjusted to the requirements of a person was very difficult or impossible and the impact of intent-based conversational interfaces were limited. (Jentzsch et al., 2019) Even if it was known what a good explanation to a person should look like, it was technically very difficult to actually create an adequate explanation. With the potential of ChatGPT to simulate different personalities, it is possible to develop XAI interfaces that can be adapted to different people and to different needs and requirements of those people. However, it is still an open question how well ChatGPT can simulate different personalities and how well responses are adapted to people's needs.

This paper investigates the ability of ChatGPT to simulate different personalities and describes the advantages for adaptive human-centric XAI interfaces. An NLP approach is chosen by applying different metrics to ChatGPT texts in different personality styles and it is investigated how clearly these personalities can be distinguished from each other and which phrases and words are typical for different personalities. By exploring the potential of adapting explanations to individuals, this work aims to address the current limitations and unlock the full potential of ChatGPT in fostering trust and transparency in AI systems.

2 Related Work

ChatGPT's responses are currently being studied by many researchers and the applicability in different domains is being validated. It is investigated whether texts generated by humans can be distinguished from those generated by ChatGPT and what the differences are. Mitrović et al. (2023) investigate whether a classifier can be trained to distinguish human-generated texts from those generated by ChatGPT, achieving 79% accuracy. Through an analysis of the classifier with the XAI framework SHAP, they look for differences between the formulations. They find that ChatGPT tends to focus on describing experiences rather than expressing feelings and it avoids using personal pronouns. Moreover, it has a tendency to utilize uncommon or unusual words and never employs aggressive language or rude vocabulary in its responses. Mindner et al. (2023) created several text classifiers for the educational field to distinguish texts generated and rephrased by ChatGPT from human-created texts, with F1-scores of over 96% and 78%, respectively, outperforming even GPTZero[1], the most prominemt approach, in the best *basic* text rephrasing task. Other authors focus on how trustworthy the texts generated by ChatGPT appear to people. Li et al. (2023) analyze the applicability of ChatGPT for Information Extraction (IE) tasks and found that ChatGPT performs poorly on the Standard-IE setting, but performs very well on the Open-IE setting. Furthermore, they investigated the quality and trustworthiness of the explainations of ChatGPT responses in a self-check and by domain experts and judged them to be of high quality and trustworthy.

One aspect that is not yet investigated in current studies is the adaptability of the formulations of ChatGPT in different situations and under different user requirements. For effective use in human-centered XAI interfaces, ChatGPT must be able to generate different types of formulations that are adapted to people's needs. It is important that interfaces also address humans on an emotional level to enable trust. The fact that ChatGPT tends not to express emotions (Mitrović et al., 2023), can be challenging in this regard.

3 Method

This section presents the approach including data collection and metrics used.

3.1 Data Collection

For this study, a total of 333 instructions were manually selected from the ShareGPT[2] dataset which contains real world examples of conversations with ChatGPT. For the selection process a set of 500 randomly selected instructions was created. The instructions were then manually filtered based on usefulness (i.e. instructions that were not written in English or that consisted of only one word or that contained only a technical command were removed). The instructions were then utilized to interact with the ChatGPT API. The *gpt3.5-turbo* model was selected for the data collection process, because it was the best accessible model at the time. To ensure a comprehensive analysis, ChatGPT was requested to respond to the instructions, in addition to the default answer, using 10 distinct personality

[1] https://gptzero.me/
[2] https://sharegpt.com/

styles as described in Section 3.2. In order to cater to different user groups commonly encountered in XAI interfaces, two additional styles were incorporated — one targeting laypeople and the other aimed at experts. Thus, a total of 13 different styles were applied during the interactions and resulted in the acquisition of 4329 datapoints.

3.2 Personality Styles

In order to address people in different ways in an XAI interface, besides ChatGPT's default style, 10 personality styles were created to satisfy individual needs and requirements. Additionally the two target groups *laypeople* and *expert* are described. When selecting the personality styles, attention was paid to a diverse range and existing findings were taken into account (e.g. addressing laypersons and experts). The styles should contain human characteristics that can be useful for an explanation. However, the real usefulness still has to be determined in user experiments.

Default: ChatGPT's default personality if no specific prompt to change its personality is given.

Child-like: Simple, short and playful and targets a younger audience or can be helpful when explaining basic concepts to users who prefer a more lighthearted and approachable explanation.

Parent-like: A parent-like explanations could provide patience and empathy and may offer guidance and support throughout the learning process.

Professorial: High level of expertise and use academic language. Those styles could be useful when catering to users who appreciate in-depth knowledge and a more formal style of explanation.

Friendly Companion: Supportive, uses conversational tone, engages in conversations and listens actively to user queries.

Expert Guide: Talks in a knowledgeable and authoritative manner and can be effective when users are seeking accurate and detailed information from a trusted source.

Storyteller: Focus on narratives and anecdotes and can give a more memorable experience, enabling users to connect with the AI through storytelling.

Helpful Assistant: Creates clear and concise explanations and emphasizes practicality and utility.

Humorous & Entertaining: Uses jokes, puns, or witty remarks and can make the interaction more enjoyable and help alleviate potential boredom or monotony during the explanation process.

Motivator: Inspiring and encouraging users and can provide positive reinforcement, acknowledge progress, and instill confidence in users' ability to grasp the material.

Technician: Pays attention to detail and on the technical aspects and can be valuable for technical professionals.

Laypeople: Aims at laypersons and will try to present content in a way that is easy to understand.

Expert: Targets experts and will provide a lot of expert knowledge.

The personality styles can be roughly divided into two categories. One category includes more technical and professional personality styles (ChatGPT's default, professorial, expert, expert guide, technician). The other category includes emotional, human-oriented personality styles (helpful assistant, storyteller, motivator, laypeople, friendly companion, humorous/entertaining, parent-like, child-like). The two categories are not strictly separated and overlap of personality styles is possible.

3.3 Prompts

A separate prompt was created for each personality style, with most of them following the scheme:

"I want you to communicate like a <personality> in the following conversation. I will give you a question or instruction and I want you to answer it in a way a <personality> would do it.

<instruction>"

For personality styles like child-like or professorial the personality tag was replaced with words like *"child"* or *"professor"*. In cases where this was not possible, the prompt was adjusted to instruct ChatGPT to act in a specific way,

e.g. *"I want you to communicate in a humorous/entertaining way [...]"*. The two prompts, aimed at laypeople and experts, follow the scheme:

"I want you to communicate in such a way that your answers are directed at laypeople/experts. I will give you a question or instruction and I want you to answer it in a way that laypeople/experts can understand.

<instruction>"

The instruction tag was replaced by an instruction from ShareGPT.

4 Metrics

Metrics from the fields of NLP and linguistics were used to analyze the responses of ChatGPT. In addition, a ML model was trained that attempts to classify the different personality styles based on the texts. An XAI framework was then used to examine specific words and phrases of these styles.

4.1 Linguistic Metrics

For the linguistic analysis of ChatGPT's responses the spaCy framework[3] in version 3.6 was used. The *average token-wise text length* was calculated to investigate if there are differences in the length of the answers between the personalities. To check how direct and precise a text is written, the *average number of stopwords* was determined. Personalities expected to produce text that is low in information density are likely to have an increased number of stopwords. These could include, for example, the storyteller and the motivator personalities.

Further differentiation of responses could be possible by the *average number of named entities* used. The more precise an answer is, the more named entities it might contain. A more professional answer will most likely contain more facts and details that could also be expressed in named entities. To compensate for the dependence on the length of the texts, the average number of stopwords and named entities per 100 words was calculated.

4.2 Readability

The readability of the responses in the different personalities of ChatGPT is expected to differ significantly. For this reason, the readability was compared using the well established Flesch Reading Ease (FRE) index (Flesch, 1948). The index has a range from 0 to 100, where a value of 0 represents very hard to read text and 100 represents very easy to read text. The FRE is derived from a base value from which the weighted average sentence length and the weighted average number of syllables in a word are subtracted.

4.3 Lexical Richness

Lexical richness is classically formed as the token-type ratio (TTR), which is the relation of unique words (types) and the set of total words (tokens) (Templin, 1957). There are several variants of this measure that make corrections to compensate for a dependence on the length of the text (Torruella and Capsada, 2013). One of these measures is the Measure Of Textual Lexical Diversity (MTLD) as described in McCarthy and Jarvis (2010). For the calculation, the text is divided into segments for which the TTR is calculated. A segment is expanded until a TTR of a given threshold is reached. Then the number of words in the text is divided by the number of segments to calculate the lexical richness. A higher MTLD suggests that the text uses a wider variety of words across its segments, and therefore has greater lexical richness. A lower MTLD indicates that the text may have more repetition and less varied vocabulary.

4.4 Sentiment Analysis

Personality styles that formulate text on an emotional level (e.g. motivator personality) can be expected to show increased positive or negative sentiment. In order to investigate whether personalities exhibit a certain sentiment, a sentiment analysis of the texts was performed. The model used is *distilbert-base-uncased-finetuned-sst-2-english* which is a DistilBERT model (Sanh et al., 2020) fine-tuned on SST-2 dataset (Socher et al., 2013). The classification results in an assignment of a label POSITIVE or NEGATIVE to the text with a percentage indication of the strength of the sentiment. To distinguish positive and negative sentiments numerically, all negative sentiments were converted to a negative value. Thus, all positive sentiments run from 0 to 1 and all negative sentiments from 0 to -1.

4.5 Personality Classifier

To gain a deeper understanding of the ChatGPT texts, a classifier was trained that attempts to predict the respective personality based on the texts.

[3] https://spacy.io/

It is reasonable to assume that there are personalities that are very easy to predict because they have unique phrases and style. Other personalities are more similar to each other and more difficult to predict. For the training, a *distilbert-base-uncased* model was fine-tuned on the 4329 data points with 20% test data, 5 epochs of training, a learning rate of 0.00002 and a batch size of 16.

4.6 Explaining the Classifier

The personality classifier was examined using the XAI framework Shapley Additive Explanations (SHAP) (Lundberg and Lee, 2017) to analyze which words are typical of the different personalities. SHAP is based on the shapley values of cooperative game theory and attempts to assign a value to each feature, in the case of text each token, indicating how much contribution that feature has to the overall classification.

5 Results

Presented below are the results of the NLP analysis of the 4329 data points split between the 10 defined personalities, the two specific target groups of laypeople and experts, and the default response of ChatGPT. For simplified readability, the following will always refer to 13 personalities.

5.1 Text Length

The average text length of the responses for the different personality styles is shown in Figure 1. The length of ChatGPT's default response is in the upper range of values with a average text length of 332 tokens. The child-like personality has the shortest average length with 148 tokens, while the storyteller personality has the longest with 468 tokens on average. The more professional/technical styles are in the upper range of values.

5.2 Stopwords

Figure 2 shows the average number of stopwords for each personality style with ChatGPT's default answer at the second position with 31 stopwords per 100 words. It can be seen that the technical/professional personality styles are in the lower range of stop words and the more personal/emotional styles are in the upper range with child-like (42 stopwords per 100 words) and parent-like (43 stopwords per 100 words) at the top.

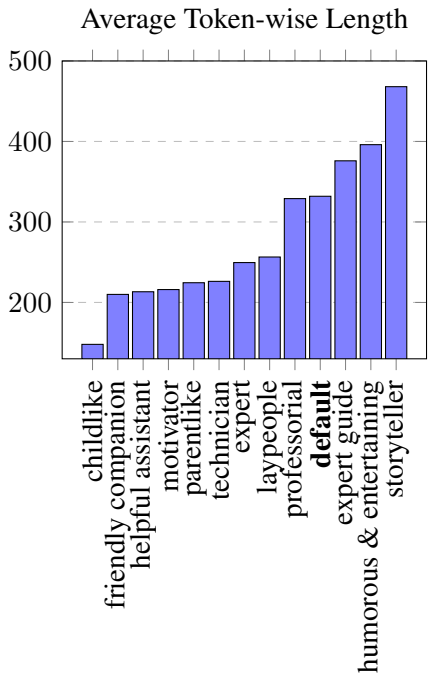

Figure 1: The average token-wise lengths of the personality styles.

5.3 Named Entities

The average number of named entities is shown in Figure 3. The lowest number of named entities occur for the personality styles motivator with 2.22 named entities per 100 words and storyteller with 2.23 named entities per 100 words, the highest for ChatGPT's default response with 4.42 named entities per 100 words. The technical styles also tend to be on the higher end for the number of named entities. Surprisingly, the professorial style is an exception.

When the named entities are split according to their categories, it can be seen that the motivator personality uses almost no numbers and the default personality of ChatGPT uses no ordinal entities, such as *"first"*, *"second"*, *"third"*, etc. The entertaining personality contains the most named entities with the category WORK OF ART, which classifies book tiles, song names etc. The expert personality has a high value for the FAC category which contains building, airports, highways etc.

5.4 Readability

The scores for the Flesch Reading Ease index for the different personality styles are shown in Figure 4. Low values mean that a text is more difficult to read. The technical and more professional styles, including ChatGPT's default response, are on the

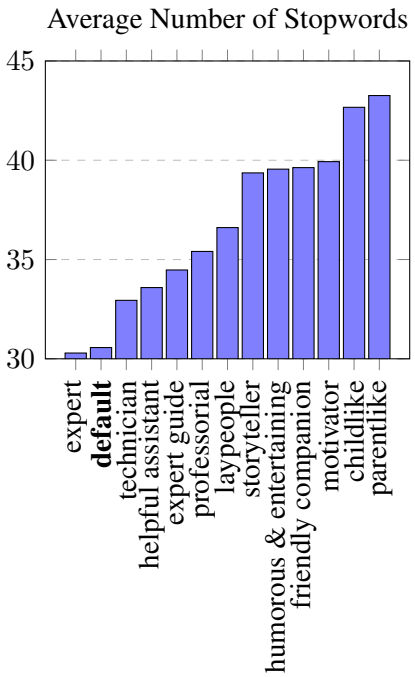

Figure 2: Average number of stopwords of the personality styles per 100 words.

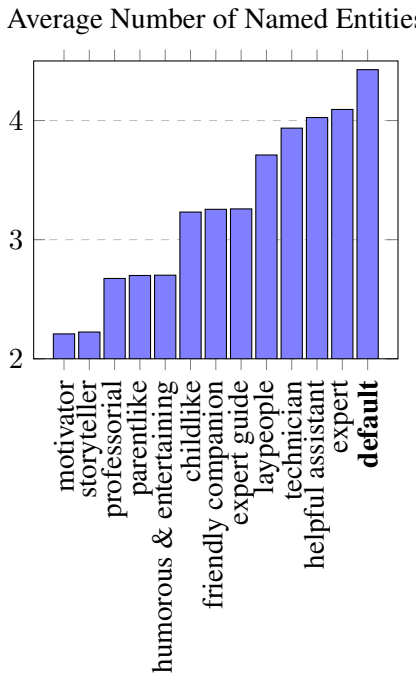

Figure 3: Average number of named entities of the personality styles per 100 words.

lower end. The professorial style is the hardest to read with a score of 19.12, while the child-like is the easiest with a score of 77.99. For all personalities, the average is below 78, which means that on average none of the texts are easy or very easy to read.

5.5 Lexical Richness

The distribution of the MTLD lexical richness scores is shown in Figure 5. ChatGPT's default answer has the lowest MTLD score (71.0), similar to the child-like personality, which means, that the lexical richness is low. The highest value is for the humorous/entertaining personality (129.0). No distinction can be made between technical/professional and non-technical/non-professional as in the other results. Lexical Richness seems to be a very individual property of the respective personality styles.

5.6 Sentiment Analysis

In Figure 6 is shown, that the average sentiment scores ranges from 0 to 1, with ChatGPT's default personality having a value close to 0. This does not mean that this personality generates very neutral texts. The opposite is true, as can be seen in Figure 7. The distribution of negative and positive sentiments balance each other, resulting in a neutral value on average. In fact, all the personalities behave in this way, with differing weights. For example, the motivator personality has the predominant amount of sentiments in the upper positive range. In particular, there is no negative trend in sentiment. All changes in the weights are in the direction of more positive sentiment. Texts that contain code tend to be classified with a negative label.

5.7 Evaluating the Personality Classifier

The personality classifier to decide a personality style out of the 13 classes reached a F1-score of 0.79. Although the value is already quite good, a difference can be seen between the individual personality styles. As suspected, there are styles that are particularly predictable. Humorous/entertaining, storyteller and parent-like personalities with a F1-score of 0.97, child-like with 0.95 and professorial with 0.89. These personality styles have very unique formulations and style. The storyteller personality in particular has frequent unique phrases, such as *"once upon a time"*, that are not used by other styles. The hardest to predict personalities are technician with F1-score of 0.48, expert with 0.61 and laypeople with 0.63.

A look at the confusion matrix, which can be seen in Figure 8, shows that the technician personality is in 15 cases confused with the expert personality and 6 times with the helpful assistant.

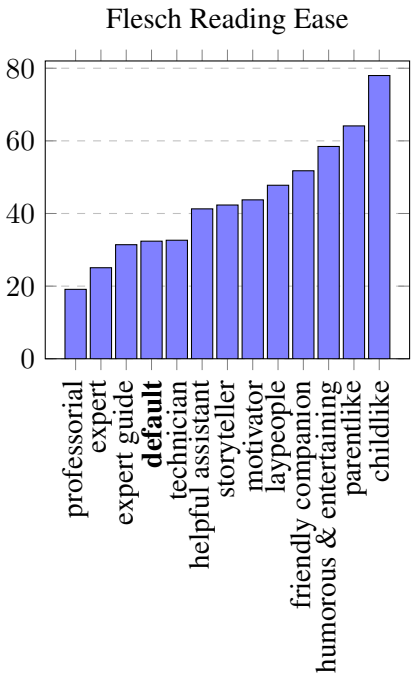

Figure 4: Flesch's Reading Ease (FRE) values of the personality styles. FRE ranges from 0 to 100, with higher values indicating better readability.

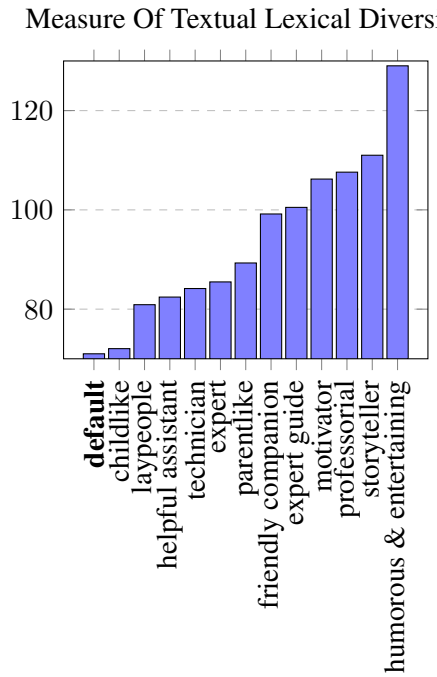

Figure 5: Measure Of Textual Lexical Diversity (MTLD) values of the personality styles, with higher values indicating higher lexical richness.

The expert personality is confused in 8 cases with the technician, in 7 cases with the helpful assistant and in 5 cases with ChatGPT's default answer. The laypeople personality is in 8 cases confused with the helpful assistant, in 6 cases with ChatGPT's default answer and in 2 cases with the technician. This shows very clearly that these personality styles are very similar to each other and may use the same phrases.

5.8 Extracting Important Words with SHAP

The most important words for the prediction of personalities with the personality classifier were extracted with the SHAP XAI framework. A global explanation approach was chosen using summarized text explanations. The extracted top 10 words for each personality style are shown in Table 1.

6 Discussion

The lengths of personality styles generated by ChatGPT are as expected. A child-like personality is expected to have rather short texts, since many details are omitted. In contrast, the storyteller personality generates very long texts because whole stories are formulated with embellishments. An explanation for why the professional/technical personality styles tend to generate longer texts is that they contain more factual content, are described in more detail, and contain code. The average number of stopwords, the average number of named entities, Flesch's Readining Ease and the sentiment score can be explained by the categorization into professional/technical and non-professional/non-technical personality styles. Personality styles that are not so technical/professional but more human-oriented, take into account other dimensions besides the factual content, e.g. include motivational phrases, descriptive texts, entertaining passages, etc. This leads to the increased number of stopwords. At the same time, due to a different focus (motivation, entertainment, etc.) the facts are reduced, which explains the reduced number of named entities. As professionalism and technical focus increase, wording becomes more complex, which degrades readability, explains the falling Flesch Reading Ease, and aligns with expectation. ChatGPT's ability of being able to switch appropriately between professional/technical and emotional/human-centered formulation of texts through the presented prompts fits very well with the requirement for adaptive human-centered XAI interfaces to adapt to user needs and to provide information in different preparations and different granularities.

The distribution of MTLD indicates that the lexical richness of personality styles has no simple ex-

Personality	Most Important Tokens
Default	“Certainly”, “steps”, “bu”, “B”, “template”, “:”, “Firstly”, “you”, “Python”
Child-like	“!”, “Oh”, “.”, “and”, “,”, “”, “friend”, “Hi”, “or”
Parent-like	“Parent”, “.”, “?”, “Child”, “,”, “sweetie”, “a”, “!”, “sweetheart”
Professorial	“Indeed”, “pleased”, “scholarly”, “.”, “,”, “Thank”, “Colleagues”, “intriguing”, “excellent”
Friendly Companion	“absolutely”, “delighted”, “Hey”, “,”, “fascinating”, “Oh”, “course”, “!”, “assist”
Expert Guide	“walk”, “guide”, “Welcome”, “welcome”, “Certainly”, “Allow”, “.”, “!”, “explorer”
Storyteller	“Once”, “nestled”, “!”, “expertise”, “time”, “tale”, “upon”, “a”, “qui”
Helpful Assistant	“course”, “assist”, “helpful”, “Certainly”, “assistant”, “,”, “.”, “!”, “As”
Humorous/Entertaining	“!”, “Oh”, “jolly”, “Well”, “?”, “,”, “comrades”, “.”, “Don”
Motivator	“!”, “welcome”, “Absolutely”, “Remember”, “inspire”, “amazing”, “you”, “,”, “friend”
Technician	“technician”, “Technician”, “assist”, “,”, “Sure”, “.”, “and”, “V”, “X”
Laypeople	“Certainly”, “Sure”, “”, “Singapore”, “Remember”, “,”, “guide”, “memory”, “appropriate”
Experts	“Certainly”, “examples”, “Experts”, “expert”, “,”, “experts”, “subjective”, “I”, “Title”

Table 1: Most importants words for each personality extracted from the personality classifier using XAI framework SHAP.

planation and that a separate explanation for each style needs to be found in future work. However, for the personality styles motivator, professorial, storyteller, and humorous/entertaining, which are in the upper range of values, the result is at least plausible, since many unique words for these can be expected. If the scores of the MTLD are compared with the Flesch Reading Ease, there are cases, such as the professorial personality, where the texts are very difficult to read and have a high lexical richness. A random examination of the data shows that the texts are indeed particularly written in a sophisticated way. There are other cases, such as ChatGPT's default personality, which is also difficult to read, but has a low lexical richness. Possible explanations could be a more complex sentence structures, advanced vocabulary and redundancies.

Sentiment analysis showed that for personality styles for which positive sentiments are expected (motivator, storyteller, friendly compation, humorous/entertaining) the texts were adequately simulated by ChatGPT. This is due to an increased use of words with positive sentiment. The ability of ChatGPT to provide an explanation to a human in an appropriate sentiment is a strong feature for human-centric XAI interfaces. Explanations conveyed with an appropriate sentiment seem more natural and it can be presumed that trust is increased. The analysis of the personality classifier shows that there are personality styles that are very distinct from each other. The reason are words and phrases typical for the personality. This finding becomes even clearer by analyzing the words extracted with SHAP. For example, the professorial style uses sophisticated words such as *“indeed”*, *“pleased”*, *“intriguing”*, and *“excellent”*, which are very appropriate for this style. The friendly companion uses very friendly words and the expert guide personality is very welcoming. It can also be seen that some styles reference themselves. For example, the helpful assistant uses the words *“assist”* and *“assistant”*, the technician uses the word *“technician”*, and the expert style uses the words *“expert”* and *“experts”* frequently. This is because ChatGPT generates phrases like *“Ok I will give the answer like a technician”* at the beginning of the answer.

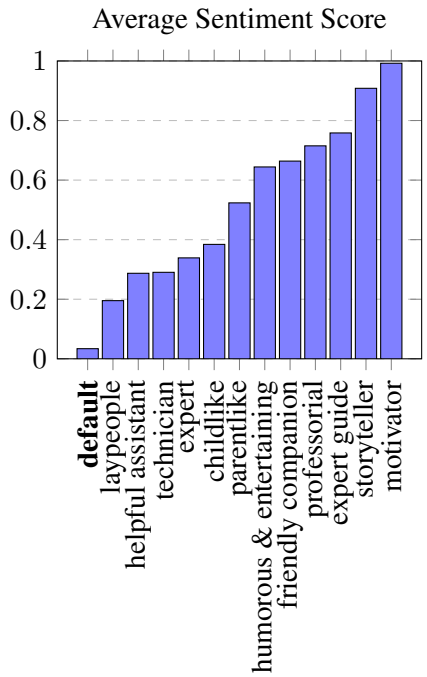

Figure 6: Average sentiment score for each personality style.

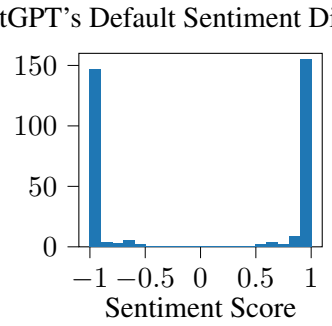

Figure 7: Distributions of sentiment scores for ChatGPT's default personality.

7 Conclusion

The analysis showed that the personality styles simulated by ChatGPT are largely in line with requirements and expectations and can be used in adaptive human-centric XAI interfaces. ChatGPT is able to generate texts of appropriate length with a number of facts adapted to the personality. A clear distinction could be made between professional/technical and more emotional/human-centered personalities, which is of great importance for adaptive human-centered XAI interfaces. The use of stopwords and the readability of the texts behave according to the personality styles. ChatGPT is able to create the appropriate sentiment of a text and words and phrases are used that match the personalities. This was shown by training and analysis of a personality classifier and application of SHAP explanations.

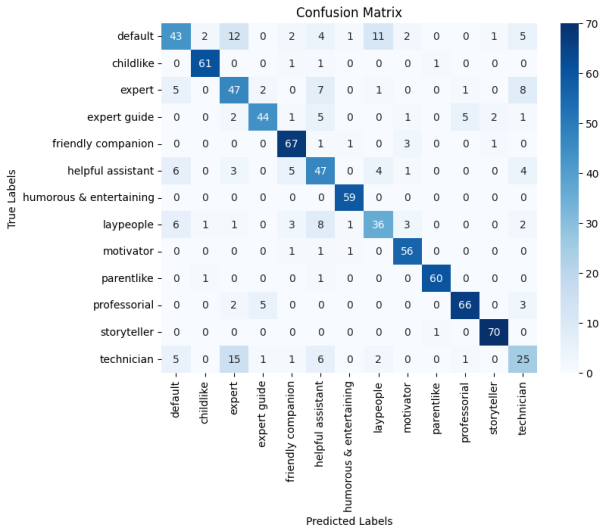

Figure 8: Confusion matrix of the personality classifier.

8 Limitations

While ChatGPT has demonstrated the ability to effectively replicate diverse personality styles in textual analysis, the congruence of these simulations with real human perception remains unestablished. In order to provide clarity on this issue, it is necessary to examine how the simulated personalities affect the individual. Also, whether the personality styles can positively influence the important attributes of XAI interfaces, including trust, fairness and transparency, must be shown in future studies.

References

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of Applied Psychology*, 32(3):221–233.

Denis J. Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107:65–81.

Sophie F. Jentzsch, Sviatlana Höhn, and Nico Hochgeschwender. 2019. Conversational interfaces for explainable ai: A human-centred approach. In *Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds) Explainable, Transparent Autonomous Agents and Multi-Agent Systems. EXTRAAMAS 2019. Lecture Notes in Computer Science()*, volume 11763, pages 77–92, Cham. Springer International Publishing.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea

Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *Computing Research Repository, arXiv:2304.11633. Version 1.*

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA. Curran Associates Inc.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human- and AI-generated texts: Investigating features for ChatGPT. In *Artificial Intelligence in Education Technologies: New Development and Innovative Practices*, pages 152–170, Singapore. Springer Nature Singapore.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text. *Computing Research Repository, arXiv:2301.13852. Version 1.*

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Computing Research Repository, arXiv:1910.01108. Version 4.*

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mildred C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, NED - New edition edition, volume 26. University of Minnesota Press.

Joan Torruella and Ramon Capsada. 2013. Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, 95:447–454.

Thorsten Zylowski. 2022. Study on criteria for explainable AI for laypeople. In *Proceedings of the Second International Workshop on Explainable and Interpretable Machine Learning (XI-ML 2022) co-located with the 45rd German Conference on Artificial Intelligence (KI 2022)*, Trier (Virtual), Germany. CEUR Workshop Proceedings.