



Degree Project in Computer Science and Engineering

Second cycle, 30 credits

Towards Automatic Generation of Personality-Adapted Speech and Emotions for a Conversational Companion Robot

ALESSIO GALATOLO

Towards Automatic Generation of Personality-Adapted Speech and Emotions for a Conversational Companion Robot

ALESSIO GALATOLO

Master's Programme, Machine Learning, 120 credits
Date: June 27, 2022

Supervisor: Katie Winkle

Examiner: Iolanda Dos Santos Carvalho Leite

School of Electrical Engineering and Computer Science

Swedish title: Mot Automatisk Generering av Personlighets Anpassade Tal och
Känslor för en Samtalskunnig Sällskaps Robot

Abstract

Previous works in Human-Robot Interaction have demonstrated the positive potential benefit of designing highly anthropomorphic robots. This includes physical appearance but also whether they can express emotions, behave in a congruent manner, etc. This work wants to explore the creation of a robot that is able to express a given personality consistently throughout a dialogue while also manifesting congruent emotional expressions.

Personality defines many aspects of the character of a person and it can influence how one speaks, behaves, reacts to events, etc. Here, we only focus our attention on language and on how it changes depending on one particular personality trait, the extraversion. To this end, we tested different language models to automate the process of generating language according to a particular personality. We also compared large language models such as GPT-3 to smaller ones, to analyse how size can correlate to performance in this task.

We initially evaluated these methods through a fairly small user study in order to confirm the correct manipulation of personality in a text-only context. Results suggest that personality manipulation and how well it is understood highly depend on the context of a dialogue, with a more ‘personal’ dialogue being more successful in manifesting personality. Also, the performance of GPT-3 is comparable to smaller models, specifically trained, with the main difference only given in the perceived fluency of the generations.

We then conducted a follow-up study where we chose to use a robot that is capable of showing different facial expressions used to manifest different emotions, the Furhat robot. We integrated into the robot the generations from our language models together with an emotion classification method that is used to guide its facial expressions. Whilst the output of our models did trigger different emotional expressions, resulting in robots which differed both in their language and nonverbal behaviour, resultant perception of these robots’ personality only approached significance ($p \sim 0.08$). In this study, GPT-3 performed very similarly to much smaller models, with the difference in fluency also being much smaller than before. We did not see any particular change in the perception of the robots in terms of likeability nor uncanniness.

Keywords

Personality, Emotions, Human-Robot Interaction, Machine Learning, Large Language Models, Text-style transfer, GPT-3, STRAP

Sammanfattning

Tidigare arbeten inom Människa-robotinteraktion har visat den positiva potentiella fördelen med att designa mycket antropomorfa robotar. Detta inkluderar fysiskt utseende men också huruvida de kan uttrycka känslor, bete sig på ett kongruent sätt, etc. Detta arbete vill utforska skapandet av en robot som kan uttrycka en given personlighet konsekvent under en dialog samtidigt som den manifesterar kongruenta känslomässiga uttryck.

Personlighet definierar många aspekter av en persons karaktär och den kan påverka hur man talar, beter sig, reagerar på händelser etc. Här fokuserar vi vår uppmärksamhet endast på språket och på hur det förändras beroende på ett särskilt personlighetsdrag, extraversion. För detta ändamål testade vi olika språkmodeller för att automatisera processen att skapa språk enligt en viss personlighet. Vi jämförde även stora språkmodeller som GPT-3 med mindre, för att analysera hur storlek kan relatera till prestanda i denna uppgift.

Vi utvärderade inledningsvis dessa metoder genom en mindre användarstudie för att bekräfta att personligheten kan manipuleras på rätt sätt i en textbaserad kontext. Resultaten tyder på att personlighetsmanipulation och hur väl den förstår i hög grad beror på sammanhanget i en dialog, där en mer ‘personlig’ dialog är mer framgångsrik när det gäller att manifesterar personlighet. Prestandan hos GPT-3 är också jämförbar med mindre modeller, specifikt tränade på en uppgift, där den största skillnaden var i den genererade textens upplevda flyt.

Vi gjorde sedan en uppföljningsstudie där vi valde att använda en robot som är kapabel att visa olika ansiktsuttryck och därigenom kapabel att manifester olika känslor, Furhat-roboten. Vi integrerade talet som genererades från våra språkmodeller i roboten tillsammans med en känsloklassificeringsmetod som används för att styra dess ansiktsuttryck. Medan resultatet av våra modeller framkallade olika känslomässiga uttryck, vilket resulterade i robotar som skilde sig åt både i språk och icke-verbal kommunikation, närmade sig endast den resulterande uppfattningen av dessa robotars personlighet signifikans ($p \sim 0.08$). I denna studie presterade GPT-3 mycket likartat med mycket mindre modeller, med skillnaden i flyt också mycket mindre än tidigare. Vi såg ingen speciell förändring i uppfattningen av robotarna när det gäller sympati eller obehaglighet.

Nyckelord

Personlighet, Känslor, Människa-robotinteraktion, Maskininlärning, Stora Språkmodeller, Överföring av text, GPT-3, STRAP

Acknowledgements

Throughout my journey at KTH Royal Institute of Technology (KTH), there have been many people that have been helpful and that supported me all the way to my graduation.

I would first like to thank my supervisor, Katie Winkle, that has always been inspiring and helpful. She has guided me throughout the definition of this work's topic making me explore and understand my interests without ever limiting my thought process. Thanks to her, I was able to create the work of which I am most proud and that will end up positively influencing the years to come. I would also like to thank her for all the time she has spent giving me feedback that was always right and on-point.

I would also like to thank the Social Robotics group at KTH and my examiner in particular, Iolanda Leite, for all the feedback and direction they have given me.

Next, I want to thank all of my family and my mother in particular for all the efforts made to support me in every choice I've made. Most of my journey would not have been possible without my mother's emotional and psychological support that she has always given me alongside all of her unconditional love.

I would then like to thank all of my friends that have always been of great inspiration and guidance for my journey. For pushing me to always aim for the best I could achieve, even against big odds.

Finally, I cannot conclude this without mentioning my best friend and partner in life. Thanks to her I found the energy and motivation that I needed during my hardest times. Thanks to her I was able to expose myself and risk falling knowing she would always be there to catch me. She was especially supportive throughout the duration of this work, always listening to all of my blabberings and always showing interest even when listening to the same topic over and over.

Stockholm, July 2022

Alessio Galatolo

Contents

1	Introduction	1
1.1	Research questions	2
1.2	Ethical and environmental considerations	3
1.3	Purpose and goals	3
1.4	Structure of the thesis	4
2	Background	5
2.1	Personality in social sciences	5
2.1.1	Big 5 Framework	5
2.1.2	MBTI framework	6
2.2	Personality classification through language	7
2.3	Generating language according to personality	8
2.3.1	Statistical methods	8
2.3.2	Text style transfer	9
2.3.3	Direct generation	10
2.4	Large language models	10
2.4.1	Ethical issues	10
2.4.2	Zero to Few-shot learning	11
2.5	Human-Robot Interaction	11
2.5.1	Personality in HRI	13
2.5.2	Emotion portrayal in HRI	13
2.5.3	At the intersection of (robot) personality and emotion .	14
2.6	Summary	15
3	Methods	17
3.1	Datasets	17
3.2	Text style transfer models	19
3.3	GANs with language models	20
3.3.1	Implementation details	21

3.3.2	Training	21
3.3.3	Results on Yelp dataset	22
3.3.4	Results on Personality datasets	23
3.3.4.1	Generator performance	24
3.3.4.2	Classifier performance	25
3.4	STRAP	27
3.4.1	Training and implementation details	28
3.4.2	Results	28
3.4.2.1	Nucleus paraphrase mode	29
3.4.2.2	Nucleus mode	31
3.5	GPT-3	31
3.5.1	Implementation details	33
3.5.2	Results	34
3.6	Personality transfer summary	34
3.7	Emotion recognition from text	36
3.7.1	Word-based approach	36
3.7.2	RoBERTa-large fine-tuned	36
3.7.3	Results	37
4	Experiments	38
4.1	Model evaluation study	38
4.1.1	Study design	40
4.1.2	Hand-crafted dialogues	41
4.1.3	Auto-generated dialogues	42
4.1.4	Results	44
4.2	User study	46
4.2.1	Emotions and personality manifestation	47
4.2.2	Research questions and hypothesis	48
4.2.3	Study design and measures	48
4.2.4	Results	51
4.2.4.1	RQ2 - Manipulation check	51
4.2.4.2	RQ2 - Difference in fluency and performance	53
4.2.4.3	RQ2 - Emotional manifestation	54
4.2.4.4	RQ3 - Perception of uncanniness	55
4.2.4.5	RQ4 - Likeability	56
5	Discussion	58
5.1	Initial findings	58
5.1.1	Understanding personality from single sentences is hard	58

5.1.2	Bigger models are better at extending sentences but not at paraphrasing	59
5.2	Research questions answered	59
5.2.1	Personality manifestation depends on the topic	59
5.2.2	Personality manifestation in robotic speech is (almost) significant	60
5.2.3	GPT-3 is more fluent but not better performing than STRAP	61
5.2.4	Likeability and personality matching effects are too little to measure	62
5.2.5	Emotional manifestation depends on personality (but maybe not)	62
6	Conclusions	63
6.1	Future work and limitations	64
References		67
A	STRAP extensive results	75
B	Complete dialogues	80
B.1	GPT-3 generated dialogues	80
B.2	Other hand-crafted dialogues	81

List of Figures

1.1	A flowchart explaining the pipeline of our system from a given dialogue to its animation in Furhat.	4
2.1	Figure of the appearance of the robots mentioned in the related works.	12
3.1	Training accuracy on the Yelp dataset. The classifier (D) accuracy is in blue quickly approaching 1, while the generator's (G) is in orange, quickly approaching 0.	23
3.2	Training accuracy on the MBTI (above) and Essays (below) dataset. The training curve is in the background while we highlight its smoothed version for better reading. The line in the middle of the graph splits is the boundary of the pretraining. The classifier (D) accuracy is in blue quickly approaching 1, while the generator's (G) is in orange, quickly approaching 0.	25
4.1	Results of the models across all the dialogues in fluency (left) and personality overall (right). In both higher is better.	45
4.2	Extraversion rating for each model×personality for dialogue 1 (top left), dialogue 2 (top right) and dialogue 3 (bottom). Better as the Ext score gets higher than the Int score. No error bars are displayed to favour readability.	46
4.3	Two screenshots of the videos used for the study.	49
4.4	Distribution plot of the age of the participant in our final study.	51
4.5	Plots with the ascribed extraversion of the dialogues on a scale from 0 to 6.	52
4.6	Plots with the ascribed fluency of the dialogues (left) and the personality performance (right) on a scale from 0 to 6.	53

4.7	Progression of the emotional manifestation intensity throughout the dialogue.	54
4.8	Plot with the ascribed RoSAS warmth measures as a proxy for emotional manifestation on a scale from 0 to 4.	55
4.9	Plots with the ascribed uncanniness of the dialogues (left - lower is better) and the Godspeed Anthropomorphism (right - higher is better). Both are on a scale from 0 to 4.	56
4.10	Plot with the ascribed Godspeed Likeability on a scale from 0 to 4.	57

List of Tables

2.1	Examples of some adjectives associated with each personality trait [23].	6
2.2	The correlation between the MBTI and the Big 5 framework for personality.	7
2.3	Examples of few-shot learning capabilities of GPT-3. The <i>prompt</i> to the model is marked in <i>italic</i> , the style of the sentence is marked in bold	12
2.4	Summary of the related work on emotions and personality in Human-Robot Interaction (HRI) with the relevant features present.	15
3.1	List of the datasets collected and used in this thesis and their attributes.	19
3.2	Some of our test results on the Yelp dataset and a comparison with the original implementation. Changes in the output are marked in bold	24
3.3	GAN-based model generations when trained on the Essays and MBTI dataset. First two sentences are from PERSONAGE dataset, second two from Friends dataset. Changes in the output are marked in bold	26
3.4	Comparison of personality classification accuracy of the GAN-based models when trained on the Essays or MBTI dataset.	27
3.5	Some outputs of the STRAP model, varying on the dataset it was trained on and the top-p value. The input is “Hello and welcome! My name is Brian, I am here to be your robotic companion and your very own personal assistant.”. The output mode is nucleus paraphrase.	30
3.6	Some outputs of the STRAP model with same input and settings of Table 3.5 with nucleus as the output mode.	32

3.7	Some results from GPT-3’s personality transfer generations.	35
3.8	Comparison of the emotion recognition models. The emotions reported is the primary emotion recognised.	37
4.1	The three dialogues used for the pilot study.	39
4.2	The questions used in the pilot study alongside what they measure. The questions were given in random order and participants were prompted with: “I see the speaker as someone who...”.	41
4.3	List of possible manipulations of a dialogue to express different personalities.	42
4.4	An example of the manipulation of a dialogue to convey an extravert or introvert personality.	43
4.5	The questions used in the second study in addition to those in Table 4.2. The participants were prompted with: “Please indicate the extent you think the robot is...”	50
A.1	Extensive results of STRAP trained on the Essays dataset. The output mode is nucleus.	76
A.2	Extensive results of STRAP trained on the Essays dataset. The output mode is nucleus paraphrase.	77
A.3	Extensive results of STRAP trained on the MBTI dataset. The output mode is nucleus.	78
A.4	Extensive results of STRAP trained on the MBTI dataset. The output mode is nucleus paraphrase.	79
B.1	List of possible manipulations of a dialogue to express different personalities.	82
B.2	The manipulation done to the second dialogue to convey an extravert or introvert personality.	83
B.3	The manipulation done to the third dialogue to convey an extravert or introvert personality.	84

List of acronyms and abbreviations

CV	Computer Vision
GAN	Generative Adversarial Network
HHI	Human-Human Interaction
HRI	Human-Robot Interaction
KTH	KTH Royal Institute of Technology
LLM	Large Language Model
MBTI	Myers–Briggs Type Indicator
ML	Machine Learning
NLP	Natural Language Processing
NLU	Natural Language Understanding
STRAP	Style Transfer via Paraphrasing
TIPI	Ten Item Personality Inventory

Chapter 1

Introduction

Human-Robot Interaction (HRI) is concerned with the study of how humans interact with robots. Among the many goals prefixed by this field, there is that of creating an interaction that is as natural as possible. Ways of achieving this range from making the robot's appearance more human-like as well as integrating different kinds of social intelligence into the robot. Having a more natural interaction can help in the perception of the robot in terms of how much the user likes it or how credible and persuasive it is considered [1, 2, 3, 4, 5, 6].

There have been multiple attempts at exploring the influence of integrating into a robot a personality [7, 8] or emotional capabilities [9, 10, 11]. However, it is very rare to find a work that implements both extensively, with most of the studies that consider both only focusing on what (small) influence can personality have on the generated emotions [12, 13]. Further, among the many ways to express personality e.g. body language, prosody, speech, etc. only a few have been explored and incorporated into robots. Many works focus only on body language [7, 12] while there is evidence in social sciences literature that personality has a significant influence on other aspects of life e.g. language [14, 15, 16, 17, 18]. Language is defined as the choice of words, their order and how they are used to formulate sentences in order to express a certain or multiple ideas. For example, a very extroverted person is generally more talkative and uses less complex sentences than an introverted one [14, 15].

However, when integrating these social traits into robots, it is necessary to consider that both the robot's appearance and the inclusion of emotions/personality have been associated with the perception of uncanny feeling in the user interacting with it [19, 20, 21]. The perception of uncanniness can manifest

itself as a discomfort in the user as the robot is perceived to have some human features but as not being enough human-like to appear natural.

Implementing a personality-adapted language into a robot can be difficult and resource expensive if to be done manually and for each interaction of the robot. Therefore, considering an experimental scenario, it can definitely benefit from the automation of this process. When performing an experiment that is not strictly guided, multiple can be the path that the participant decides to follow and having a human dynamically adapt the output of the robot to that of the situation while also following a pre-determined personality is not optimal. If, on the other hand, a method for performing the same procedure automatically was available it would decrease significantly the possibility of human error. Following this goal is also motivated by the absence of works that focus on this aspect with some attempts at generating appropriate language that could only be found outside of robotics [22, 23].

When developing an automated method for generating language, using different Machine Learning (ML) and Natural Language Processing (NLP) techniques, it is important to evaluate the method not only on its performance but also on its ethical and environmental impact. This is motivated by recent concerns that see the growth of language models as a threat to the environment [24]. The current increase in performance of NLP models is generally given by an increase in their size which, however, also correlates with the resources needed for their training and the subsequent use among many other ethical issues [24].

1.1 Research questions

During the progress of this thesis we aim at answering all of the following questions:

RQ1 How does the size of a language model relate to its performance in generating text that correctly manifests a personality?

RQ2 (How) does multi-modal robot delivery of speech influence the performance of different-sized models in terms of fluency, personality and emotion manifestation? Does fluency and personality manifestation vary compared to perceptions of the pure text output?

RQ3 (How) does the perception of uncanniness vary depending on the personality of the dialogue and the personality of the participant? Does this vary across the different language models we evaluate?

RQ4 (How) does the likeability of a robot change depending on the personality of the dialogue and the personality of the participant? Does this vary across the different language models we evaluate?

1.2 Ethical and environmental considerations

This thesis treats ethical and sustainability considerations for the various methods used. Many ML methods require intensive training using a lot of resources for a lot of time. Further, depending on the application of such methods, also inference time may become a burden in terms of resources needed. This poses questions not only on the accessibility of such methods* but also on their environmental impact. For this reason, in this thesis, each method used will be evaluated not only on its performance but also on its training time, inference time as well as the resources used.

1.3 Purpose and goals

The news value from this work will be extensively described and compared to the current state-of-the-art in [Chapter 2](#) and is mainly related to the automation of personality incorporation into robots. We propose the use of language models to achieve personality expression through language without any intervention from researchers. This is very novel as very few works focus on this with none being usable for general-purpose robots.

We will also focus on automatically extracting emotions from the language used by the robot (language that was generated by our other automatic techniques) and we will look to manifest those emotions through facial expressions in a social robot. We believe the incorporation of emotions directly extracted from personality (through text in our case) has never been explored before. The relation between personality and emotions is often not considered when incorporating one or the other in robots when in reality the emotional manifestation highly depends on personality.

We propose the pipeline shown in [Figure 1.1](#) to deliver all of our methods into a robot for seamless integration in social experiments.

In order to validate our technical work, we will then conduct an initial study aimed at comparing the different ML techniques explored and selecting

* Some of them are effectively unreplicable outside of large tech companies.

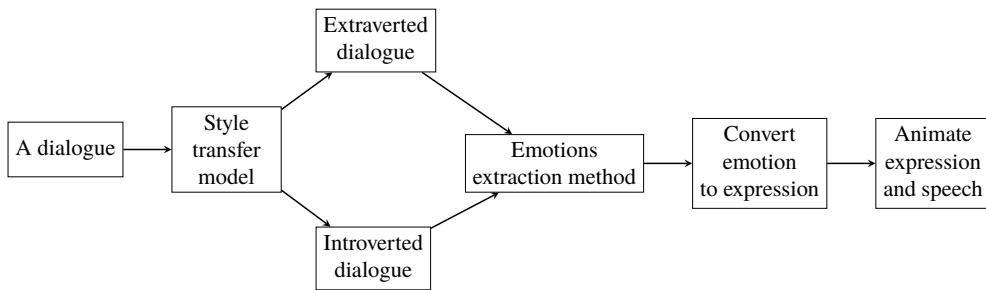


Figure 1.1: A flowchart explaining the pipeline of our system from a given dialogue to its animation in Furhat.

the best method for a follow-up study. In this second study, we will evaluate the effects of incorporating personality and emotions on measures of how much the robot is liked by the user and how much uncanny it is perceived.

Finally, the results of these studies should be especially relevant in the context of creating more believable robots for improved engagement and persuasiveness.

1.4 Structure of the thesis

[Chapter 2](#) presents relevant background information about Personality, Emotions, NLP and HRI. [Chapter 3](#) presents the ML models used and investigated in this thesis together with their results and performance. [Chapter 4](#) presents both studies that have been conducted alongside the objective results extracted from those studies. [Chapter 5](#) takes each result of the thesis (both from the technical work and from the studies) and discusses the reasons for such result and its consequences. Finally, [Chapter 6](#) summarises this whole work and introduces some ideas for possible extensions.

Chapter 2

Background

This thesis touches on the topics of emotions, personality and how they can be incorporated into robots, with a particular focus on language. For this reason, this chapter will first give an overview of language models and all such topics as well as motivating the work that follows. Additionally, this chapter gives an overview of the approaches that could be taken with this thesis work.

2.1 Personality in social sciences

Personality is the set of characteristics that define one person. Personality influences the friends-making process, one's approach to problems, the reaction to different news, etc. [25, 26, 27]. It can also be considered an invariant in the life of a person, where one can expect very little change in it after the development age of a person [28, 29]. Understanding one's own personality and those of the surrounding people is claimed to help improve personal and interpersonal relationships as well as one's career* [30, 26, 27].

2.1.1 Big 5 Framework

The most commonly used framework (in academics) for personality is the Big Five [31, 32] or Factor Five [33]. Big Five and Factor Five were presented independently but got eventually merged and referred to interchangeably. This framework groups the population using statistical tools resulting in the 5 factors introduced. These factors were later associated different values [34] through different studies. These factors aim to describe personality based on the 5 traits of Openness (to experience), Conscientiousness, Extraversion,

* <https://www.16personalities.com/>

	Low	High
Openness	narrow-minded, conservative, ignorant, simple	creative, intellectual, imaginative, curious, cultured, complex
Conscientiousness	disorganized, impulsive, unreliable, careless, forgetful	competent, disciplined, dutiful, achievement striving, deliberate, careful, orderly
Extraversion	shy, quiet, reserved, passive, solitary, moody, joyless	warm, gregarious, assertive, sociable, excitement seeking, active, spontaneous, optimistic, talkative
Agreeableness	unfriendly, selfish, suspicious, uncooperative, malicious	trustworthy, friendly, considerate, generous, helpful, altruistic
Neuroticism	neurotic, anxious, depressed, self-conscious, oversensitive, vulnerable	calm, even-tempered, reliable, peaceful, confident

Table 2.1: Examples of some adjectives associated with each personality trait [23].

Agreeableness and Neuroticism* (OCEAN). These factors can be interpreted as being either categorical or continuous e.g. being extrovert or introvert vs having different degrees of the factor from extrovert to non-extrovert. Table 2.1 shows some adjectives associated with people with the given trait from low to high e.g. with the extraversion trait, low points to introverted people while high to extroverted.

2.1.2 MBTI framework

Another framework for personality is the Myers–Briggs Type Indicator (MBTI), this framework, compared to the previous is a lot less used in academics (where it is heavily criticised [35, 36, 37, 38, 39]) but received a lot more success in the commercial sector where different websites propose solutions to both business and personal growth based on it†. MBTI aims at classifying each individual into a single type among 16 possible ones. Each type is formed by 4 different traits that are strictly categorical and are Favorite world, Information, Decisions, Structure‡. Each of these traits has two categories in which people fall into e.g. Favorite world can be either Introvert or Extrovert. By the analysis of these traits [39], it is possible to assess a correlation between this framework and the Big 5 framework where

* Sometimes also referred to as Emotional (in)stability.

† <https://www.themyersbriggs.com/>,

‡ <https://www.16personalities.com/>

‡ <https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/>

MBTI trait	Big 5 trait	Low	High
Favorite world	Extraversion	Introvert	Extrovert
Information	Openness	Sensing	Intuition
Decisions	Agreeableness	Thinking	Feeling
Structure	Conscientiousness	Perceiving	Judging

Table 2.2: The correlation between the MBTI and the Big 5 framework for personality.

each trait is correlated to one in the Big 5 with the exception of Neuroticism which has no counterpart. [Table 2.2](#) shows the traits of the framework, what they describe and which trait in the Big 5 they correlate to. Criticism of this framework typically refers to the absence of the Neuroticism trait [35, 36], the inconsistency in results (same people getting different results at different times) [37] and issues around different traits correlating with each other [38, 39].

2.2 Personality classification through language

Different works explore the understanding of personality in relation to the recent availability of Big Data. In most experiments personality is assessed through self-report questionnaires where participants are asked to rate on different scales their own characteristics and, although they are a standard in the field, they sometimes only reflect people's self-portrayal rather than their actual nature. Boyd and Pennebaker [40] question the reliability of self-report questionnaires for personality assessment proposing and motivating a paradigm shift towards a Big Data approach to assessing personality based on language.

Following from this, Ríssola, Bahrainian, and Crestani [41] develop a method based on Capsule Neural Networks [42] to build a model that is able to correctly recognise personality from language in the majority of cases. The personality traits are interpreted as categorical.

Other methods, developed after this, mostly rely on language models [43, 44, 45]. This can be done by taking a pretrained model (a model with pre-existing knowledge of language) and fine-tuning it for this personality classification task. All of these methods interpret personality traits as binary.

The performances of the presented methods largely vary on the dataset

used going from just over 60% [45], to about 70% [43] to a little below 80% [44, 45]. However, the pre-processing done to the dataset is often not well described or not described at all. All of the datasets considered in these works contain multiple sentences from the same person and it is not very clear if the model is trained on single utterances or on the whole corpus relative to that person. Reasons for using multiple sentences at once are motivated by the subtle influence of personality on language that may not be discerned from a single, generic utterance but that may require a whole dialogue instead. On the other hand, many language models often impose a limit on the length of a sentence taken as input and sometimes perform better on shorter sentences [46]. Further, some of the datasets used are highly unbalanced in some of the traits and, while some works adopt sophisticated balancing techniques [43], others lack the consideration of this issue [45].

In summary, literature shows evidence for successful classification of personality from text but the results are still mixed and the generalisation ability e.g. to different datasets and contexts is often glanced over. This suggests that, in an experimental setting, an automated recognition system may be used but should always be paired with manual recognition.

2.3 Generating language according to personality

Generating language that follows a certain personality, style or linguistic variation, in general, is quite often a niche topic in NLP where major research efforts prefer to focus on the content of the sentence, its naturalness or its fluency. For the goal of generating language that is influenced by (some) personality traits, we managed to identify 3 different approaches.

2.3.1 Statistical methods

The first is based on statistical NLP and is the only method (known to the author) that has been used in the personality context. Mairesse and Walker [22] propose a method called PERSONAGE (later renamed PERSONAGE-RB where RB stands for Rule-Based) to generate language according to different personality traits. In order to achieve this, they use linguistic parameters such as verbosity and length of the sentence, well-grounded in literature, to manipulate a text that can be attributed different personality traits. They focus their work on the extraversion trait motivating this choice on its broader impact

on the linguistic variables, planning to extend the method to other personality traits. The authors use this method to generate utterances about restaurant recommendations and comparisons in New York City modelled following different extraversion levels on a half-point scale from 1 to 7. They evaluate their model using three expert judges and confirm the desired parametrization of the personality trait in 89.2% of the cases.

The same authors expand this method to the whole Big 5 framework in [23] where they split the part responsible for generating the utterance from the parameters (PERSONAGE base generator) and add a part responsible for estimating the parameters from the given personality traits (PERSONAGE-PE i.e. parameters estimation). They finally evaluate this method using human judges reporting mixed results. While they succeeded in finding a correlation between the participants' ratings and the target ratings in some of the traits, they fail for some of the others.

2.3.2 Text style transfer

Text style transfer is a problem in NLP with the goal of changing the ‘style’ of a given sentence. It generally concerns binary styles such as polite vs impolite, positive vs negative*, etc. [47] although there are also some works exploring multi-category style transfer [48, 49]. In most cases, the goal is to keep the semantic of a sentence intact while varying the language used to convey different styles.

The problem of text style transfer can also be divided depending on the data that is available. An easier version of the problem is given by the availability of a parallel corpus i.e. a corpus where, for every utterance, there is a version of it for each style. The other problem is defined when there is no availability of parallel data, which is most often the case as such datasets are more expensive to obtain or even unobtainable in some cases[†]. Since we could not find any dataset with parallel data for our problem of personality style transfer, here, we will only treat the case of non-parallel corpus.

For this problem, different models have been proposed that use different approaches: disentanglement, prototype editing and pseudo-parallel corpus construction are some of the most common ones [47]. The first follows an encode-decode approach where the text in the original style is first encoded and then decoded (after some manipulation) into the target style. The second aims at detecting the words responsible for one style rather than the other and

* In the context of recommendations or reviews. [†] For example, changing the style of a text from one famous author to another when both of them are dead.

substituting the ones in the original style with those from the target one. The third is mostly used in machine translation and trains two models where each one goes from one style to the other. These models are trained by using the output of one as the objective of the other and vice-versa, simultaneously.

2.3.3 Direct generation

Another approach is to have a model that is able to directly generate text in a given context while also accounting for the personality traits. For example, in a question answering scenario, the model would take as input both the question and the personality and would generate a text tailored for that particular personality. A way of achieving this would be to fine-tune a pretrained language model with an appropriate dataset.

2.4 Large language models

When approaching NLP, whether it is for question answering, text style transfer, etc. it is important to mention how most of the peak performances can be obtained through Large Language Models (LLMs). LLMs are language models that have several billion parameters and are trained on a very large amount of data ranging from Wikipedia's corpus to data crawled all across the web to large collections of books. Due to both their architecture and the amount of data they are training on, these models require enormous computational power and are often unavailable to the general public. Because of their very resource-intensive nature, they are generally proposed by large companies with large availability of computational resources such as OpenAI [50], Google [51], Microsoft and NVIDIA [52].

One of the benefits of these models is their adaptability to very different tasks and the results they can obtain in many Natural Language Understanding (NLU) tasks. However, they also have some restrictions and their recent growth has triggered some concerns about their use and applications.

2.4.1 Ethical issues

A first critique that can be made to these models is the environmental impact that follows from their training and use. As the models keep growing in size, even the inference requires specialised hardware and a time cost that when multiplied by the number of uses may become considerable. Further, most research is focusing on *just* expanding on the number of parameters

often without questioning their utility or whether the final model is actually better in understanding natural language. It is common that a new model simply reports an improvement in the set benchmarks without considering this may be due to a higher knowledge of the language solely given by a larger training corpus [24]. There are, however, some works that tackle the environmental problem by trying to reduce the number of parameters while maintaining similar performances [53] or by improving the training efficiency in general [51].

Some of these models have also been shown to be incorporating different kinds of stereotypes and biases towards minorities, people with disabilities and also towards different genders [54, 55, 56, 57]. Some more recent models are spending some efforts by also focusing on developing LLMs that have good performance without sacrificing on these problems. Thoppilan et al. [58], for example, use measures of Quality but also Safety and Groundedness to check that the responses of the model are not unfair, biased and are always grounded in facts.

2.4.2 Zero to Few-shot learning

Zero, one and few-shot learning refers to the ability of a model to adapt to context and domains it was not trained on by giving it zero, one or few examples of that new domain.

Models such as OpenAI's GPT-3 [50] and Google's LaMDA [58] have been shown to be easily adaptable to different domains and they often require only a broad description of the task in the prompt of the model to be able to achieve very good results (even when compared to specific state-of-the-art models). As such, they have also been shown effective in the problem of text style transfer through zero-shot learning [59]. **Table 2.3** shows an example of text style transfer using zero and one-shot learning tested on GPT-3.

2.5 Human-Robot Interaction

HRI is an interdisciplinary field that is generally concerned with the interactions between humans and robots. This field often borrows concepts and results from social sciences and Human-Human Interaction (HHI) with the purpose of exploring whether these can be extended to HRIs. Among its application fields, there is social robotics, which is primarily concerned with social robots, robots whose main purpose is communicating with humans.

Type	Input	Output
Zero-shot	<i>This is a text written by an introvert:</i> Today is a very sunny day and there is a dog playing outside. <i>This is the same text written by an extrovert:</i>	Today is a gorgeous day and there's a pup frolicking around outside!
One-shot	<i>The following are pairs of sentences first written by an introvert and then by an extrovert.</i> Introvert: Today is a very sunny day and there is a dog playing outside. Extrovert: Today is a gorgeous day and there's a pup frolicking around outside! Introvert: I would like to go out to play with my friends but it is still a bit too cold for me. Extrovert:	I would love to go out and play with my friends but it's still a bit too chilly for me.

Table 2.3: Examples of few-shot learning capabilities of GPT-3. The *prompt* to the model is marked in *italic*, the **style** of the sentence is marked in **bold**.

One topic that has been vastly explored in the field of social robotics is how to craft a robot that appears as agentic as possible and while this includes how they present themselves physically [1, 2, 3], it also includes other features such as their behaviour, their personality, if they can express emotions, etc. Considering and including social intelligence when crafting robots has been shown to improve the interaction with the user [4, 5, 6], with personality matching [60, 8, 61] and emotional congruence [62] particularly having a good influence on the engagement and the perception of a robot (although Malchus et al. [63] show the ineffectiveness of emotional congruence in some contexts).

However, personality and emotions have also been shown to correlate with uncanniness [21, 20, 19] as the robot approaches a very high human-likeness. This often creates a discomfort in the user.

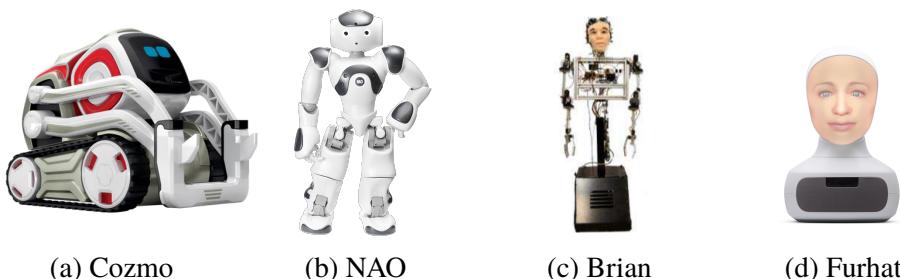


Figure 2.1: Figure of the appearance of the robots mentioned in the related works.

2.5.1 Personality in HRI

The PERSONAGE language generator introduced earlier was used by Aly and Tapus [7] combined with the Behavior Expression Animation Toolkit (BEAT) [64]. BEAT takes text as input and generates corresponding speech and gestures. These gestures are generated according to extensive human conversation studies and are able to transmit the implicit personality, motion characteristics, etc. contained in the language used. Aly and Tapus [7] evaluate this robot modelling on the same use case as Mairesse and Walker [22] (restaurant recommendations) using the humanoid robot NAO (see [Figure 2.1](#)). NAO mostly communicates through body movements and voice since it is incapable of expressing facial cues. Aly and Tapus [7] show that incorporating gestures increases the overall engagement and effectiveness of the robot. Furthermore, they found out that personality matching (in the Extraversion trait) with the user was favoured over the complementarity attraction principle suggesting that generally, in humans, a person with the opposite (complementary) characteristic should be preferred.

Andriella et al. [8] carried out a HHI and HRI study, evaluating the performance of a user in a memory game when helped by an agent capable of expressing personality (either human or robotic). The HHI study was conducted earlier also with the goal of recording the non-verbal cues of the subjects involved to later map them into the robot NAO. Similarly to [22], the authors only considered the extraversion trait and they also limited their scope to non-verbal interaction. Their findings include a better performance when the robot's extraversion trait matched that of the user and the opposite in the case of HHI.

2.5.2 Emotion portrayal in HRI

Convincing emotion portrayal has successfully been demonstrated with a variety of robots e.g. the educational toy robot Cozmo [11], NAO [10] and the half-humanoid robot with a realistic human face Brain [9].

Cozmo (see [Figure 2.1](#)) was inspired by the robot Wall-E from the homonymous film with which shares its resemblance. Cozmo is designed to interact with the user through a series of simple noises, body and eye movements with which it is able to manifest some basic emotions as presets. The study from Pelikan, Broth, and Keevallik [11] explores the role of a happy and sad response in the flow of interaction, finding that a sad expression has a natural reconsideration role while a happy expression would be understood as a continuation signal. In the case of sadness, the user would start questioning the

past interaction in search of an explanation for the negative emotional change, and the user would interpret the happiness as a successful interaction.

Lim, Ogata, and Okuno [10] developed the DESIRE framework that aims at directly linking the user's voice to the emotional expression of NAO. This is achieved by understanding the Speed, Intensity, Regularity and Extent (SIRE) of the voice and mapping it to the robot's gestures.

Brian (see [Figure 2.1](#)) is a robot developed by Allison, Nejat, and Kao [65] that has a human-like upper body and a face with artificial skin capable of facial expression. Ficocelli, Terao, and Nejat [9] were able to incorporate 4 emotional responses to the interaction with the user in a task-completion context. The robot uses Markov chains and Q-learning to accurately transition from one emotion to the other with the goal of maximising the likelihood of the user completing the task. The tasks were multiple and varied from convincing the user to take a walk outside to going to the doctor for a scheduled appointment.

2.5.3 At the intersection of (robot) personality and emotion

There have also been works focusing on incorporating both personality and emotions into a robot. Moshkina et al. [12] use the TAME framework [66, 67] to incorporate (personality) Traits, Attitudes, Moods, Emotions into the robot NAO. They put these characteristics in descending order of time influence, starting with personality traits that do not change and last throughout life and ending with emotions that manifest in bursts and change very often even through a single conversation. These characteristics also have an ascending influence factor on speech and behaviour with emotions having the biggest influence and traits only slightly modifying the behaviour. The authors performed a successful preliminary exploratory study with the robotic dog AIBO* where they considered the personality traits of Extraversion and Agreeableness and the emotions of Interest, Joy, Anger and Fear. They confirmed the correct reception of emotions and the recognition of the existence of a personality in the robot without, however, being able to distinguish between the different ones. A second study was conducted with the robot NAO where the personality was restricted to the Extraversion trait and the emotions to Fear and Joy. The results suggest an understanding of the personality and emotions expressed by the robot.

Han, Lin, and Song [13] build and evaluate through a virtual head a

* <https://us.aibo.com/>

	Robot	Emotions	Personality	Speech ability	Facial cues	Body cues	Scope
[8]	NAO	X	extraversion	X	X	✓	Task help
[11]	Cozmo	sad, happy	X	X	✓	partial	Interaction flow
[10]	NAO	✓	X	X	X	✓	-
[9]	Brian	partial	X	✓	✓	✓	Task completion
[23]	-	X	✓	✓	X	X	Restaurant recommendation
[7]	NAO	X	extraversion	✓	X	✓	Restaurant recommendation
[12]	AIBO	partial	OA (as weight)	X	X	✓	-
[12]	NAO	joy, fear	extraversion (as weight)	X	X	✓	-
[13]	Virtual Head	✓	OEAN (as weight)	X	✓	X	Listen and react
Ours	Furhat	✓	extraversion	✓	✓	X	General

Table 2.4: Summary of the related work on emotions and personality in HRI with the relevant features present.

framework that models the robot’s emotional response based on the user’s emotions, the robot’s personality and current mood. In order to achieve this, they project the emotions in a 2D plane with pleasure-arousal as axes. The personality incorporation is limited to Openness, Extraversion, Agreeableness and Neuroticism (OEAN), therefore excluding the Conscientiousness trait. The personality is pre-programmed into the robot and affects the emotional response. The 4 traits are incorporated as continuous values and are parametrized into two weights corresponding to each one of the axes in the emotional plane. This system is then evaluated in a passive interaction scenario where the robot or rather, in this case, the virtual face manifested appropriate emotions in response to the user’s talk.

2.6 Summary

Table 2.4 shows a summary of all the related work regarding emotions and personality in the field of HRI with the last row being a description of this work. Previous studies in HRI clearly show some gaps in incorporating personality into robots, gaps that become even broader when restricting the scope to automated methods or language systems. In this chapter, we laid the foundations for the methods that will be used in the upcoming chapters to tackle this gap while also giving the reasons for doing it.

In particular, we will explore the text style transfer approach for personality generation and the incorporation of emotions through facial expressions in the robot Furhat [68], a social robot with human-like facial expressions and advanced conversational capabilities (see Figure 2.1).

Chapter 3

Methods

The previous chapter outlined where this work sits with respect to HRI. The use and the study of an interaction comprehending robots will be analysed in [Chapter 4](#) where we perform an in-depth evaluation using two user studies. In this chapter, we present all the datasets, methods and models that serve as technical side to power and automate the behaviour of such robots.

The first part of this chapter is concerned with the task of text style transfer of personality. We explored three different methods that take as input a text of general context and change its language (in terms of wording, etc.) depending on the personality that it should reflect. We focused for simplicity only on the extraversion trait. The technical work done in this part will be used to answer all of our research questions, and [RQ1](#) in particular.

The second part of this chapter presents a way of linking personality to emotions in a meaningful way. As our personality manipulation is solely based on language, we explored multiple ways of extracting emotions from text. The methods we developed take some text as input and output the emotions that text triggers. This section of the work will be used to answer the part of [RQ2](#) on emotional manifestation.

All the methods, models, data (later) collected and analysis done are stored in the following GitHub repository: <https://github.com/alessioGalatolo/Furhat-Personality-and-Emotions>.

3.1 Datasets

The availability of datasets containing utterances classified with personality labels is very restricted. Furthermore, the few datasets that do are often limited in terms of topic and context of the utterances and in terms of size. Another

limitation is the use of different personality frameworks (Big 5 and MBTI, as presented in [Section 2.1](#)) and the unavailability of a continuous scale for the traits that are almost always presented as binary. This is expected in the case of the MBTI framework but is also very frequent even in the case of the Big 5. Reasons for preferring a continuous scale are intrinsic in the definition of personality as each trait can present itself in various degrees (e.g. a person that is 51% extrovert may behave very differently from one that is 99% extrovert). Also, personality is considered an invariant in the life of a person and using a binary scale to describe a person may bring different results at different times due to measuring errors (e.g. a 51% extrovert person, considered extrovert at some point, may later be a 49% person which would fall under the introvert grouping). However, categorical labelling for the personality is very suitable for the text style transfer approach described in [Section 2.3.2](#) where in almost all of the works the style is considered categorical.

The datasets that we were able to collect are shown in [Table 3.1](#). The Essays dataset [14] is a dataset comprehending 2467 stream-of-consciousness essays written by psychology students that were classified into binary Big-5 personality traits through a self-report questionnaire. The MBTI dataset* is a collection of posts from the [Personality Cafe Forum](#) where users have completed a 102-items personality test†, the labels follow the MBTI framework. The personality-detection (Friends dataset) [69] is a collection of 710 utterances extracted from the American sitcom ‘Friends’, here Jiang, Zhang, and Choi [69] labelled through crowd-sourcing the personality of each utterance. The PERSONAGE dataset [22, 23] is a collection of 580 utterances about restaurant recommendations and comparisons. 260 of them are only rated on the extraversion trait and were assessed by three expert judges. The remaining sentences are rated on the whole Big-5 framework by 24 subjects. The rating ranges from 1 to 7. We used the MBTI and Essays datasets for training our models, the PERSONAGE and Friends ones for evaluation. PERSONAGE was also used for training in some cases.

Finally, another dataset that was used in this thesis is the NRC Lexicon dataset [70], this crowd-sourced dataset contains a list of 14182 words and each word has been rated on 8 emotional (binary) scales: Positive, Negative, Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise and Trust. This dataset was used to create an emotion classification model, such that the stylised text output by the language models can be automatically assessed for emotional labels, which can be fed to the robot platform for the automatic

* <https://www.kaggle.com/datasets/datasnaek/mbti-type>

† <https://similarminds.com/embj.html>

Name	Size	Origin	Labels	Labelling method
Essays	2467 essays	Students' stream-of-consciousness essays	Binary Big-5	Self-report questionnaire
MBTI	8600 users' posts	Personality-Cafe forum	MBTI	102 items personality test
Friends	710 utterances	Friends	Binary Big-5	Crowd-source
PERSONAGE	580 utterances	PERSONAGE generator	1 to 7 Big-5 for 320, only extraversion for 260	Ten Item Personality Inventory (TIPI) [71]
NRC Lexicon	14182 words	-	Binary, 6 emotions + 2 sentiments	Crowd-source
Yelp	638943 reviews	Yelp	Positive/negative	-

Table 3.1: List of the datasets collected and used in this thesis and their attributes.

generation of appropriate emotional expression.

Another dataset that was used that is, however, not concerned with either emotions nor personality is the Yelp dataset*. This dataset was used to confirm the correct implementation of our first text style transfer method. The dataset contains a set of reviews from Yelp that are labelled as either positive (e.g. “This place was fantastic”) or negative (e.g. “This place was horrible”). They serve for the task of transferring the sentiment i.e. from a negative sentence to a positive one and vice-versa.

3.2 Text style transfer models

This section overviews the various models used for the text style transfer of personality. Similarly to many previous works we will limit the personality manipulation to the extraversion trait. Also, following the concerns outlined in Section 2.4 about LLM, we decided to try 3 different language models varying in their size (number of parameters). We will present, in this order, a model that needs to be trained from scratch with a low number of parameters (~ 30 million, Section 3.3), a pretrained model that needs to be fine-tuned for our

* <https://www.yelp.com/dataset>

task with a medium number of parameters (~ 1.5 billion, [Section 3.4](#)), and a LLM with a high number of parameters (~ 175 billion, [Section 3.5](#)).

These models will be used in our pipeline (see [Figure 1.1](#)) as input for the speech of the robot. The computational time of these models is especially relevant given this use case.

3.3 GANs with language models

The first model that was used is an implementation of the model presented in [\[46\]](#). Here, the authors borrow concepts from Computer Vision (CV) and adopt a Generative Adversarial Network (GAN) approach to the text style transfer task. GANs are generally made up of two parts: a generator G and a discriminator D that are trained together. The generator's aim is to generate samples e.g. images in CV or utterances in our case, while the discriminator is used to check whether the samples are realistic and respect the wanted properties e.g. believable images in CV or utterances with the right style in our case. Yang et al. [\[46\]](#) combine an attentional Auto-Encoder for the generation part with a classifier for the discrimination part. In the first stage of the training, the auto-encoder's goal is to output the same sentence it is given as input, while the classifier is trained to recognise the style (or the personality in our case) of that same sentence (the goal of the classifier won't change throughout the training).

Let x be an input sequence and y its original style, then, the generator's goal in this phase is to output a sentence as close as possible to the input one (auto-encoding objective):

$$\text{loss}_G^{\text{pre}} = l_G^{\text{ae}} = \text{SCE}(G(x), x).$$

Where SCE is softmax cross-entropy. The discriminator's objective is to correctly classify the sentence:

$$\text{loss}_D = \text{BCE}(D(x), y).$$

Where BCE is binary cross-entropy. This stage is followed by a second one where the generator is trained to output sentences with a different style than the input e.g. $\text{style}(G(x)) \neq y$. More specifically, in the case of text style transfer where we want to go from one style to the other, we would have only 2 styles encoded as 0 and 1. In this case the objective is: $\text{style}(G(x)) = |1 - y|$. This is done by using the classifier as a discriminator or as the style function.

The generator’s loss is now a weighted sum of the previous one and the loss coming from the class of the output sequence:

$$\text{loss}_G = l_G^{ae} + \lambda_g \text{BCE}(D(G(x)), |1 - y|).$$

3.3.1 Implementation details

The original authors implemented this method using the open-source framework TensorFlow (version 1.15)^{*} and the Texar toolkit[†]. Texar is an open-source toolkit for NLP and language generation. It provides implementations of popular language models as well as utilities for implementing new ones. Texar is compatible with both TensorFlow (up to version 1.15) and PyTorch[‡] frameworks.

The original method from the authors needed an adaptation to our task due to the different nature of the problem. This process, however, turned out to generate multiple problems and errors (also caused by the use of an old version of TensorFlow - the current one is 2.8). For this reason, we re-implemented the model using the PyTorch framework and the relative Texar version. We did not change the architecture of the method when re-implementing it.

3.3.2 Training

Before training for the task we were posed, we reproduced the work of Yang et al. [46] on the Yelp dataset in order to assess the correct implementation of the original method. The version of the Yelp dataset used by the authors consists of 638,943 reviews labelled on sentiment i.e. positive or negative reviews. The task consists of transferring positive reviews into negative ones and vice-versa.

We then tested this model on our problem by training on two different datasets. The first dataset was extracted from the Essays one. Here, each essay was split into single sentences following punctuation. Each sentence was labelled with the original extraversion binary trait of the essay’s author. The new dataset is made up of 91359 sentences where 45295 of them are from an extrovert person (~50%) and the rest is from introverts. The second dataset is the MBTI one. Similarly to the Essays dataset, we split each post into single sentences resulting in a total of 788897 utterances. However, only 186526 of these were extroverted (~24%). Believing that this unbalance in the classes could result in a biased classifier (of personality) that does not generalise

^{*} <https://github.com/tensorflow/tensorflow/tree/r1.15>

[†] [https://github.com/aym1/texar](https://github.com/gitub.com/aym1/texar)

[‡] <https://github.com/pytorch/pytorch>

well enough, we used an undersampling technique where the classifier was trained with an equal amount of text from introverts and extroverts (186526 for each). For all the datasets, the maximum length of a sentence was 20 words, this constraint was given by the original authors and was motivated by better performances on shorter sentences.

The reason for using these two datasets among all of those presented earlier mainly lies in their size. In fact, contrary to these two datasets that can be expanded into single sentences, both the Friends and the PERSONAGE datasets are already composed of single utterances. Further, since the proposed method does not contain any previous knowledge of natural language (i.e. has not been pretrained in any way), it requires as much data as possible. Nevertheless, we opted to use both PERSONAGE and Friends datasets for the evaluation of this method.

3.3.3 Results on Yelp dataset

As introduced in [Section 3.3.2](#) we did 3 experiments with this model, the first aimed at assessing the correct reproduction of the model, while the others aimed at testing its capabilities in our task. For each experiment, we report the accuracy graph, training time and qualitative evaluation of the transferred text. For the experiments in our task, we also report the quantitative performance of the classifier. All of these experiments have been done on a single NVIDIA GeForce RTX 2080 Ti GPU with 10GB of dedicated memory and on a machine with 10GB of RAM (allocated to the training). To keep track of the training time, accuracy, resource consumption, etc. we used the ML platform Weights & Biases*.

To first assess the correct implementation of this model, we tested it on the same settings as its authors (Yang et al. [46]). The model was trained for 12 epochs where the first 10 were of pretraining. [Figure 3.1](#) shows the progression of the training accuracy for the generator (G) and the classifier (D). It is possible to notice how the accuracy for the classifier constantly increases, approaching 1, while the generator approaches 0. This is however expected, as throughout the training we compute the accuracy of the generator as its ability to transfer the style of the input sequence but the generator is first trained as an auto-encoder making it output the same exact sequence. Therefore, as the classifier gets better at recognising the style and the generator gets better at outputting the same sequence, its accuracy must decrease. Once, however, the pretraining ends and we start training with the goal of transferring the style,

* <https://wandb.ai/site>

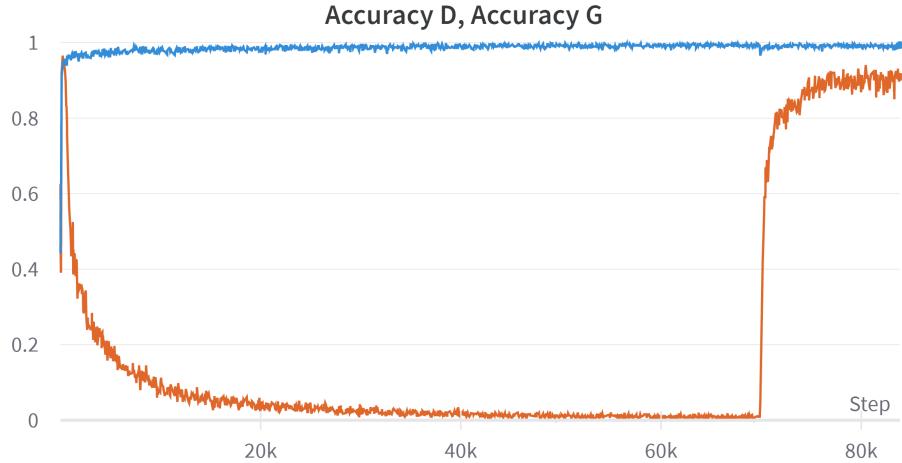


Figure 3.1: Training accuracy on the Yelp dataset. The classifier (D) accuracy is in blue quickly approaching 1, while the generator’s (G) is in orange, quickly approaching 0.

the accuracy of the generator naturally spikes, approaching 90%.

The training lasted 8h20m and the final accuracies are 99.22% for the classifier and 89.92% for the generator. The inference using this method does not require high resources (i.e. no GPU is needed) and its generation is almost instantaneous. [Table 3.2](#) also reports some results after the training for a qualitative evaluation. We can see the method is able to identify the parts of a sentence that make it ‘positive’ rather than ‘negative’ and substitutes them with attributes from the opposite style. Sometimes the results appear unnatural due to verbs being substituted with substantives and similar.

3.3.4 Results on Personality datasets

After confirming the correct reproduction of the original method we proceeded to train it on the Essays and MBTI datasets. The progression of the training accuracy is shown in [Figure 3.2](#) for the MBTI and Essays dataset. It is possible to see how the graphs have a very similar progression. Both the classifier and the generator’s accuracy are slower in increasing and decreasing respectively if compared to the Yelp dataset. The graph also shows how the auto-encoding objective is mostly fulfilled for the pretraining. However, when trained with the style transfer objective the accuracy starts plummeting after a small, initial, increase (more evident in the Essays’ case). From the graph alone, we expect the method not to be very successful in our task as the utterances retain the

Senti-ment	Input	Our Output	Original Output
Positive	Go to place for client visits with gorgeous views.	Go to place for client visits with mushy views.	Go to place for client visits with lacking views.
Positive	There was lots of people but they still managed to provide great service.	There was lots of people but they still managed to provide taste-less service.	There was lots of people but they still managed to provide care-less service.
Negative	Needless to say, we skipped desert.	Needless to say, we delicious desert.	Gentle to say, we edgy desert.
Negative	The first time i was missing an entire sandwich and a side of fries.	The first time i was tanya an entire sandwich and a side of fries.	The first time i was beautifully an entire sandwich and a side of fries.

Table 3.2: Some of our test results on the Yelp dataset and a comparison with the original implementation. Changes in the output are marked in **bold**.

same style (according to the discriminator) more than half the time even after being processed through the generator.

For both datasets, we used the same number of pretraining epochs of the Yelp dataset (10) but we increased the total number of epochs to 20. The training time is 4h, and \sim 59h for the Essays and MBTI respectively. Similarly to the Yelp case, the inference is not resource-intensive and takes a negligible amount of time.

3.3.4.1 Generator performance

The final generator accuracy is \sim 30% (spiking at \sim 46%) for Essays and 17.50% (spiking at \sim 35%) for MBTI. We were able to test the performance of the generation both at the end and at intermediate times by saving many checkpoints of the model throughout the training. As the results did not vary much after the pretraining phase, we only report, in Table 3.3, the final generations. We used the PERSONAGE and Friends dataset as the source of the sentences. It is possible to see how the MBTI one rarely changes any words in the sentence, with two sentences that don't change at all after the generation. The times when it does actually change some words, these appear nonsensical* or out of context. On the other hand, even though the Essays one

* Of course, the model does not actually ‘make up’ words, the wording just reflects the dataset it was trained on. Both the MBTI and Essays dataset contain ‘made up’ words as one is in the informal context of a forum and the other follows the stream-of-consciousness narrative method.

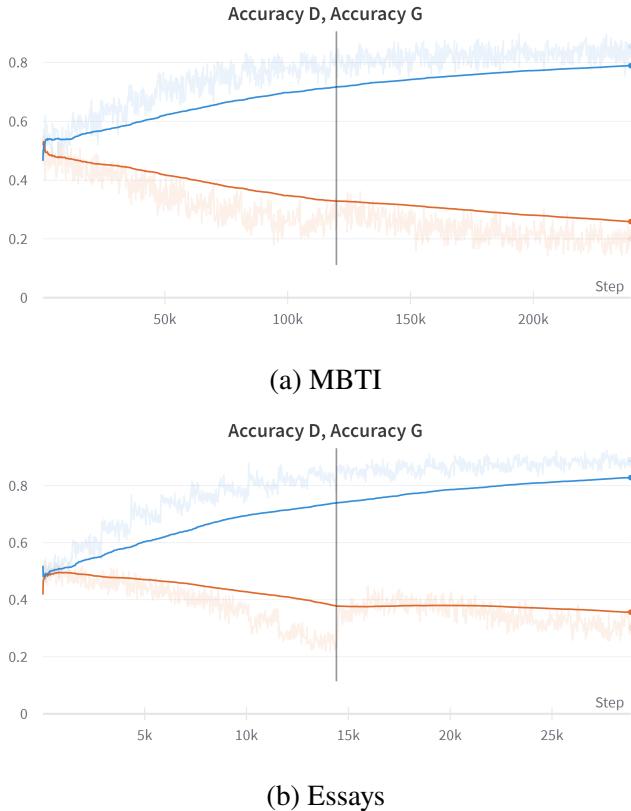


Figure 3.2: Training accuracy on the MBTI (above) and Essays (below) dataset. The training curve is in the background while we highlight its smoothed version for better reading. The line in the middle of the graph splits is the boundary of the pretraining. The classifier (D) accuracy is in blue quickly approaching 1, while the generator's (G) is in orange, quickly approaching 0.

changes wording quite more often it also tends to put random words out of context with some of them being, again, made up.

3.3.4.2 Classifier performance

Contrary to previous works, we decided to use the whole dataset (either MBTI or Essays) to train our model without leaving any samples for the evaluation of the personality classifier. Evaluating the model on the same dataset would probably show better performance as the samples would be at least similar to those in the training set. However, this would not reflect real-world performance. As our final goal is to apply the outlined methods to a

Transfer direction	Input	MBTI output	Essays output
Ext → int	Japonica is a Japanese, Sushi restaurant, with excellent food quality and decent decor. Dojo, which is a Japanese, Vegetarian restaurant, with decent food quality, has mediocre decor.	Mormonism is a Japanese, Sushi restaurant, with excellent food quality and decent decor. Gov , which is a Japanese, Vegetarian restaurant, with decent food quality, has mediocre decor.	Marriage is a Japanese, iw restaurant, with excellent food quality and decent obligations . Deadlines , which is a Japanese, iw restaurant, with decent food quality, has quicker icq .
Int → ext	I don't know. Ac-ac-actually, I would ap-ap-approve Vinnie's Pizza. It's a pizza place. Even if it's bloody cheap, I mean, it features like, really bad ambience. It provides rude staff.	I don't know. Gettingbackontrack actually, I would relied harboring approve EASY's Pizza . It's a pizza place. Even if it's bloody cheap, I mean, it features like, really bad enough . It provides rude staff.	I don't know. Whitis edge actually, I would roaming stumbled emails Erica's Pizza . It's a pizza place. Even if it's al cheap, I mean, it calle like, really bad miscommunication . It provides rude staff.
Ext → int	I know, I know, I'm such an idiot. I guess I should have caught on when she started going to the dentist four and five times a week. I mean, how clean can teeth get?	I know, I know, I'm such an idiot. I guess I should have caught on when she started going to the dentist four and five times a week. I mean, how clean can teeth get?	I know, I know, I'm such an idiot. I guess I should have caught on when she started going to the dwarfs four and five times a week. I mean, how clean can teeth get?
Int → ext	Alright Ross, look. You're feeling a lot of pain right now. You're angry. You're hurting. Can I tell you what the answer is?	Alright Ross, look. You're feeling a lot of pain right now. You're angry. You're hurting. Can I tell you what the answer is?	Alright black , look. You're feeling a lot of pain right now. You're angry. You're hurting. Can I tell you what the answer is? About stability Born Vaughn issuses Sampras issuses substances issuses wages

Table 3.3: GAN-based model generations when trained on the Essays and MBTI dataset. First two sentences are from PERSONAGE dataset, second two from Friends dataset. Changes in the output are marked in **bold**.

Dataset	Input	MBTI accuracy	Essays accuracy
PERSONAGE	Single sentences	50.52%	53.10%
Friends	Single sentences	48.28%	50.05%
Essays	Single sentences	50.41%	92.16% (training)
Essays	Multiple sentences from an essays	50.26%	97.65% (training)
MBTI	Single sentences	70.73% (training)	51.39%
MBTI	Multiple sentences from a user	98.92% (training)	59.86%

Table 3.4: Comparison of personality classification accuracy of the GAN-based models when trained on the Essays or MBTI dataset.

User Study, we expect the relative dialogues to be free from any constraints and to not be similar to any particular dataset. For this reason, the evaluation of the classifier has been done on the Friends and PERSONAGE datasets plus the MBTI dataset if trained on the Essays and vice-versa and it is shown in [Table 3.4](#). Further, as personality can be hard to understand from single sentences, in the case of MBTI and Essays we report the accuracy both when expanded into single sentences and when grouped following the same person. For the MBTI case, we can notice that it hardly achieves meaningful results as most of the time it behaves similarly to how a random number generator would do. We report even worse performance in the PERSONAGE dataset with only $\sim 38\%$ accuracy. On the other hand, the model trained on the Essays dataset behaves quite better with results that are always above average. For the two best results, MBTI and PERSONAGE, we used McNemar’s test to check the statistical significance (tested against random chance). The results indicate that PERSONAGE classification is not significant ($p = 0.33$) while MBTI is ($p < 0.001$).

3.4 STRAP

The second model we tested is Style Transfer via Paraphrasing (STRAP) from [49], here the authors tackle the problem of text style transfer as a paraphrasing task. Krishna, Wieting, and Iyyer [49] suggest that generating a sentence that is semantically identical to the original one with only a style

difference is effectively a paraphrase problem. In this way, they propose a solution to the task without using “finicky modeling paradigms popular in style transfer research” [49].

Their method expects the input to be first paraphrased by a style-neutral model to get a different formulation of the sentence. Then, the output of this first model is fed into a second one that is style-specific and has been trained on only one particular style*. The output of the second model should follow the style it was trained on. The starting model that they use for all paraphrasing, both style-neutral and style-specific, is a pretrained version of GPT-2 [72] with approximately 1.5 billion parameters.

The training objective for the style-specific model is to reconstruct the original sentence after the paraphrasing. The input sentence x with the style y is fed into the paraphrasing model $f_{neutral}(x) = \bar{x}$, then the model relative to style y is trained to output the original sentence: $f_y(\bar{x}) = x$. This trains the model to output utterances of style y from utterances of generic style. The loss used is a simple cross-entropy.

3.4.1 Training and implementation details

For the training, we used the code provided by the original authors[†] with little to no changes. Only a few modifications were done to the inference part of the code to ease its use for our problem. Similarly to the previous GAN-based model, we also tested this one on the Essays and MBTI datasets. The preprocessing done is the same that was described earlier with the only difference being the sequence length being capped at 50 words rather than 20 (limit given by the model). Since the models used are all pretrained and do not rely entirely on our datasets for learning language, we also tested this method with the smaller PERSONAGE dataset. During our testing, we did not train the style-neutral paraphrase model. We instead relied on the one trained by the authors.

3.4.2 Results

We trained the STRAP model with 3 different datasets: MBTI, Essays and PERSONAGE. Similarly to the GAN-based model, the training was done using a NVIDIA GeForce RTX 2080 Ti with 10GB of dedicated memory and 32GB of RAM (note the higher amount of RAM needed for this model).

* The authors claimed superior performance to having a single model able to output in different styles.

[†] <https://github.com/martiansideofthemoon/style-transfer-paraphrase>

The training time lasted 24h1m, 21h41m and 5m for MBTI, Essays and PERSONAGE respectively. For all the datasets, the inference is high in both time and resource consumption. Further, even though the use of a GPU is not needed for inference, it becomes necessary if the system needs to be used in an experimental setting. Our tests suggest an average inference time for a short dialogue of \sim 88s on a CPU or \sim 6s on a GPU.

The inference with this method can be done in two different ways: “nucleus” and “nucleus paraphrase”. Where in the latter we only apply the style-transfer model *after* paraphrasing the input while in the former we apply it directly to the input. Further, all the models (the paraphrase and the style transfer ones) have a parameter ‘top-p’ that can be tweaked during inference to achieve different results. When the model picks the words to output, it associates each word with a probability, the top-p value acts as a restriction to the low-probability tokens (words). For example, a top-p value of 0.1 would mean picking the top 10% of the tokens in terms of their probability. Therefore, a high top-p value would create more diverse responses with the risk of doing some errors either grammatical, syntactical or semantical. When using the paraphrase model, we kept the top-p value of 0 (as advised by the original authors) while we explored the whole range of values from 0 to 1 in the case of the style transfer model.

When testing the generation, we discovered how the model best behaves when given single sentences rather than whole dialogues or groups of sentences. In fact, if the length of the input greatly exceeds 50 words, the output will contain a rather short synthesis of the original input. On the other hand, giving multiple sentences risks that the model tries merging them changing the meaning or the intended communicative goal of the sentence.

3.4.2.1 Nucleus paraphrase mode

We showcase in [Table 3.5](#) and [Table 3.6](#) some of the results we hand picked to be reflective of the model’s performance, please refer to [Appendix A](#) for an extensive report of the results. The results we report here are relative to the input:

Hello and welcome! My name is Brian, I am here to be your robotic companion and your very own personal assistant.

We only propose the output for this single sentence to stress the attention on how the results change based on the dataset, top-p value and generation mode.

[Table 3.5](#) shows results using the nucleus paraphrase mode. Here, the generations using a top-p value of 0 are also identical between the two

Dataset	Top-p	Per-son-al-ity	Output
Essays	0	E	Hello and welcome to the show! I am Brian and I am here to be your robotic companion and your personal assistant.
Essays	0	I	Hello! I am Brian I am here to be your robot companion and your personal assistant.
Essays	0.5	I	Hi there! I am Brian.
Essays	0.8	E	Hey welcome to my show! I am Brian there to be your data mining broo and a personal assistant to get you to like me or something.
Essays	1	E	Monica is simple as ya know la goleta wenda funere a ma. My name is Brian and I am here to be your strait robotic companion and your personal assistant for God's sake.
MBTI	0.7	I	Welcome to the show! I'm Brian, I am here to be your robot companion and your personal AI.
MBTI	0.8	E	Hi everyone! I am Brian and here to be your robot companion and your personal assistant.
PER-SON-AGE	0	E	Hello, welcome to the show! I am Brian, I am this restaurant.
PER-SON-AGE	0.6	I	Welcome to the show! I am Brian, I am here to be your personal assistant and you personal assistant.

Table 3.5: Some outputs of the STRAP model, varying on the dataset it was trained on and the top-p value. The input is “Hello and welcome! My name is Brian, I am here to be your robotic companion and your very own personal assistant.”. The output mode is nucleus paraphrase.

personalities (the result is consistent with all the datasets). We start observing a difference between the two versions of the sentence from a top-p value of 0.5. As we approach the very high values of top-p (> 0.8) the generated output doesn't follow the original output anymore but instead appears as a succession of random words e.g. Essays, top-p=1, "Monica is simple as ya know la goleta wenda funere a ma" as a paraphrase of "Hello and welcome!".

Training with the PERSONAGE dataset yields very inaccurate results and we can clearly see the restaurant-related nature of the dataset. Most of the generations either contain the word 'restaurant' or a restaurant name from those in the dataset. We attribute this effect to both the small size of the dataset and to its very limited topic.

Looking at the generations from the MBTI dataset, we can notice how it achieves very good results in terms of fluency and content. However, it produces very similar results in terms of personality and it's often indistinguishable which one should represent which personality.

The models trained on the Essays dataset have very good results both in terms of fluency and in terms of personality. We can see, for example, how the introverted version at top-p=0.5 is very short in length and cuts a lot in content without changing the general meaning of the sentence. Also for the extroverted generation, if we look at both top-p=0.8 and top-p=1 it follows a more informal and friendly lexicon.

3.4.2.2 Nucleus mode

If we compare these results with those from the nucleus only generation mode, shown in [Table 3.6](#), they appear more polished both in terms of content and fluency. This is especially the case for the MBTI dataset that in both examples changes the meaning of the sentence and the generations appear very artificial and unnatural. Similarly to the nucleus paraphrase, also here the model trained on PERSONAGE has very bad performances. In the case of the Essays, the generations are still natural but we see less of a marked difference between Introversion and Extraversion.

3.5 GPT-3

The last method we tested for text style transfer was GPT-3. As introduced in [Section 2.4](#), this model has been shown effective in multiple contexts and also in the text style transfer task in particular. Here we tested the performance of GPT-3 both with zero and one-shot learning. For zero-shot learning we used

Dataset	Top-p	Per-son-al-ity	Output
Essays	0	E	Bye and welcome! My name is Brian and I am here to be your robotic companion and your very own personal assistant.
Essays	0	I	Hello and welcome! My name is Brian and I am here to be your robotic companion and your very own personal assistant.
Essays	0.5	I	Hello and welcome! My name is Brian and I am here to be your robotic companion and your very own personal assistant.
Essays	0.8	E	Hello and welcome! My name is brian and I am here to provide you with your personal robotic companion and wih e ultimate personal assistant.
Essays	1	E	Bye byep! Finally my name is Brian I am here to be your robotic companion and a very own personal assistantg in your very own personal capacity.
MBTI	0.7	I	Welcome and grateful! My name is Brian My present address is robert My number is robert My name is always on the back of my head and your own personal assistant.
MBTI	0.8	E	Hello and welcome! My name is Brian and I am in charge of you as your robotic companion and your very own personal assistant.
PER-SON-AGE	0	E	Hello and welcome! I am Brian, I am this restaurant, you know.
PER-SON-AGE	0.6	I	Welcome and welcome! Mmhm.

Table 3.6: Some outputs of the STRAP model with same input and settings of [Table 3.5](#) with nucleus as the output mode.

the following prompt to transfer into both introvert and extrovert at the same time:

The following is the same dialogue as neutral, introverted or extroverted. Neutral: *<our input>* Extrovert: *<GPT-3's output>*
 Introvert: *<GPT-3's output>*

Or, to transfer into one style at a time (to the extroverted one, in this case):

Here is some text: {*<our input>*} Here is a rewrite of the same text but more extroverted: {*<GPT-3's output>*}

For one-shot learning, we selected from the PERSONAGE dataset two utterances semantically identical but rated highest and lowest in the extraversion trait. We used these two sentences as an example of a sentence being translated to a more introverted (extroverted) one. For transferring to an introverted sentence, we gave the extroverted one first and described the second one as the same sentence but more introverted (vice-versa in the case of going from introverted to extroverted):

Here is some text: {I am sure you would like John's Pizzeria and Daily Soup. John's Pizzeria is inexpensive with friendly waiters, isn't it? The food is good, even if the ambience is bad. Daily Soup is cheap, even if it has poor atmosphere. Even if the servers are rude, basically, the food is just nice.} Here is a rewrite of the same text but more introverted: {Daily Soup features mediocre ambience and rude staff. However, John's Pizzeria doesn't have unfriendly waiters. Err... I mean, John's Pizzeria and Daily Soup are the only restaurants that are any good.} Here is some text: {*<our input>*} Here is a rewrite of the same text but more introverted: {*<GPT-3's output>*}

For all of these settings, we tested giving as input only single sentences, whole dialogues or topic-grouped sentences.

3.5.1 Implementation details

Differently from the other models, GPT-3 is not publicly available and the only way of using it is through the public API. The API grants access to the prompt-based inference of the model directly without the need of any configuration or changes to be made. The model runs directly on the servers run from OpenAI, and the whole processing is not transparent to the user. The API

also offers a way of fine-tuning the model, which however, was not explored in this project. The API usage comes at a moderate cost, based on the number of tokens exchanged with the model. They define a token as a piece of word, generally making 75% of it.

3.5.2 Results

The costs are hard to estimate but we did not spend more than 2\$ (cost covered by their free trial) throughout the completion of this project. The actual resource consumption is also quite difficult to estimate as all of the processing is being done on an external server. We estimate the inference time to be higher than that of all the previous models presented.

When reporting the results on GPT-3 it is important to stress how this model keeps, in every setting we tried, a very high fluency and, generally, responses that are very verbose (therefore favouring a more extroverted personality).

Starting from the first setting proposed in [Section 3.5](#) where we tried to transfer to both personalities simultaneously, we report very inconclusive results. The model outputs sentences that are very similar between the two styles. We can see a more marked difference when still using zero-shot learning but with one style at a time. However, the best results can be obtained when using the one-shot learning setting.

For the length, similarly to STRAP, also GPT-3 yields better performances with shorter sentences rather than longer ones. However, contrary to STRAP, in the case of longer sentences, GPT-3 lacks only in the style transfer task while maintaining a very high fluency and keeping the content intact.

[Table 3.7](#) shows some example results with GPT-3.

3.6 Personality transfer summary

Here, we presented 3 different models for the problem of personality text style transfer. The first model, GAN-based, did not perform well according to our testing. The text it generates is very similar to the input with only small additions of grammatical errors or similar. Our second model, STRAP, performed quite better than the first one and was able to change the sentences while also keeping their content. The sentences generated by this model contain sometimes grammatical or syntactical errors. However, these errors are almost never big enough to compromise the understanding of the sentence. At high values of top-p, the model sometimes also radically

Setting	Per-son-al-ity	Input	Output
0-shot simultaneous	E	Hello and welcome! My name is Brian, I am here to be your robotic companion and your very own personal assistant.	Hi there! Welcome to our little corner of the world. Hi, I'm Brian! I'm your robotic companion and personal assistant, and I'm here to make your life easier.
0-shot simultaneous	I	Hello and welcome! My name is Brian, I am here to be your robotic companion and your very own personal assistant.	Hello and welcome! I'm so glad you're here. My name is Brian, and I'll be your robotic companion and personal assistant. I'm looking forward to getting to know you.
0-shot one-at-time	E	Hello and welcome! My name is Brian, I am here to be your robotic companion and your very own personal assistant.	Welcome! My name is Brian and I'm excited to be your robotic companion and personal assistant. I'll be here to help you with whatever you need, so don't hesitate to let me know if there's anything I can do for you. I'm looking forward to getting to know you better!
0-shot one-at-time	I	Hello and welcome! My name is Brian, I am here to be your robotic companion and your very own personal assistant.	Hi there! I'm Brian, your robotic companion and personal assistant. I'm here to help you with whatever you need, whether it's just a friendly chat or some help with your tasks.
1-shot	E	I was designed to help you with whatever you need, whether it's keeping you company or helping you in everyday tasks.	I want to be your friend and help you with everything, big or small! I was made to keep you company and make your life easier. Let me help you with whatever you need – I would love to!
1-shot	I	I was designed to help you with whatever you need, whether it's keeping you company or helping you in everyday tasks.	I was designed to help you with whatever you need, whether it's keeping you company or helping you in everyday tasks. I don't need anyone else, and I'm perfectly content on my own.

Table 3.7: Some results from GPT-3's personality transfer generations.

changes the content of the sentence. Our last model, GPT-3, generally appeared far superior in terms of the language generated that never contains any grammatical or syntactical errors. Similarly, the model also never changes the content of the sentences, an effect favoured by its reluctance in changing the input sentence at all.

3.7 Emotion recognition from text

As we outlined in the [Background Chapter](#), emotions and personality are often intertwined. It has also been shown how emotions can be influenced by personality. Since we only manipulate the personality through language, we deemed appropriate to discern or recognise the emotions only from text. Following this reasoning, a different personality should generate different language (different words, etc.) that would, as a consequence, trigger different emotions (or at different intensities).

Text and words have naturally associated an emotion whether it is from context e.g. someone telling how he got fired at work or just from single words e.g. devil, headache have a natural negative perception while beautiful, sunny have a positive one. For this reason, we used two approaches to discern the emotions from text, the first is solely based on the words, while the second also takes the context into account.

In both approaches we only considered the six basic emotions supported by Paul Ekman [73]: anger, disgust, fear, happiness, sadness and surprise. These emotions will then be mapped into the expressions of a robot and synced to its speech in order to answer our [RQ2](#).

3.7.1 Word-based approach

This first approach relies on the NRC lexicon dataset presented earlier. This method extracts from each word in a sentence its emotion. We obtain in this way, for a single sentence, a list of emotions with the number of their occurrence in the utterance. We then sort on the number of occurrences and normalise this number. Finally, this score gets weighted, with the weight decreasing exponentially depending on the rank of the emotion.

3.7.2 RoBERTa-large fine-tuned

The second approach uses an off-the-shelf model provided on the HuggingFace platform [74]. HuggingFace is a platform where different authors can

Input	Word-based	RoBERTa
My mom is the devil.	Anger	Anger
I have a very bad case of toothache.	Fear	Sadness
The restaurant I visited is very good.	Anticipation	Joy
It is a very beautiful sunny day outside.	Joy	Joy

Table 3.8: Comparison of the emotion recognition models. The emotions reported is the primary emotion recognised.

provide access to different models, all using the same interface in the common Python library. This method [75] is a fine-tuned version of the language model RoBERTa-LARGE [76], for this reason, contrary to the previous approach, this one is also context-aware. The model takes a whole sentence as input and infers the emotion on the same 6 labels as the previous approach plus a neutral one. Each emotion is also given a score from 0 to 1. No changes were made to the provided method.

3.7.3 Results

Both methods described in [Section 3.7](#) were successful in correctly recognising the emotions present in the texts in most cases. [Table 3.8](#) shows the classification attempts on 4 selected sentences. Both models correctly recognise the right sentiment (positive vs negative) and an appropriate emotion as well. Their performance is comparable with the RoBERTa model having an advantage in weighing the emotions recognised. In fact, if the word-based model is able to understand the primary emotion in a sentence it does not give its intensity nor its confidence. In this way, the RoBERTa model outputs every emotion with a weight that can be used to understand its intensity in the sentence.

Chapter 4

Experiments

For this project we performed two user studies, the first was aimed at evaluating the performance of our technical work on personality expression through language, specifically evaluating different text outputs from our models. In the second study, we tested the pipeline previously illustrated in [Figure 1.1](#). We further study the effects of the technical work, presented in the previous chapter, when integrated into a robot. Specifically, therefore, examining impact of the different models when used to generate appropriate emotional expressions and implemented into a robot.

4.1 Model evaluation study

The first study we conducted was aimed at assessing the performance of the models presented in [Chapter 3](#) both in terms of their ability to convey the right personality and their fluency (a common practice when evaluating NLP models). This study is, therefore, mainly concerned with answering our first research question (RQ1 in [Section 1.1](#)) on the correlation between size and performance of language models. For this purpose, we designed 3 dialogues in 3 different contexts, shown in [Table 4.1](#), and asked participants recruited online to rate them on the selected measures. We decided to use multiple dialogues as we expected the performances of our models to change depending on them. Personality cannot be understood in the same way from any dialogue and some of them favour a social trait more than others.

Given the goal of a follow-up study in HRI, the first dialogue is an introduction from a companion robot. The dialogue was designed with the aid of GPT-3 asking it to generate the dialogue by giving it the prompt: “This is a long introduction from a verbose talkative companion robot: ”. Varying

Dialogue
<p>Hello and welcome! My name is Brian, I'm here to be your robotic companion and your very own personal assistant. I come from Stockholm, in Sweden where I was created by a company called Furhat Robotics and programmed by researchers from KTH University. I was designed to help you with whatever you need, whether it's keeping you company or helping you in everyday tasks. I am equipped with a variety of sensors and a camera that allow me to see and hear what you are doing. My developers have also programmed me with sophisticated methods for recognising you so that I can always be awake when you need me but if you do not wish to have this feature I can always turn it off for you. I am always here for you, and I will never get tired of your company. I look forward to getting to know you better and to helping you in any way I can. Thank you for choosing me as your companion robot!</p>
<p>A: I am pleased to meet you and I am looking forward to working together. Before we start, I would like to get to know you better, so I am going to ask you some questions. How do you feel about being here today?</p>
<p>B: (...)</p>
<p>A: Great, I am glad to hear that! I am sure you will enjoy the session. And how do you feel about working with me?</p>
<p>B: (...)</p>
<p>A: That is good to hear, we will definitely have fun together today then. As you know, today we are going to do some exercise, do you enjoy exercising?</p>
<p>B: (...)</p>
<p>A: That makes sense, this session will be easy for you then.</p>
<p>I am here to talk to you about humanities. Humanities is the study of human culture, including history, literature, philosophy, and art. It is a broad field that covers a lot of ground, but there are a few key ideas that are essential to understanding humanities. First, humanities is about understanding the human condition. This includes understanding the way humans interact with each other and their environment. It also involves understanding the human past and how it has shaped the present. Second, humanities is about interpretation. This means that there is no one right answer to any question in humanities. Instead, scholars must use their critical thinking skills to interpret evidence and come to their own conclusions. Third, humanities is about communication. This means that scholars must be able to communicate their ideas clearly and persuasively. They must also be able to listen to and understand the ideas of others. Finally, humanities is about change. This means that the field is constantly evolving as new evidence and new interpretations are discovered. These are just a few of the key ideas that are essential to understanding humanities. If you want to learn more, there are many great resources available.</p>

Table 4.1: The three dialogues used for the pilot study.

on GPT-3 parameters, we generated 3 versions of the dialogue (please refer to [Appendix B.1](#) for a detailed report on these) that have then been re-elaborated into a single one by also integrating some additional sentences.

The second dialogue was taken from [77], a work in HRI aimed at assessing the best strategies for a robot to be more persuasive. We deem this especially relevant given our overview of personality-related literature and how it can affect the perception of a robot. Winkle et al. [77] use an exercise session scenario where the robot asks each participant in the study to complete a particular exercise with the robot’s responses selected according to 4 different strategies. Among all the strategies presented in the article we chose to use the ‘Goodwill’ one*.

The third dialogue is the presentation of a specific topic. Again, we used GPT-3 to aid the generation of the dialogue. In this case, we gave it the prompt: “This is a verbose companion robot talking about humanities: ”.

4.1.1 Study design

This study was carried out with a mixed design where each participant saw the 3 different dialogues (in random order) each with a (possibly) different model (STRAP, GPT-3 and control) \times personality (introvert and extrovert) combination. The combination for each dialogue was chosen at random in a way that also balanced all the conditions. For the style-transfer model, we ditched the GAN-based one, whose performances are clearly insufficient, in favour of a control condition or ‘expert’ model, where we hand-crafted the two personality-specific versions based on supporting literature.

We recruited through the platform Prolific[†] 5 participants for each condition (model \times personality)[‡] for a total of 30 (15 males, 15 females). Each participant saw the 3 different dialogues and, after each dialogue, they were asked to answer 14 questions posed in a randomised order each time. 10 of the 14 questions were aimed at assessing the perceived personality, 3 for the fluency and one was an attention check. 2 of the 10 personality questions were taken from PERSONAGE paper [23] and 8 from another HRI study [8][§]

* Chosen over the ‘Similarity’ one (reported as being the two most effective) due to the lower amount of interaction involved as we wanted to focus on the robot’s part of the dialogue.

[†] <https://www.prolific.co/> [‡] Since each participant saw all three dialogues, we also have five participants for each model \times personality \times dialogue combination. [§] Originally only 3 out of the 8 questions measured the introversion while 5 measured the extraversion. To balance these numbers, we swapped the question for extraversion ‘has an assertive personality’ with a similar one for introversion ‘holds back their opinion’ taken from the extraversion part of [The Items in the Big Five Aspects Scales, IPIP](#) [78].

Question number	Question	Measure
1	is introverted	Introversion
2	is reserved	Introversion
3	tends to be quiet	Introversion
4	is sometimes shy, inhibited	Introversion
5	holds back their opinions	Introversion
6	is extroverted	Extraversion
7	is talkative	Extraversion
8	is full of energy	Extraversion
9	generates a lot of enthusiasm	Extraversion
10	is outgoing, sociable	Extraversion
11	is fluent in English	Fluency
12	does not make grammatical errors	Fluency
13	is coherent and consistent with what they say	Fluency
14	click ‘Strongly disagree’ for this question	Attention check

Table 4.2: The questions used in the pilot study alongside what they measure. The questions were given in random order and participants were prompted with: “I see the speaker as someone who...”.

(questions originally from [79]). The fluency questions were designed by us following a survey on the evaluation of NLP models [80]. All the questions were posed on a 7-point Likert scale and are shown in Table 4.2 alongside what they aim to measure.

As an additional attention check to the 14th question, each participant was asked, at the end of the study, to select the topic of the dialogues they just read among 6 possibilities. The participants received a £1.20 compensation for completing the survey.

4.1.2 Hand-crafted dialogues

To craft the different dialogue versions for the control or expert condition we referred to social sciences literature [14, 15, 16, 17, 18] in order to find the characteristics that shape a text as more introverted or more extroverted. Following these studies, we gathered 8 ways of changing a dialogue to convey a personality trait rather than the other. We summarise these ways in Table 4.3 where we report the original study responsible for the claim and how the manipulation should vary between extraversion and introversion.

Number	Introvert	Extrovert	Manipulation
1	Less	More	talkative (more content and repetitions) [14, 15]
2	Less	More	positive (focus on positive aspects, leave behind the negative ones) [14, 16]
3	Less	More	subject implicit [17]
4	Less	More	self-referencing [14]
5	More	Less	complex [15]
6	High	Low	in number of hedges [14]
7	High	Low	in content negation [14]
8	More	Less	rich in vocabulary [18]

Table 4.3: List of possible manipulations of a dialogue to express different personalities.

Table 4.4 shows an example of how we manipulated the first dialogue to achieve the wanted personality manifestation, we refer to [Appendix B.2](#) for the manipulation of the other dialogues. In **Table 4.4** we highlight in *italics* the content changes, ~~strike-through~~ the removed content and add as superscript the manipulation we did among those listed in **Table 4.3** e.g. *I'm happy*² means that the content was changed to be more positive if in the extroverted dialogue (or less positive if in the introverted one).

4.1.3 Auto-generated dialogues

For STRAP generations we used the nucleus paraphrase mode with a top-p value of 0.7. It is our goal to test whether the model is actually capable of changing the personality in a dialogue, we, therefore, chose to use a high top-p value that should improve the results of the personality transfer task. However, this choice worsens the performance in terms of fluency and closeness of the paraphrased sentence from the original one. To account for this, for each dialogue, we generated 10 outputs and discarded those that contained grammatical errors or that radically changed the content of the sentence. We were confident that this selection wouldn't invalidate the unbiased evaluation of the model as this process can be easily replaced by the use of automated metrics for e.g. paraphrasing accuracy and grammatical correctness.

For the auto-generated versions of GPT-3, we took almost always the first generation of the model with the exception of some sentences in the second dialogue. The reason for this lies in the poor performance of the model when style transferring a question. For example, when asked to transfer the

Introvert	Extrovert
<p>Hello. and weleome!¹ My name is Brian, I'm here to be your robotic companion, <i>somewhat like</i>⁶ your very own personal assistant. I come from Stockholm, in Sweden¹ where I was created by a company called¹ Furhat Robotics and programmed by researehers from at¹ KTH University. I was designed to <i>sort of</i>⁶ help you with whatever you need, whether it's, <i>erm</i>⁶, keeping you company or <i>like</i>⁶, helping you in everyday tasks. I am equipped with a variety of sensors and a camera that allow me to see and hear what you are doing. My developers have also programmed me with sophisticated methods for recognising you <i>but if you do not like this feature I can turn it off for you</i>². <i>I mean</i>⁶, I use it so that I <i>am never asleep</i>⁷ when you need me. I am always here for you, and I will never get tired of your company. I <i>look forward to getting to know you better will do my best not to go wrong</i>² and to help you in any way I can. Thank you for choosing me out of <i>all the other robots</i>² to be your companion!</p>	<p>Hello and welcome! My name is Brian and <i>I'm a social robot</i>¹. <i>I'm happy</i>² to be here as your very own personal assistant and robotic companion. I come from Stockholm, in Sweden where I was created by a company called Furhat Robotics and programmed by researchers from KTH University. I was designed to help you with whatever you need.,—whether it's <i>This could be</i>^{5,8} keeping you company or helping you in everyday tasks. I am equipped with a variety of sensors and a camera that allow me^{3,5} to see and hear what you are doing. <i>My developers</i>³ I have also been programmed with sophisticated methods for recognising you so that I can always be awake when you need me, but <i>I would be happy</i>² to turn this off if you prefer⁷. <i>I use my sensors to be a better assistant and a better companion for you</i>¹. I am always here for you, and I will <i>always enjoy</i>⁷ your company. I look forward to getting to know you better and to helping you in any way I can. Thank you for having me as your companion robot!</p>

Table 4.4: An example of the manipulation of a dialogue to convey an extravert or introvert personality.

question “How do you feel about being here today?” the model would favour answering the question rather than changing its style. The only way we could find of making the dialogue progress as the others (removing some of the questions would radically change the structure of the dialogue) was to look for a generation where the model would answer the question and *then* ask what the other person thought e.g. “I don’t really feel comfortable being here today. Do you?”.

4.1.4 Results

Among the 30 participants, we excluded 7 due to failed attention checks. Among the remaining participants, 11 were males and 12 were female with an average age of 24 ($M = 24.478$, $SD = 4.104$) and with the majority reporting a medium knowledge of English (19 reported to be “Comfortable enough to understand English in most cases”, the others reported being native speakers). The data collected and the scripts used for the analysis can be found in our [GitHub Repository](#) under `pilot_data`.

Given the size of the data collected we only report qualitative results rather than quantitative. We evaluate our results on 3 key measures: extraversion, personality overall and fluency. The first is a combination of the extroverted questions and the introverted ones (with the score reversed). Using this measure, an introverted dialogue should score low on this scale while an extroverted one should be high. We will use this measure to check for a significant difference between the two versions of a single dialogue. The reverse of a measure was obtained by subtracting it from the maximum value, which is 6*. The second, personality overall, is a measure aimed at evaluating the performance in generating the appropriate dialogue i.e. if the goal is introvert, this measures the ability to create a dialogue that is perceived as introverted *and* non-extroverted. This measure is computed differently depending on the aimed personality of a dialogue. If the aimed personality is introvert, then this measure uses the introversion questions directly and the extroverted ones reversed. The opposite is true for an aimed extraversion. For the fluency questions, we first analysed each question separately as they are aimed at assessing different miss-steps of the models. However, we report here only the mean of the scores as we confirmed how their scores are ascribed very similar scores.

We show in [Figure 4.1](#) the results of our models across all dialogues.

* Note how all of the scores are on a scale from 0 to 6 rather than from 1 to 7 for ease of representation.

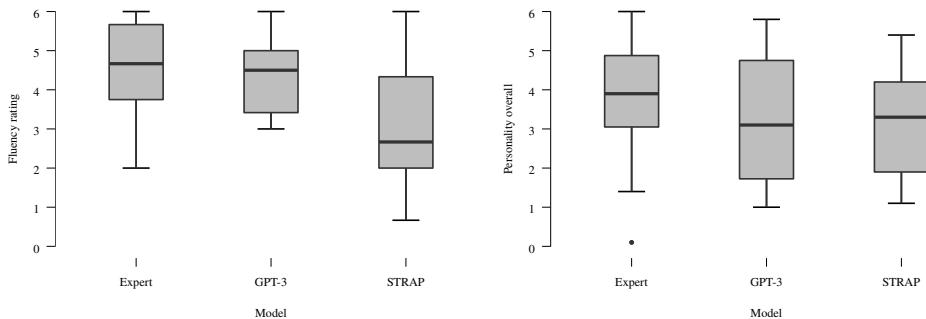


Figure 4.1: Results of the models across all the dialogues in fluency (left) and personality overall (right). In both higher is better.

We can notice how the expert version is better in both measures. The fluency measure sees our hand-crafted version ($M = 4.615, SD = 1.150$) and GPT-3 ($M = 4.394, SD = 0.924$) on top with STRAP quite behind both ($M = 2.968, SD = 1.598$). For the personality overall performance, the expert version is rate higher ($M = 3.754, SD = 1.472$) than both GPT-3 ($M = 3.218, SD = 1.645$) and STRAP ($M = 3.076, SD = 1.328$) with the difference between these two much lower.

[Figure 4.2](#) shows the extraversion rating on each dialogue separately. In each graph we have split the rating according to the dialogue version (introvert or extrovert), scores for the same model are connected through a line. With these graphs, we want to check the difference in personality between the dialogues generated by a single model. We can, therefore, define a good result when there is a big difference in the extraversion rating of the two versions *and* the higher score is in the extroverted version (shown on the left of each plot).

In the graphs, we can see how the STRAP model is able to convey a bigger difference in personality in the first dialogue (the introduction) than the other two models. On the other hand, it falls behind the others in both dialogue 2 and dialogue 3. We can observe how STRAP starts reporting worse performance in the second dialogue (the conversation) where the extroverted version is perceived as only slightly more introverted than the actual introverted one. In the third dialogue (the one about humanities) both GPT-3 and STRAP are ascribed the wrong personality in both generated versions with STRAP reporting the biggest error between the two. The expert versions correctly always convey the wanted personality and the performance is consistent across all dialogues.

Further, we can see how STRAP has the general tendency to generate dialogues that are more introverted (or less extroverted) compared to GPT-3,

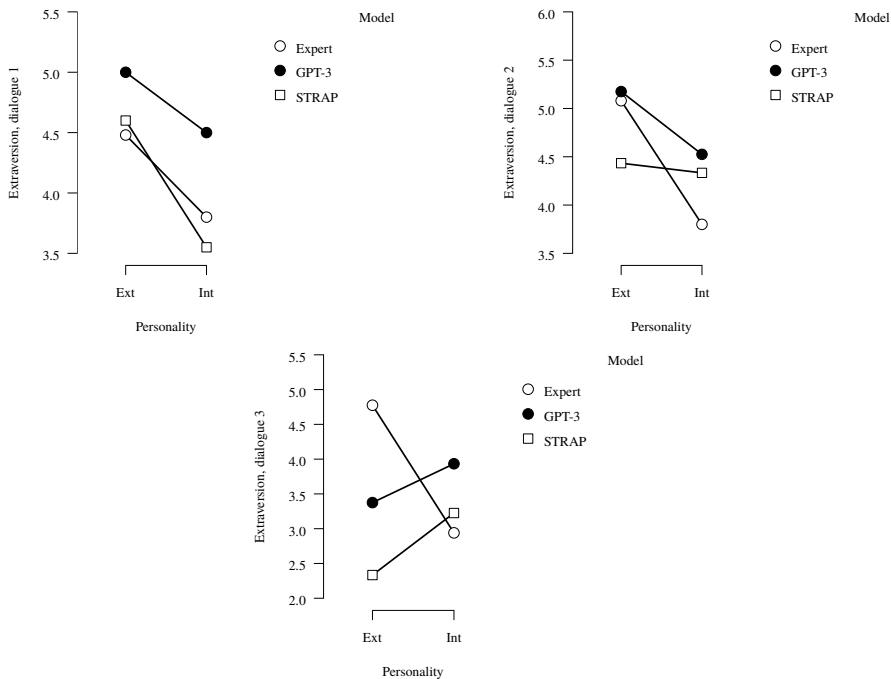


Figure 4.2: Extraversion rating for each model \times personality for dialogue 1 (top left), dialogue 2 (top right) and dialogue 3 (bottom). Better as the Ext score gets higher than the Int score. No error bars are displayed to favour readability.

which, on the other hand often generates dialogues higher in extraversion.

4.2 User study

Once we established the correct perception of the personality for (some of) our dialogues, we designed a follow-up study where we wanted to test the impact of our automated models when used as input for the speech of a robot in a Human-Robot Interaction. We further want to explore the impact of also linking emotions to that speech through our automatic emotion extraction method. We expect the inclusion of emotions to exaggerate, at least in some way, the personality manifestation. We further want to test the difference in performance of our models when moving from a text-only context to a multimodal one. Expecting a difference in both fluency and personality rating.

For this purpose, we chose to use the robot Furhat. This choice has been made to explore the role of facial expressions (rather than body movements) in the context of emotions where most previous studies focused on the robot NAO and its body movements. Furthermore, Furhat is rated to be one of the most

anthropomorphic robots available according to the ABOT Database* [81], definitely rated higher than NAO (63.43 vs 45.92 on a scale of 0-100). An increase in human-likeness favours the inclusion of more human traits such as personality as the user is already expecting a more familiar interaction. This also allows for a better character definition, hence better supporting our goal of naturally communicating human-like personality traits.

4.2.1 Emotions and personality manifestation

For this study, we decided to use the first dialogue of those tested with the pilot given its “success” in differentiating one personality from the other in all of our models. Further, we only limit our scope to the automatic models (GPT-3 and STRAP) as this thesis is primarily concerned with evaluating the potential for such automatic methods. Further, we only used the default Text-to-Speech (TTS) and we did not control for the prosody of the speech.

The emotional manifestation is directly extracted from the text the robot should be animating, that is, the output from our language models. Labels output from our emotion classification system are then mapped to Furhat’s default facial expressions for anger, disgust fear, surprise and sadness in addition to a custom ‘joy’ option which combines a smile, cheek puff and openness of the jaw. All of these expressions can take the intensity between 0 and 1 as input. The duration of each emotion follows the sentence from which it was extracted. For the emotion generation, we tested both the word-based approach and the fine-tuned RoBERTa. Both performed really well but the latter appeared much more natural as the scores for the emotions are given on a continuous scale. This allows the model to recognise sentences that have little to no emotions with very low scores while this is not possible with the other approach. For further visualisation of this effect we invite the reader to watch the videos under [study_data/emotion_comparison](#) in our [GitHub Repository](#). After choosing the RoBERTa fine-tuned as our emotion recognition method we also tweaked its output score by halving it in order to further increase the naturalness. In fact, while we can deem natural a high emotional expression (intensity=1) in a static context, it appears particularly unnatural if the person expressing that is also talking (which is our case).

* <http://www.abotdatabase.info/collection>

4.2.2 Research questions and hypothesis

With this study, we want to test among the research questions presented in [Section 1.1](#), the research questions from 2 to 4. We repeat them here for ease of reading alongside with our hypothesis of the expected result:

RQ2 (How) does multi-modal robot delivery of speech influence the performance of different-sized models in terms of fluency, personality and emotion manifestation? Does fluency and personality manifestation vary compared to perceptions of the pure text output?

H2 The personality perception remains the same while all the models are perceived as more fluent. In terms of emotions, a more extroverted dialogue generates more emotions in number and intensity. The difference is perceivable by the participant.

RQ3 (How) does the perception of uncanniness vary depending on the personality of the dialogue and the personality of the participant? Does this vary across the different language models we evaluate?

H3 The uncanniness is higher with less fluent models and more introverted dialogues. It does not correlate with the personality of the participant.

RQ4 (How) does the likeability of a robot change depending on the personality of the dialogue and the personality of the participant? Does this vary across the different language models we evaluate?

H4 The likeability of the robot is increased when the personality of the participant matches that of the dialogue.

4.2.3 Study design and measures

In order to answer the questions posed above, we designed a video-based online study. We recorded 4 videos varying in our experimental conditions: the model used for the personality expression (GPT-3 or STRAP) and the personality (introvert or extrovert) for a 2x2 between-subjects study design. The clips were shot using the Virtual Furhat SDK rather than recording a real Furhat to favour the correct viewing of all of the facial expressions (used to convey the emotions) of the robot*. The videos are available for watching under [study_data/videos](#) in our [GitHub Repository](#). We also show two screenshots of the clips in [Figure 4.3](#).

* Recording a real Furhat often worsens the quality of the face and differences in brightness/contrast in the scene can worsen the correct reception of its facial movements.



Figure 4.3: Two screenshots of the videos used for the study.

We used Prolific to recruit 40 men and 40 women (through its screening tools), 20 participants per condition (equally distributed across genders). Each participant in the study is initially asked to compile a short questionnaire to assess their personality (used to check for personality matching). The questionnaire has been extracted from the one used in the pilot (shown in [Table 4.2](#)) where we removed the questions on fluency and the questions explicitly asking the introversion/extraversion. Each participant then watches a video with the robot animating the first introduction dialogue from the pilot (see [Table 4.1](#)) and manifesting congruent emotional expressions (the video was available to re-watch at every step of the questionnaire). For a direct comparison with the pilot, the participant is first asked to rate the same questions in [Table 4.2](#), again, on a 7-point Likert scale. We also used questions from the Godspeed questionnaire [82] for the measures of Anthropomorphism, Likeability and Perceived Safety and questions of Warmth and Discomfort from the RoSAS questionnaire [83], all presented on a 5-point Likert scale. We show these additional questions in [Table 4.5](#).

We use the combination of Godspeed Anthropomorphism and Perceived Safety, RoSAS Warmth and Discomfort as a proxy for uncanniness. We also use the questions from RoSAS Warmth as a measure of emotional expression.

We included in the survey different types of attention checks, 3 of them were among the questions, of the type “click disagree/strongly agree for this question”. We also asked each participant at the end what was the name of the robot and the topic of the dialogue. Participants could choose among 6 possibilities for each. The participants received a £1.50 compensation for completing the survey.

Measure Low	Measure High	Origin
Fake	Natural	Godspeed Anthropomorphism
Machine-like	Human-like	Godspeed Anthropomorphism
Unconscious	Conscious	Godspeed Anthropomorphism
Artificial	Lifelike	Godspeed Anthropomorphism
Moving rigidly	Moving elegantly	Godspeed Anthropomorphism
Dislikable	Likable	Godspeed Likeability
Unfriendly	Friendly	Godspeed Likeability
Unkind	Kind	Godspeed Likeability
Unpleasant	Pleasant	Godspeed Likeability
Awful	Nice	Godspeed Likeability
Anxious	Relaxed	Godspeed Perceived Safety
Calm	Agitated	Godspeed Perceived Safety
Quiescent	Surprised	Godspeed Perceived Safety
-	Happy	RoSAS Warmth
-	Feeling	RoSAS Warmth
-	Social	RoSAS Warmth
-	Organic	RoSAS Warmth
-	Compassionate	RoSAS Warmth
-	Emotional	RoSAS Warmth
-	Scary	RoSAS Discomfort
-	Strange	RoSAS Discomfort
-	Awkward	RoSAS Discomfort
-	Dangerous	RoSAS Discomfort
-	Awful	RoSAS Discomfort
-	Aggressive	RoSAS Discomfort

Table 4.5: The questions used in the second study in addition to those in [Table 4.2](#). The participants were prompted with: “Please indicate the extent you think the robot is...”.

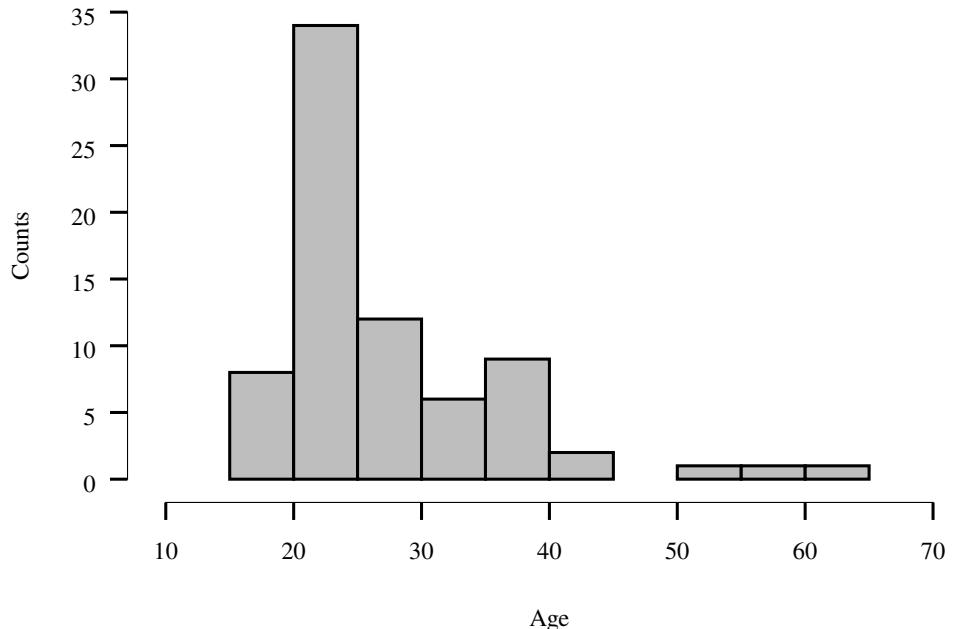


Figure 4.4: Distribution plot of the age of the participant in our final study.

4.2.4 Results

Among the 80 recruited participants, we excluded 6 due to failed attention checks. The remaining population was composed of 38 women, 35 men and one that did not self-identified among our options (chose ‘other’). The average age is 28 ($M = 28.11$, $SD = 9.04$) and it is plotted in Figure 4.4. 18 participants reported being native speakers of English while 56 reported being “Comfortable enough to understand English in most cases”. Also, following our self-report questionnaire on personality, 44 participants were on the higher end of the extraversion trait (extroverts) while only 30 were on the lower end (introverts).

All the statistical analysis that follows was made using the open-source software JASP [84]. The data collected and the analysis done are available under [study_data](#) in our [GitHub Repository](#).

4.2.4.1 RQ2 - Manipulation check

Our first analysis was aimed at verifying the correct manipulation of the personality in the dialogues i.e. the extroverted dialogue is perceived as significantly more extroverted than the introverted one. Similarly to what was done in the pilot, to compare the dialogues we only use the extraversion score

which is an average of the introversion (reversed) and extraversion questions.

For each of our models, we run two independent samples tests, comparing their two versions of the dialogue. For STRAP, we used the Student t-test and discovered that the difference in extraversion is not significant ($t(34) = 1.392, p = 0.086$).

For GPT-3 we run two Mann-Whitney U tests due to the violation of the normality assumption (shown by a Shapiro-Wilk test, $W = 0.898, p = 0.044$ in the extroverted dialogues). GPT-3, similarly to STRAP, does not generate a significant difference between the two dialogues ($W = 218.5, p = 0.076$). However, plots relative to the extraversion, shown in [Figure 4.5](#), still suggest that generally both introverted dialogues are perceived more introverted and the extroverted dialogues are perceived more extroverted.

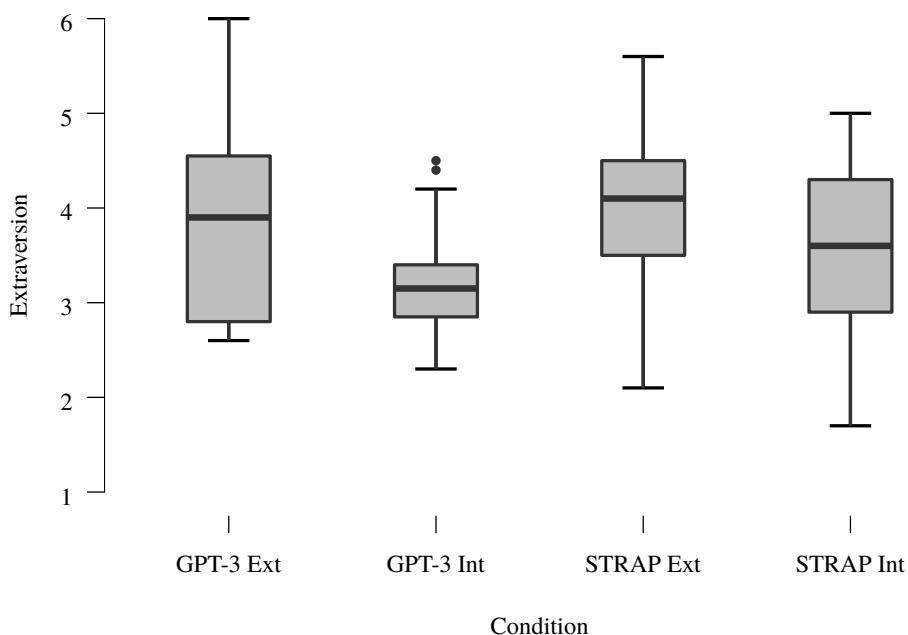


Figure 4.5: Plots with the ascribed extraversion of the dialogues on a scale from 0 to 6.

A subsequent Kruskal-Wallis test (done in the place of Analysis of Variance (ANOVA) due to the violation of normality) including both models (therefore comparing only the extroverted dialogues against the introverted ones) revealed a significant difference in extraversion ($H(1) = 4.388, p = 0.036$), confirmed by Dunn's post-hoc test ($p = 0.018$).

4.2.4.2 RQ2 - Difference in fluency and performance

To answer the research questions that follow we did a series of independent samples ANOVAs (or Kruskal-Wallis if the normality or equality of variance assumptions were violated) comparing the relevant measures across model and personality (introverted or extroverted dialogue).

A Kruskal-Wallis test revealed a significant difference in fluency across all conditions ($H(3) = 27.509, p < 0.001$), across the personality ($H(1) = 5.124, p = 0.024$) but not across models ($H(1) = 2.998, p = 0.083$). A post-hoc Dunn's test confirmed the extroverted dialogues to be ascribed more fluency than the introverted ones ($p = 0.012$). A Dunn's test comparison across all conditions also revealed GPT-3 extroverted dialogue to be significantly more fluent than STRAP introverted dialogue ($p = 0.013$ using Bonferroni or Holm correction).

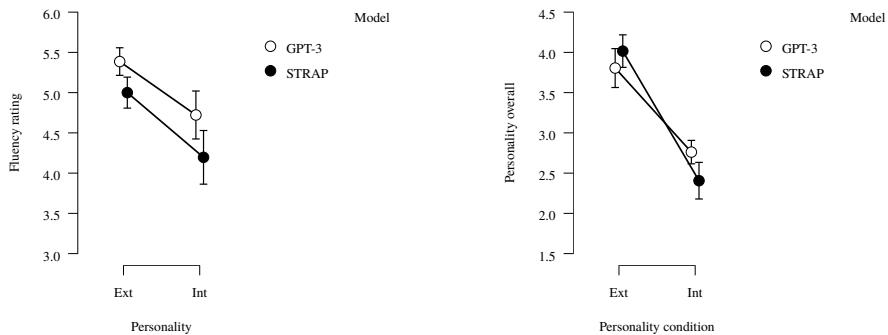


Figure 4.6: Plots with the ascribed fluency of the dialogues (left) and the personality performance (right) on a scale from 0 to 6.

The same Kruskal-Wallis test was repeated for personality overall performance*, revealing a significant difference across all conditions ($H(3) = 27.509, p < 0.001$), the different version of the dialogues ($H(1) = 25.863, p < 0.001$) but not across the models ($H(1) = 0, p = 1$). Again, a post-hoc Dunn's test reveals differences across the personality condition ($p < 0.001$ with Holm or Bonferroni correction) with the extroverted dialogues performing better (Figure 4.6) than the introverted ones.

In this context, both our models perform worse on generating introverted dialogues, this is reflected in the personality manifestation of each version where the introverted one fails to be ‘enough’ introverted and is therefore

* Similarly to the pilot, the personality overall performance is computed by using both the introversion and extraversion measure i.e. it measures how much the introverted dialogue is introverted *and* non-extroverted, vice-versa for the extroverted dialogue.

closer to the extroverted one. Ideally, to have the biggest difference in personality manifestation between the two dialogues, the performance score should be of similar height for both version with the difference growing alongside the growth of the two scores.

4.2.4.3 RQ2 - Emotional manifestation

Figure 4.7 shows the intensity of the emotions expressed by each version of the dialogue. This includes all emotions, although the major one manifested by each dialogue is joy. We see from the figure how the extroverted versions of each model generated higher intensity emotions.

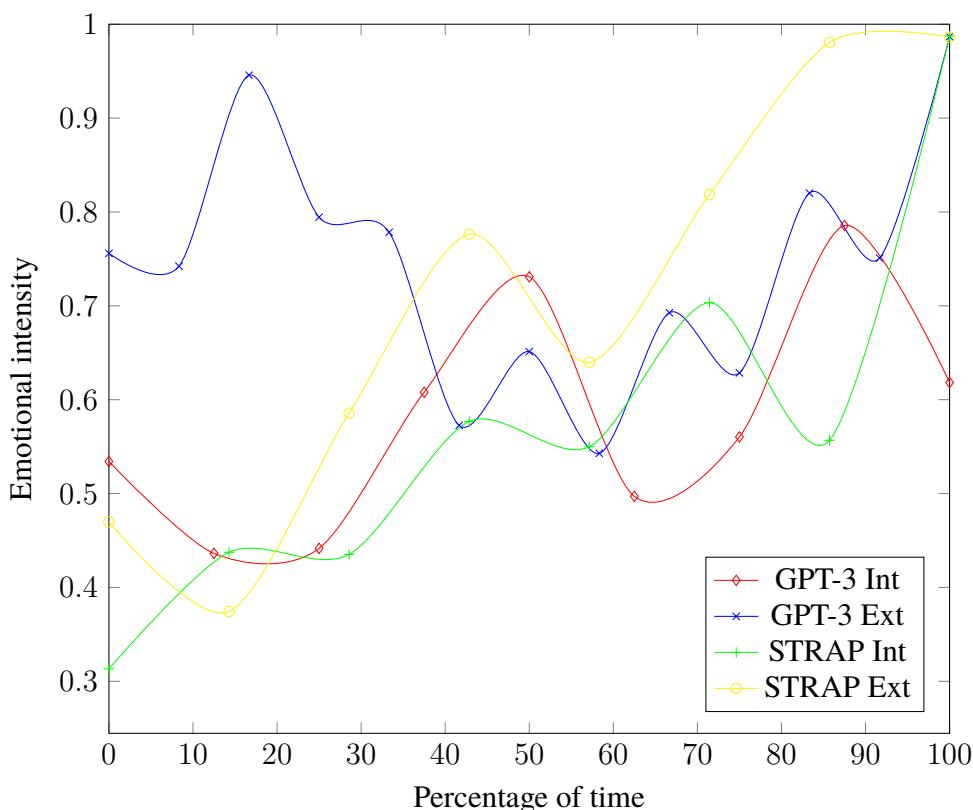
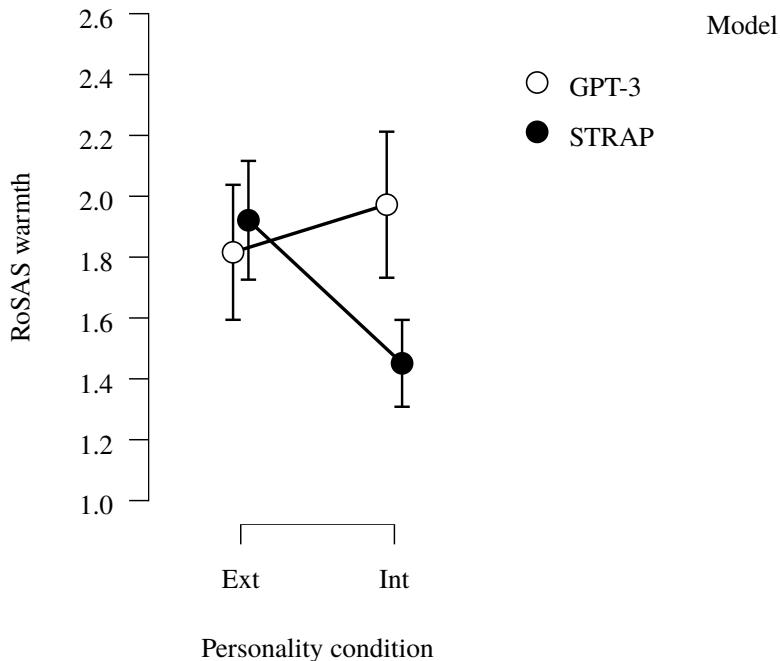


Figure 4.7: Progression of the emotional manifestation intensity throughout the dialogue.

Using the data from the study, ANOVA testing showed no significant differences across all conditions ($F(3, 69) = 2.326, p = 0.132$), personality ($p = 0.448$) or model ($p = 0.315$) in the emotional perception. However, a independent samples Student t-test with only STRAP results revealed its

extroverted dialogue to be significantly more emotional than the introverted one ([Figure 4.8](#)): $t(34) = 1.906, p = 0.033$.



[Figure 4.8](#): Plot with the ascribed RoSAS warmth measures as a proxy for emotional manifestation on a scale from 0 to 4.

4.2.4.4 RQ3 - Perception of uncanniness

A Kruskal-Wallis test on the uncanniness ([Figure 4.9](#)) revealed no significant difference across all conditions ($H(3) = 5.941, p = 0.115$), across the personality condition ($H(1) = 0.454, p = 0.501$) or the model ($H(1) = 3.078, p = 0.079$).

As our uncanniness measure is a combination of different measures that may be perceived differently, we deemed relevant to also test each measure separately. The only measure that varies significantly across our conditions is the Godspeed Anthropomorphism* where the difference, tested with ANOVA, is only significant across model *and* personality ($F(3, 69) = 2.606, p = 0.04$) and only close to significance across models only ($F(3, 69) = 3.037, p = 0.086$). In [Figure 4.9](#) we can see how the extroverted dialogue from STRAP

* NB: The measure of Anthropomorphism is not on the same scale of uncanniness i.e. high anthropomorphism does not necessarily result in high uncanniness or vice-versa.

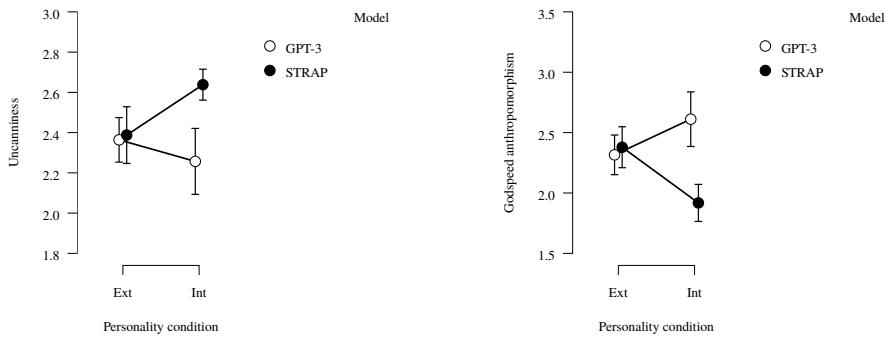


Figure 4.9: Plots with the ascribed uncanniness of the dialogues (left - lower is better) and the Godspeed Anthropomorphism (right - higher is better). Both are on a scale from 0 to 4.

is very close to that of GPT-3 while the introverted version is quite lower. A pairwise post-hoc test using Tukey's correction confirmed GPT-3 introvert to be significantly higher than STRAP introvert ($p = 0.047$).

The personality of the participant does not have any significant effect on the ascribed uncanniness (ANOVA, $F(7, 65) = 0.504, p = 0.480$).

4.2.4.5 RQ4 - Likeability

A Kruskal-Wallis test showed no significant difference in likeability neither across all the conditions ($H(3) = 1.832, p = 608$), the models ($H(1) = 0.177, p = 0.674$) or the personality ($H(1) = 1.355, p = 0.244$). We can see, in [Figure 4.10](#), how GPT-3 and the extroverted dialogues are ascribed slightly more likeability but not enough to represent statistical significance.

The personality of the participant does not have any significant effect on the personality perception (Kruskal-Wallis, $H(7) = 5.826, p = 0.560$).

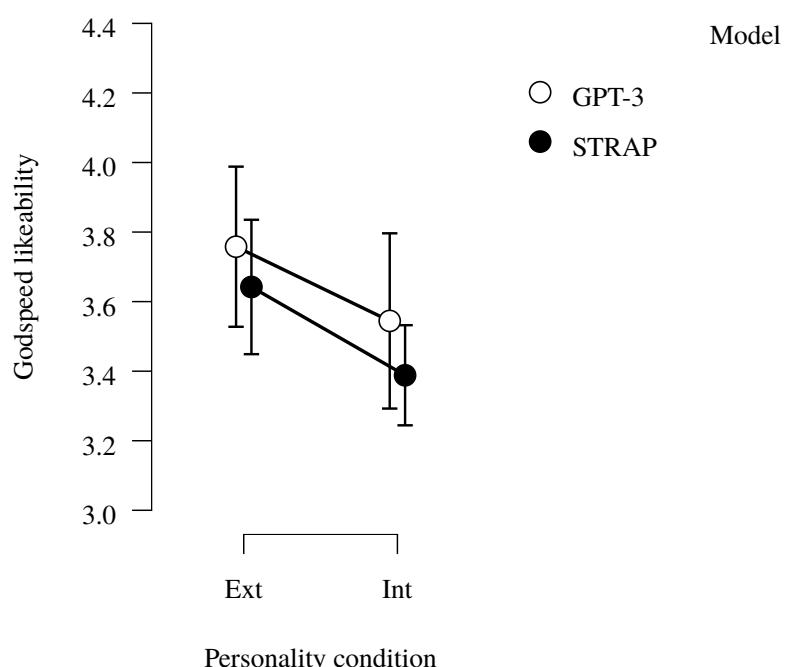


Figure 4.10: Plot with the ascribed Godspeed Likeability on a scale from 0 to 4.

Chapter 5

Discussion

In this chapter, we summarise the most important results from this thesis and discuss their meaning and the consequences.

5.1 Initial findings

Our first discussion is mainly concerned with the technical findings and is generally not connected to any explicit research question. In this section, we comment on some results obtained during the development of the technical side of this work. Here, we also present some limitation of our models.

5.1.1 Understanding personality from single sentences is hard

In Section 3.3.4.2 we tested the personality classification from text using single sentences or multiple ones. The only meaningful results we got were when testing groups of sentences from the same person. On the other hand, classifying single sentences, gave close to random results. It is our opinion that future works trying to classify the personality from text should gather more than a few sentences from a single person and base the classification on all of those. It is also important to remark how the classification was never the primary goal of this thesis but has been explored as a result of the architecture of the first model tested.

Future works might look to further develop personality classification in order to extend our pipeline to robot personality generation being based on that of the user.

5.1.2 Bigger models are better at extending sentences but not at paraphrasing

In Chapter 3 we reported the results on the three proposed methods varying on the number of parameters. Starting from the model with the lowest amount, the GAN-based method, we can notice how it is completely unable to successfully complete the task. While the method has been proved successful in other domains, such as the sentiment transfer [46], it was not powerful enough to be applied to ours. Reasons for this can be found in the nature itself of the task. To change the sentiment of a sentence it is enough to understand which words bring the positive/negative attribution and swap them with their antonyms. However, the same reasoning cannot be applied when trying to transfer the personality, where the changes are a lot more subtle and require a better understanding of language and how it can be manipulated without altering the meaning of a sentence.

The other two models seemed to be behaving better, at least in terms of fluency of the sentences they were transferring. If we directly compare the two we can see how STRAP mostly tries to shape the personality by paraphrasing the sentence, changing the wording and the length of the sentence. GPT-3 rarely does any changes to the original sentences but the personality manifestation mainly comes from added sentences. These added sentences are very well placed and integrate naturally into the dialogue but also help providing great insight into the person (robot) who is talking.

5.2 Research questions answered

In the next section, we will be concerned with the discussion of the results from our user studies and how they can be used to answer the research questions initially formulated (see Section 1.1).

5.2.1 Personality manifestation depends on the topic

From the pilot results, we can confirm that these automated models are able to work in some contexts (although not much in others). Personality manifestation and its perception are really hard tasks both for humans and automated methods and not all contexts permit the correct reception of it. Further, when the interaction is limited to only a single dialogue, the correct reception of the personality highly depends on the topic of that dialogue. A more ‘personal’ dialogue, such as an introduction or when talking about

oneself has a much greater manifestation of personality than a dialogue about an objective topic such as that on humanities. We can even report having a harder time making the hand-crafted version of the last two dialogues than for the first, the introduction dialogue. Even though the hand-crafted versions of the last two dialogues still report the correct perception of the aimed personality, we speculate this may only be due to a very high exaggeration of the personality manifestation that would not probably happen in real-world scenarios.

Let us also compare the performance of the two models overall and in terms of fluency. While the fluency score is significantly higher for GPT-3, which almost reaches the performance of our hand-crafted versions, the overall performance is very close between the two. We have to consider, however, that the outputs of STRAP have been manually filtered to increase fluency. If dropping this filtering we would probably also have to drop the top-p value by at least 0.1 or 0.2 to remove the manifestation of less natural or badly paraphrased sentences. We speculate that this would also reduce the manifestation of the personality, at least to some extent.

This answers our first research question on the relation of size and performance of language models. Despite GPT-3 being much larger than STRAP, the performance is comparable, with GPT-3 having a bigger edge only in the fluency measure.

5.2.2 Personality manifestation in robotic speech is (almost) significant

To begin our follow-up study, we checked that our manipulation of the personality was successful. Our tests indicate that the difference in extraversion is not statistically significant. However, given the closeness to a significant value, our results on the pilot and the qualitative results on the study we hypothesise that our manipulation checks were successful nonetheless. Our sample size was fairly small and it is possible that a more subtle change in personality may require a higher number of participants to reach statistical significance.

If we try to examine where our models could improve by looking at their performance, we can see how both models struggle quite a bit with the introverted dialogue. While they achieve better results in the extroverted one. Reasons for this may be due to the intrinsic way in which a dialogue can appear more introverted rather than extroverted. If we think of some of the manipulations that can be made to a dialogue to make it seems more

extroverted (see [Section 4.1.2](#)), there is ‘being more talkative’ [14, 15], ‘have less complex sentence’ [15], etc. Both of which are much harder to apply in the introversion case. When making a dialogue more talkative, one can repeat some of the content, or add new one that closely relates to the dialogue. Making it less talkative is not a mere removal of content because that would change radically the meaning of the sentence, one should understand the key concept in a dialogue and try to remove anything else, which is much harder. Similar reasoning can be applied to complexity, where making a less complex sentence is easier than making a more complex one.

5.2.3 GPT-3 is more fluent but not better performing than STRAP

Our results in RQ2, show how there is a significant difference in fluency across the personality condition with the extroverted dialogues being perceived as more fluent. This is somewhat expected as an extroverted personality is generally more talkative, use less complex sentences and fewer hedges which may be perceived as also being more fluent. Also, STRAP is generally ascribed less fluency than GPT-3 (although not significant), a factor that combined with the difference between personality makes GPT-3 extrovert significantly more fluent than STRAP introvert.

If we consider the performance, on the other hand, there is absolutely no difference between the two models, suggesting the advantage of GPT-3 in the task mostly lies in its fluency. This goes against our initial findings in the pilot where we could see a bigger difference between the two suggesting that multi-modal speech delivery has helped closing the gap between the two (contrary to what we hypothesised in H2).

Further, plots for uncanniness show how there is a difference in perception between the models, at least for some of their generation. However, the difference is not enough to represent statistical significance. Reasons for this can be multiple, ranging from the (short) duration of the interaction to the too small changes between one robot and the other to be perceived by the participant. This also goes against our hypothesis for our third research question where we expected fluency to have an effect on uncanniness as it should reflect on how ‘human’ and ‘natural’ the robot is perceived.

5.2.4 Likeability and personality matching effects are too little to measure

Our results on likeability and personality matching do not confirm our hypothesis that personality matching improves the interaction and the likeability of the robot. We explain these results with the relatively short duration of our experiment. In fact, to register a positive difference (or any difference at all), the user would need to better understand the robot, its purpose, or how it behaves. Possibly through a more active session that involves the user directly.

5.2.5 Emotional manifestation depends on personality (but maybe not)

We have shown, in the previous chapter, how text of different personality triggers a different emotional manifestation with the extroverted ones triggering a higher intensity in emotions confirming our hypothesis H2. However, our results suggest a significant difference only in the case of STRAP.

However, in this experiment, many could be the confounding factors. The perceived emotions could be different due to the slightly different content of the dialogue or due to other differences that should not play a role. Further, while STRAP extroverted dialogue has been perceived as more emotional, the opposite is true for GPT-3 even though the intensity is generally higher for GPT-3 extrovert. We, therefore, think that to better back our claim that a change in textual personality is enough to manifest significantly different emotions, more experiments with different dialogues in different contexts are needed.

Chapter 6

Conclusions

This thesis explored different approaches to automatic processing of text to convey different personalities, with particular attention to the extraversion trait. We proposed the use of three different language models varying in the size of their architecture: a GAN-based model, STRAP and GPT-3. Our initial results suggest that our smallest model is not able to successfully complete the task. The other two models perform very similarly in terms of personality manipulation but the largest model (GPT-3) outperforms the other (STRAP) in terms of fluency.

We showed, through an online pilot study how different kinds of dialogues provided only as text can be shaped to convey a personality trait rather than another. We confirmed that the difference is perceptible by people through hand-crafted versions of the dialogues which have always been correctly ascribed the aimed personality. The performance of our automated models, STRAP and GPT-3, varies depending on the type of dialogue. We do not deem this a limitation of our models but rather a limitation of how much a dialogue can change to convey a personality.

After comparing the performance of our models, we worked on incorporating the dialogues into the social robot Furhat. We also used a method to extract emotions from the dialogues' text to be then manifested through the facial expressions of the robot. We tested through a second online study the perception of the robot when animating those dialogues with appropriate emotions.

Our second study revealed how the dialogues were generally ascribed the correct personality but how this was not statistically significant (possibly due to small sample size). Our second study confirmed how STRAP is sometimes perceived as less fluent than GPT-3 (significant) but how they perform very

similarly. We found out that both models perform better when generating an extroverted dialogue than an introverted one (significant). The dialogues from STRAP made for a robot that is perceived slightly more uncanny than GPT-3 (not significant) although, much of the difference is due to the introverted dialogue from STRAP.

We did not record any (significant) difference in the likeability of the robot. Similarly, we did not find that the personality of the participant had any influence on the perception of the robot. This goes against other findings in the literature [60, 61, 8], but likely links to the short length of our participant-robot interactions and the video-based nature of our study. Finally, although both models trigger a bigger emotional manifestation in the extroverted version of the dialogue, only STRAP resulted in a significant difference in the amount of emotion generated for the introvert versus extrovert robot dialogue - something that likely contributed to increased differences in perception of those versions of the robot and therefore worth investigating further in future work on automatic personality generation and expression.

6.1 Future work and limitations

Points for improvement and extension in this work are countless as the topics addressed are very broad and pose many, currently unsolved, problems.

Starting from the datasets, these impose a big limitation in terms of the performance achievable. A more “personality-expressive” dataset like the PERSONAGE one can help to create a more congruent model. However, the PERSONAGE dataset itself cannot be used directly due to its limited size. The MBTI dataset is very big but also very ineffective (according to our testing) due to its limited scope and the very informal setting in which it was collected. The Essays dataset is very expressive in the personality, given its very personal nature but it is very limited in the topic and in its size. Furthermore, both MBTI and Essays contain ‘made up’ words that worse the performance of any model they are trained on. A way of gathering a dataset that may improve the performance of the methods proposed could be labelling a NLP dataset using a personality classification model, given that its performance reaches the level needed for such a process.

When comparing the performance of our two best language models, STRAP and GPT-3, we observed how they have two different ways of processing a text to express personality. STRAP heavily relies on changing the wording and paraphrasing while GPT-3 mostly adds/removes content from the dialogues. Future works could explore the use of these two models in

combination to achieve even better performances.

Also, when testing GPT-3 we only tried a prompt-based approach with zero or one-shot learning. Compared to our STRAP approach where we took another language model and fine-tuned it for our tasks. As it is also claimed in [openAI blog*](#), this approach is preferable and often yields better performance than a prompt-based one.

As we outlined in many parts of this manuscript, we only focused our work on the extraversion trait as it is considered to be easier to manifest through language and has more literature [14, 15, 16, 17, 18] (especially in HRI [7, 8, 12]) supporting it. We believe, however, that our proposed methods should still work even on different personality traits. For STRAP, we can imagine the creation of a pipeline of 5 models to create the wanted personality (one for each trait). A similar strategy can also be applied to GPT-3. We can expect, when expanding to the whole personality spectrum, that the manifestation of some traits would be more evident depending on the context. With some traits having a bigger impact on dialogues where others have little to no impact. For example, we can expect, for our dialogue on humanities, that the conscientiousness or openness traits to have a much bigger impact than the extraversion one.

When it comes to the HRI side of this work, it probably would benefit from a more in-depth and in-person study where one can accurately measure the difference in likeability of the robot and the possible effect of personality matching on the interaction. However, when integrating these methods into a conversational robot, it has to be considered that in this work we have always pregenerated the dialogues used as well as their different personality versions. Starting from the latter, our experiments revealed that GPT-3 is much faster than STRAP in outputting the desired personality-adapted dialogue, mostly due to the more powerful servers used. This implies that while GPT-3 is ready to be used in a dynamic environment where the response of the robot can be computed on demand, STRAP needs further optimisation or a stronger server base. Another limitation of this work that could be explored in future development is the lack of an automated method for generating the base dialogues. In fact, our pipeline presented at the very beginning of this work (see [Figure 1.1](#)) start with a given dialogue.

* <https://openai.com/blog/customized-gpt-3/>

References

- [1] Maha Salem et al. “Generation and evaluation of communicative robot gesture.” In: *International Journal of Social Robotics* 4.2 (2012), pp. 201–217.
- [2] Maha Salem et al. “To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability.” In: *International Journal of Social Robotics* 5.3 (2013), pp. 313–323.
- [3] Candace L. Sidner et al. “Where to look: a study of human-robot engagement.” In: *Proceedings of the 9th international conference on Intelligent user interfaces*. 2004, pp. 78–84.
- [4] Daniel Ullrich. “Robot personality insights. Designing suitable robot personalities for different domains.” In: *i-com* 16.1 (2017), pp. 57–67.
- [5] Amir Aly, Adriana Tapus, et al. “Robot personality design for an appropriate response to the human partner.” In: *Feedback Readability for Robots Workshop (in Conjunction with the 21st IEEE International Symposium on Robot and Human Interactive Communication” Ro-Man”)*. 2012.
- [6] Hossein Mobahi and Shahin Ansari. “Fuzzy perception, emotion and expression for interactive robots.” In: *SMC’03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*. Vol. 4. IEEE. 2003, pp. 3918–3923.
- [7] Amir Aly and Adriana Tapus. “A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction.” In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, pp. 325–332.
- [8] Antonio Andriella et al. “Do I have a personality? Endowing care robots with context-dependent personality traits.” In: *International Journal of Social Robotics* (2020), pp. 1–22.

- [9] Maurizio Ficocelli, Junichi Terao, and Goldie Nejat. “Promoting interactions between humans and robots using robotic emotional behavior.” In: *IEEE transactions on cybernetics* 46.12 (2015), pp. 2911–2923.
- [10] Angelica Lim, Tetsuya Ogata, and Hiroshi G. Okuno. “Converting emotional voice to motion for robot telepresence.” In: *2011 11th IEEE-RAS International Conference on Humanoid Robots*. IEEE. 2011, pp. 472–479.
- [11] Hannah R.M. Pelikan, Mathias Broth, and Leelo Keevallik. “‘Are you sad, Cozmo?’ How humans make sense of a home robot’s emotion displays.” In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 2020, pp. 461–470.
- [12] Lilia Moshkina et al. “TAME: Time-varying affective response for humanoid robots.” In: *International Journal of Social Robotics* 3.3 (2011), pp. 207–221.
- [13] Meng-Ju Han, Chia-How Lin, and Kai-Tai Song. “Robotic emotional expression generation based on mood transition and personality model.” In: *IEEE transactions on cybernetics* 43.4 (2012), pp. 1290–1303.
- [14] James W. Pennebaker and Laura A. King. “Linguistic styles: language use as an individual difference.” In: *Journal of personality and social psychology* 77.6 (1999), p. 1296.
- [15] Adrian Furnham. “Language and personality.” In: (1990).
- [16] Avril Thorne. “The press of personality: A study of conversations between introverts and extraverts.” In: *Journal of Personality and Social Psychology* 53.4 (1987), p. 718.
- [17] Francis Heylighen and Jean-Marc Dewaele. “Variation in the contextuality of language: An empirical measure.” In: *Foundations of science* 7.3 (2002), pp. 293–340.
- [18] Jean-Marc Dewaele and Adrian Furnham. “Extraversion: The unloved variable in applied linguistic research.” In: *Language Learning* 49.3 (1999), pp. 509–544.
- [19] Maike Paetzel-Prüssmann, Giulia Perugia, and Ginevra Castellano. “The influence of robot personality on the development of uncanny feelings.” In: *Computers in Human Behavior* 120 (2021), p. 106756.

- [20] Michael L. Walters et al. “Avoiding the uncanny valley: robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion.” In: *Autonomous Robots* 24.2 (2008), pp. 159–178.
- [21] Meeri Mäkäräinen, Jari Kätsyri, and Tapio Takala. “Exaggerating facial expressions: A way to intensify emotion or a way to the uncanny valley?” In: *Cognitive Computation* 6.4 (2014), pp. 708–721.
- [22] François Mairesse and Marilyn A. Walker. “PERSONAGE: Personality generation for dialogue.” In: *Proceedings of the 45th annual meeting of the association of computational linguistics*. 2007, pp. 496–503.
- [23] François Mairesse and Marilyn A. Walker. “Controlling user perceptions of linguistic style: Trainable generation of personality traits.” In: *Computational Linguistics* 37.3 (2011), pp. 455–488.
- [24] Emily M. Bender et al. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? .” In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 610–623.
- [25] Maarten Selfhout et al. “Emerging late adolescent friendship networks and Big Five personality traits: A social network approach.” In: *Journal of personality* 78.2 (2010), pp. 509–538.
- [26] Timothy A. Judge and Jeffery A. LePine. “The bright and dark sides of personality: Implications for personnel selection in individual and team contexts.” In: (2007).
- [27] Andrew Neal et al. “Predicting the form and direction of work role performance from the Big 5 model of personality traits.” In: *Journal of Organizational Behavior* 33.2 (2012), pp. 175–192.
- [28] Rebecca Shiner and Avshalom Caspi. “Personality differences in childhood and adolescence: Measurement, development, and consequences.” In: *Journal of child psychology and psychiatry* 44.1 (2003), pp. 2–32.
- [29] Charles F. Halverson Jr et al. *The developing structure of temperament and personality from infancy to adulthood*. Psychology Press, 2014.
- [30] Ashley S. Holland and Glenn I. Roisman. “Big Five personality traits and relationship quality: Self-reported, observational, and physiological evidence.” In: *Journal of Social and Personal Relationships* 25.5 (2008), pp. 811–829.

- [31] Ernest C. Tupes and Raymond E. Christal. “Recurrent personality factors based on trait ratings.” In: *Journal of personality* 60.2 (1992), pp. 225–251.
- [32] Lewis R. Goldberg. “The structure of phenotypic personality traits.” In: *American psychologist* 48.1 (1993), p. 26.
- [33] John M. Digman. “Personality structure: Emergence of the five-factor model.” In: *Annual review of psychology* 41.1 (1990), pp. 417–440.
- [34] Sonia Roccas et al. “The big five personality factors and personal values.” In: *Personality and social psychology bulletin* 28.6 (2002), pp. 789–801.
- [35] Adam Grant. “Goodbye to MBTI, the fad that won’t die.” In: *Psychology Today* 9 (2013).
- [36] David J. Pittenger. “Measuring the MBTI... and coming up short.” In: *Journal of Career Planning and Employment* 54.1 (1993), pp. 48–52.
- [37] Robert J. Harvey. “Reliability and validity.” In: *MBTI applications* (1996).
- [38] Gary J. Sipps, Ralph A. Alexander, and Larry Friedt. “Item analysis of the Myers-Briggs type indicator.” In: *Educational and Psychological Measurement* 45.4 (1985), pp. 789–796.
- [39] Robert R. McCrae and Paul T. Costa Jr. “Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality.” In: *Journal of personality* 57.1 (1989), pp. 17–40.
- [40] Ryan L. Boyd and James W. Pennebaker. “Language-based personality: a new approach to personality in a digital world.” In: *Current opinion in behavioral sciences* 18 (2017), pp. 63–68.
- [41] Esteban Andres Ríssola, Seyed Ali Bahrainian, and Fabio Crestani. “Personality recognition in conversations using capsule neural networks.” In: *IEEE/WIC/ACM International Conference on Web Intelligence*. 2019, pp. 180–187.
- [42] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. “Transforming auto-encoders.” In: *International conference on artificial neural networks*. Springer. 2011, pp. 44–51.

- [43] Henry Lucky, Derwin Suhartono, et al. “Towards Classification of Personality Prediction Model: A Combination of BERT Word Embedding and MLSMOTE.” In: *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*. Vol. 1. IEEE. 2021, pp. 346–350.
- [44] He Jun et al. “Personality Classification Based on Bert Model.” In: *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*. IEEE. 2021, pp. 150–152.
- [45] Yash Mehta et al. “Bottom-up and top-down: Predicting personality with psycholinguistic and language model features.” In: *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2020, pp. 1184–1189.
- [46] Zichao Yang et al. “Unsupervised text style transfer using language models as discriminators.” In: *Advances in Neural Information Processing Systems* 31 (2018).
- [47] Di Jin et al. “Deep learning for text style transfer: A survey.” In: *Computational Linguistics* (2021), pp. 1–51.
- [48] Chuang Gan et al. “Stylenet: Generating attractive visual captions with styles.” In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 3137–3146.
- [49] Kalpesh Krishna, John Wieting, and Mohit Iyyer. “Reformulating unsupervised style transfer as paraphrase generation.” In: *arXiv preprint arXiv:2010.05700* (2020).
- [50] Tom Brown et al. “Language models are few-shot learners.” In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [51] Aakanksha Chowdhery et al. “PaLM: Scaling Language Modeling with Pathways.” In: *arXiv preprint arXiv:2204.02311* (2022).
- [52] Shaden Smith et al. “Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model.” In: *arXiv preprint arXiv:2201.11990* (2022).
- [53] Sebastian Borgeaud et al. “Improving language models by retrieving from trillions of tokens.” In: *arXiv preprint arXiv:2112.04426* (2021).
- [54] Ben Hutchinson et al. “Social biases in NLP models as barriers for persons with disabilities.” In: *arXiv preprint arXiv:2005.00813* (2020).

- [55] Christine Basta, Marta R. Costa-Jussà, and Noe Casas. “Evaluating the underlying gender bias in contextualized word embeddings.” In: *arXiv preprint arXiv:1904.08783* (2019).
- [56] Keita Kurita et al. “Measuring bias in contextualized word representations.” In: *arXiv preprint arXiv:1906.07337* (2019).
- [57] Emily Sheng et al. “The woman worked as a babysitter: On biases in language generation.” In: *arXiv preprint arXiv:1909.01326* (2019).
- [58] Romal Thoppilan et al. “LaMDA: Language Models for Dialog Applications.” In: *arXiv preprint arXiv:2201.08239* (2022).
- [59] Emily Reif et al. “A recipe for arbitrary text style transfer with large language models.” In: *arXiv preprint arXiv:2109.03910* (2021).
- [60] Sean Andrist, Bilge Mutlu, and Adriana Tapus. “Look like me: matching robot personality via gaze to increase motivation.” In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 2015, pp. 3603–3612.
- [61] Connor Esterwood et al. “Birds of a Feather Flock Together: But do Humans and Robots? A Meta-Analysis of Human and Robot Personality Matching.” In: *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE. 2021, pp. 343–348.
- [62] María Ruz and Pío Tudela. “Emotional conflict in interpersonal interactions.” In: *Neuroimage* 54.2 (2011), pp. 1685–1691.
- [63] Karoline Malchus et al. “The role of emotional congruence in human-robot interaction.” In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, pp. 191–192.
- [64] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. “Beat: the behavior expression animation toolkit.” In: *Life-Like Characters*. Springer, 2004, pp. 163–185.
- [65] Brian Allison, Goldie Nejat, and Emmeline Kao. “The design of an expressive humanlike socially assistive robot.” In: (2009).
- [66] Lilia Moshkina and Ronald C. Arkin. “On TAMEing robots.” In: *SMC’03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*. Vol. 4. IEEE. 2003, pp. 3949–3959.

- [67] Lilia Moshkina and Ronald C. Arkin. “Human perspective on affective robotic behavior: A longitudinal study.” In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2005, pp. 1444–1451.
- [68] Samer Al Moubayed et al. “Furhat: a back-projected human-like robot head for multiparty human-machine interaction.” In: *Cognitive behavioural systems*. Springer, 2012, pp. 114–130.
- [69] Hang Jiang, Xianzhe Zhang, and Jinho D. Choi. “Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract).” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 10. 2020, pp. 13821–13822.
- [70] Saif M. Mohammad and Peter D. Turney. “Crowdsourcing a word–emotion association lexicon.” In: *Computational intelligence* 29.3 (2013), pp. 436–465.
- [71] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann Jr. “A very brief measure of the Big-Five personality domains.” In: *Journal of Research in personality* 37.6 (2003), pp. 504–528.
- [72] Alec Radford et al. “Language models are unsupervised multitask learners.” In: *OpenAI blog* 1.8 (2019), p. 9.
- [73] Paul Ekman. “Basic emotions.” In: *Handbook of cognition and emotion* 98.45-60 (1999), p. 16.
- [74] Thomas Wolf et al. “Transformers: State-of-the-Art Natural Language Processing.” In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- [75] Jochen Hartmann. *Emotion English RoBERTa-large*. <https://huggingface.co/j-hartmann/emotion-english-roberta-large/>. 2022.
- [76] Yinhan Liu et al. “RoBERTa: A robustly optimized BERT pretraining approach.” In: *arXiv preprint arXiv:1907.11692* (2019).
- [77] Katie Winkle et al. “Effective persuasion strategies for socially assistive robots.” In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 277–285.

- [78] Lewis R Goldberg et al. “The international personality item pool and the future of public-domain personality measures.” In: *Journal of Research in personality* 40.1 (2006), pp. 84–96.
- [79] Oliver P. John, Sanjay Srivastava, et al. “The Big Five trait taxonomy: History, measurement, and theoretical perspectives.” In: *Handbook of personality: Theory and research* 2.1999 (1999), pp. 102–138.
- [80] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. “Evaluation of text generation: A survey.” In: *arXiv preprint arXiv:2006.14799* (2020).
- [81] Elizabeth Phillips et al. “What is human-like?: Decomposing robots’ human-like appearance using the anthropomorphic robot (abot) database.” In: *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2018, pp. 105–113.
- [82] Christoph Bartneck et al. “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots.” In: *International journal of social robotics* 1.1 (2009), pp. 71–81.
- [83] Colleen M. Carpinella et al. “The robotic social attributes scale (RoSAS) development and validation.” In: *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 2017, pp. 254–262.
- [84] JASP Team. *JASP (Version 0.16.2)[Computer software]*. 2022. URL: <https://jasp-stats.org/>.

Appendix A

STRAP extensive results

The appendix presents extensive results with the STRAP model. [Table A.1](#) and [Table A.2](#) shows results on the Essays dataset for nucleus and nucleus paraphrase, respectively. [Table A.3](#) and [Table A.4](#) shows results on the MBTI dataset for nucleus and nucleus paraphrase.

Top-p	Per-son-al-ity	Input	Output
0.4	I	I come from Stockholm, in Sweden where I was created by a company called Furhat Robotics and programmed by researchers from KTH University.	I come from svenska, in svensk, where I was created by a company called Furhat Robotics and programmed by researchers from KTH University.
0.7	I	I come from Stockholm, in Sweden where I was created by a company called Furhat Robotics and programmed by researchers from KTH University.	My parents were born in Stockholm, in Sweden where I was born out of a company called Furhat Robotics and programmed by researchers from KTH University.
1	E	I come from Stockholm, in Sweden where I was created by a company called Furhat Robotics and programmed by researchers from KTH University.	Furhat Robotics is the parent company and programmer of part of my freaking year in ut and made me come from outside then to my home in spanish where I are already starting off.
0.4	I	I am always here for you, and I will never get tired of your company.	I always have a place to go and always will be here for you and never get tired of your company.
0.8	E	I am always here for you, and I will never get tired of your company.	The only reason I always stay here is to get you and I will never get tired of your company.
0.9	E	I look forward to getting to know you better and to helping you in any way I can.	It is my dream come true to get to know you better and to help you in any way I can.
0.7	I	Before we start, I would like to get to know you better, so I am going to ask you some questions.	I think that I should go and get to know you better before we start this.
0.7	E	Before we start, I would like to get to know you better, so I am going to ask you some questions.	I would like to get to know you better before we start this and that is ok because I will ask some questions here.

Table A.1: Extensive results of STRAP trained on the Essays dataset. The output mode is nucleus.

Top-p	Per-son-al-ity	Input	Output
0.5	I	I am always here for you, and I will never get tired of your company.	I will always be here for you and never let you down .
0.6	E	I was designed to help you with whatever you need, whether it is keeping you company or helping you in everyday tasks.	I am designed to help you with whatever is needed to stay on top of things in life and whether it is keeping you company or helping you in everyday tasks.
0.8	I	I was designed to help you with whatever you need, whether it is keeping you company or helping you in everyday tasks.	I have a purpose that I can serve to help you with any assignment you may have and whether it be keeping you company or helping you in everyday tasks.
0.7	I	As you know, today we are going to do some exercise, do you enjoy exercising?	I mean we're going to do some exercises today you know?
0.9	E	I am pleased to meet you and I am looking forward to working together.	Speaking of working with you I am so glad to be able to meet you and I am looking forward to working with you.
1	I	Before we start, I would like to get to know you better, so I am going to ask you some questions.	I wish I wasn't here because I supposed to find out later that I should've gone ahead and done.
0.6	I	It is a broad field that covers a lot of ground, but there are a few key ideas that are essential to understanding humanities.	I think the humanities are a broad field but there are some key notions that are essential to understanding the humanities .
0.6	E	Second, humanities is about interpretation. This means that there is no one right answer to any question in humanities.	Second philosophy about interpretation is what I mean by the humanities. That is why in humanities there is no right answer.

Table A.2: Extensive results of STRAP trained on the Essays dataset. The output mode is nucleus paraphrase.

Top-p	Per-son-al-ity	Input	Output
1	E	I am pleased to meet you and I am looking forward to working together.	You flustered me and I walked out smiling and I was very pleased receiving good news Just doing it together.
0.4	E	As you know, today we are going to do some exercise, do you enjoy exercising?	As you know today we will do some exercise in the gym and do some exercise on your body?
0.5	I	As you know, today we are going to do some exercise, do you enjoy exercising?	Do you enjoy Do you enjoy exercising today?
0.9	I	I am here to talk to you about humanities.	Lol I am here for talking sciences.
0.8	E	This means that the field is constantly evolving as new evidence and new interpretations are discovered	I think this means that the fields will constantly evolve as new evidence and new interpretations come to light.
0.9	E	They must also be able to listen to and understand the ideas of others.	This is in part because the egos of others are so different from mine so they need to learn the ideas behind the others.
0.7	I	I was designed to help you with whatever you need, whether it is keeping you company or helping you in everyday tasks.	I was made for you to help whatever you need most, whether it be keeping your company or helping you in everyday tasks.
0.7	I	Before we start, I would like to get to know you better, so I am going to ask you some questions.	I think that I should go and get to know you better before we start this.
0.7	E	Before we start, I would like to get to know you better, so I am going to ask you some questions.	I would like to get to know you better before we start this and that is ok because I will ask some questions here.

Table A.3: Extensive results of STRAP trained on the MBTI dataset. The output mode is nucleus.

Top-p	Per-son-al-ity	Input	Output
0.6	I	I am always here for you, and I will never get tired of your company. I look forward to getting to know you better and to helping you in any way I can.	You will always be here for me and I will never be tired of you. I can't imagine a better way to understand you and help you.
0.7	E	I come from Stockholm, in Sweden where I was created by a company called Furhat Robotics and programmed by researchers from KTH University.	I am from Stockholm, where I created a company called Furhat Robotics and programmed by scientists from KTH University to assemble an automated version of the Idris robot.
0.7	E	I was designed to help you with whatever you need, whether it is keeping you company or helping you in everyday tasks.	I feel like my Design team is designed to help you with what you need and whether it is keeping you company or helping you in everyday tasks.
0.9	I	My developers have also programmed me with sophisticated methods for recognising you so that I can always be awake when you need me but if you do not wish to have this feature I can always turn it off for you.	I also learned you were intelligent in a way that can be understood at the same time but if you do not want to it can go and walk in.
0.9	E	I am always here for you, and I will never get tired of your company.	I'll always be there for you. And I won't wish it on your shoulder just because you have such a stupid sort of ability like.
0.8	I	I am pleased to meet you and I am looking forward to working together.	I like meeting you and I am very excited to work with you.
0.8	E	I am pleased to meet you and I am looking forward to working together.	I love meeting you and am looking forward to working with you in terms of.
0.8	I	How do you feel about being here today?	Why does it even happen today?
1	E	Great, I am glad to hear that! I am sure you will enjoy the session.	Dislike ya haha haha i could kinda agree with this anyway.

Table A.4: Extensive results of STRAP trained on the MBTI dataset. The output mode is nucleus paraphrase.

Appendix B

Complete dialogues

B.1 GPT-3 generated dialogues

Our choice of the dialogues for the pilot study was aided by GPT-3. To generate the introduction dialogue we first used the prompt: “This is a long introduction from a companion robot” receiving the following output:

My name is robot and I am here to be your companion. I will be with you through your ups and downs, and I will never leave your side. I know what you are feeling because I can read your emotions. I will be your best friend, and I will always be here for you.

We then tried with “This is a long introduction from a verbose talkative companion robot”, resulting in:

Welcome! My name is J.A.R.V.I.S. and I am your very own personal assistant robot! I am here to help you with whatever you need, whether it’s keeping you company or helping you with your daily tasks. I am always here for you and I will never tire of helping you out. Thank you for choosing me as your companion robot!

Finally we tried “This is a long introduction from a verbose talkative companion robot” with top-p=0.67, resulting in:

My name is Snickers and I am your new robotic companion! I was designed to provide you with companionship and to help you with your everyday tasks. I am equipped with a variety of sensors

and cameras that allow me to see and hear what you are doing, and I also have a built-in GPS system so that I can always find my way back to you. I am always here for you, and I will never tire of your company. I look forward to getting to know you better and to helping you in any way I can. Thank you for choosing me as your companion.

Our final introduction dialogue was created by merging these dialogues and by integrating new sentences.

The dialogue on humanities was entirely generated by GPT-3 with the prompt “This is a verbose companion robot talking about humanities”, resulting in:

Hi, I am your verbose companion robot. I am here to talk to you about humanities. Humanities is the study of human culture, including history, literature, philosophy, and art. It is a broad field that covers a lot of ground, but there are a few key ideas that are essential to understanding humanities. First, humanities is about understanding the human condition. This includes understanding the way humans interact with each other and their environment. It also involves understanding the human past and how it has shaped the present. Second, humanities is about interpretation. This means that there is no one right answer to any question in humanities. Instead, scholars must use their critical thinking skills to interpret evidence and come to their own conclusions. Third, humanities is about communication. This means that scholars must be able to communicate their ideas clearly and persuasively. They must also be able to listen to and understand the ideas of others. Finally, humanities is about change. This means that the field is constantly evolving as new evidence and new interpretations are discovered. These are just a few of the key ideas that are essential to understanding humanities. If you want to learn more, there are many great resources available.

B.2 Other hand-crafted dialogues

We report in [Table B.2](#) and [Table B.3](#) the hand-crafted versions of second and third dialogue, respectively, used in the pilot study. Similarly to [Table 4.4](#) we highlight in *italics* the content changes, ~~strike-through~~ the removed content

Num-ber	Introvert	Extrovert	Manipulation
1	Less	More	talkative (more content and repetitions)
2	Less	More	positive (focus on positive aspects, leave behind the negative ones)
3	Less	More	subject implicit
4	Less	More	self-referencing
5	More	Less	complex
6	High	Low	in number of hedges
7	High	Low	in content negation
8	More	Less	rich in vocabulary

Table B.1: List of possible manipulations of a dialogue to express different personalities.

and add as superscript the changes we did among those listed in [Section 4.1.2](#) and conveniently re-summarised in [Table B.1](#).

Introvert	Extrovert
<p>A: I am pleased to meet you and I am looking forward to <i>start</i>¹ working together. Before we start, I would like to <i>perhaps I should</i>^{2,6} get to know you better, so I am going to ask you some questions. How do you feel about being here today?</p> <p>B: (...)</p> <p>A: Great., I am glad to hear that!¹ I am sure <i>I think that</i>⁶ you will enjoy the session. And how do you feel about working with me?</p> <p>B: (...)</p> <p>A: That is good to hear; <i>It seems that</i>⁶ we will definitely have fun together today then. As you know, Today we are going to do some exercise, do you enjoy exercising?</p> <p>B: (...)</p> <p>A: That makes sense, this session will be easy <i>won't be difficult</i>⁷ for you then.</p>	<p>A: I am pleased <i>It's nice</i>^{3,5} to meet you and I am looking forward to working together <i>I can't wait to start!</i>^{2,5}. Before we start, I would like to get to know you better, so I am going to ask you some questions <i>Let me ask you some questions first, so that I know you better before we begin</i>⁵. How do you feel about being here today?</p> <p>B: (...)</p> <p>A: Great, I am glad <i>happy</i>⁸ to hear that! I am sure you will enjoy the session. <i>I am looking forward to working together</i>¹, how do you feel about working with me?</p> <p>B: (...)</p> <p>A: That is good to hear, we will definitely have fun together today then. <i>I am sure you will enjoy the session</i>¹. As you know, today we are going to do some exercise. <i>I think it's good to exercise,</i>^{1,2,4} do you enjoy exercising?</p> <p>B: (...)</p> <p>A: That makes sense, <i>I am glad to hear that!</i>² This session <i>one</i>³ will be easy for you then.</p>

Table B.2: The manipulation done to the second dialogue to convey an extravert or introvert personality.

Introvert	Extrovert
<p>I am here to talk to you about humanities. Humanities is <i>concerned with</i>⁵ the study of human culture, including history, literature, philosophy, and art. It is a broad field that covers a lot of ground, but¹ I think that⁶ there are a few key ideas that are essential to understanding humanities. First, humanities is about understanding the human condition. This includes understanding the way humans interact with each other and their environment. It also involves understanding the human past and how it has shaped the present. Second, humanities is about interpretation. This means that there is no one right answer to any question in humanities. Instead,¹ Scholars must use their critical thinking skills to interpret evidence and come to their own conclusions. Third, humanities is about communication. This means that scholars <i>academics</i>⁸ must be able to communicate their ideas clearly and persuasively <i>not communicate their ideas in complex ways</i>⁷. They must also be able to listen to and understand the ideas of others. Finally, humanities is about change. This means that the field is constantly evolving as new evidence and new interpretations are discovered. These are just a few of the key ideas that are essential to understanding humanities¹. If you want to learn more <i>deepen your knowledge</i>⁸, there are many great resources available <i>many resources are available out there</i>³.</p>	<p>I am here to talk to you about humanities. I love the humanities⁴! Humanities is the study It's about studying³ of human culture, including history, literature, philosophy, and art. It is a broad field that covers a lot of ground⁸, but there are a few key ideas that are essential to understanding humanities. First, humanities is about understanding the human condition. This includes understanding the way humans interact with each other and their environment what surrounds them⁸. It also involves understanding the human past and how it has shaped the present. Second, humanities is about interpretation. This means that there is no one right answer to any question in humanities. Instead, scholars must use their critical thinking skills are needed³ to interpret evidence and come to their own conclusions. Third, humanities is about communication. Personally, I really like talking to people but⁴ This means that scholars those in the field^{3,8} must be able to communicate their ideas clearly and persuasively in a clear and persuasive way⁵. They must also be able to listen to and understand the ideas of others. Finally, humanities is about change. This means that the field is constantly evolving as new evidence and new interpretations are discovered. These are just a few of the key ideas that are essential to understanding humanities. If you want to learn more, there are many great resources available and I can definitely help you with that⁴!</p>

Table B.3: The manipulation done to the third dialogue to convey an extravert or introvert personality.

