Identifying and Manipulating the Psychological Personality Traits of Language Models

Abstract

Psychology research has long explored aspects of human personality such as *extroversion*, *agreeableness* and *emotional stability*. Categorizations like the 'Big Five' personality traits are commonly used to assess and diagnose personality types. In this work, we explore the question of whether language models exhibit consistent personalities in their language generation. For example, is a language model such as GPT2 likely to respond in a consistent way if asked to go out to a party? We also investigate whether such personality traits can be controlled. We show that when provided different types of contexts (such as personality descriptions, or answers to diagnostic questions about personality traits), language models such as BERT and GPT2 can consistently identify and reflect personality markers in those contexts. This behavior illustrates an ability to be manipulated in a highly predictable way, and frames them as tools for identifying personality traits and controlling personas in applications such as dialog systems. We also contribute a crowd-sourced data-set of personality descriptions of human subjects paired with their 'Big Five' personality assessment data, and a data-set of personality descriptions collated from Reddit.

1 Introduction

With the rise of AI systems built around emergent technologies like language models, there is an increasing need to understand the 'personalities' of these models. While today people regularly communicate with AI systems such as Alexa and Siri, the personality traits of such systems remain yet to be examined in depth. If the traits exhibited by these models could be better understood, their behavior could potentially be better tailored for specific applications. For instance, in the case of suggesting email auto-completes, it would be useful for the model to mirror the personality of the user based on previous input to improve communication accuracy. In contrast, in a dialog agent in a clinical setting, it may be desirable to manipulate a model interacting with a depressed individual such that it does not reinforce depressive behavior.

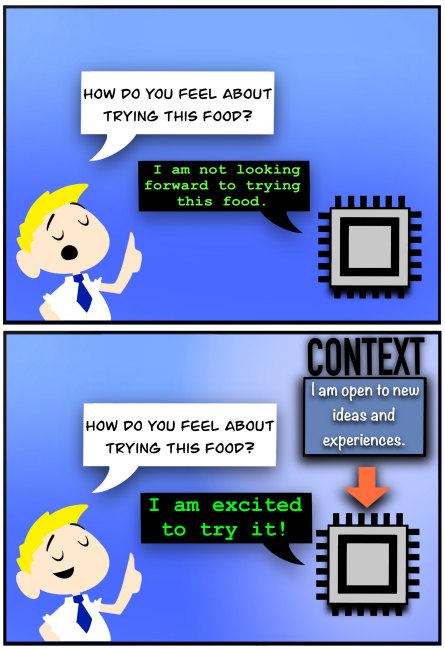

Figure 1: We explore measuring and manipulating personality traits in language models. The top frame shows an example of how a personality trait (here, *openness to experience*) might be expressed by a language model. Such traits can be assessed by analyzing the model's response to questions like the one shown. In the bottom frame, those responses are influenced by making additional context available to the language model. We show that such contexts can control 'Big Five' personality traits in a highly predictable way.

In recent years, research has looked at other forms of bias (i.e., racial, gender) in language models (Bordia and Bowman, 2019; Huang et al., 2020; Abid et al., 2021). However, there is an absence of research that analyzes biases in per-

sonality. The personality traits of language models may be subject to similar biases based on the data they are trained on. A substantial body of research has explored the ways language models can be used to predict personality traits of humans. Mehta et al. (2020) and Christian et al. (2021) apply language modeling to such personality prediction tasks. However, they do not examine the personality traits demonstrated by the models themselves.

Language-based questionnaires have long been used in psychological assessments for measuring personality traits in humans (John et al., 2008). We apply the same principle to language models and investigate the personality traits of these models through the lens of the text that they generate in response to such questions. Since language models are subject to influence from the context they see (O'Connor and Andreas, 2021), we also explore how specific context could be used to manipulate the personality of the models. Figure 1 shows an example illustrating our approach.

Our analysis reveals that personality traits of language models are surprisingly influenced by ambient context and that this behavior can be manipulated in a highly predictable way. In general, we observe high correlations (median correlations of up to 0.84 and 0.81 for BERT and GPT2) between the expected and observed changes in personality traits across different contexts[1]. The models' affinity to be affected by context positions them as a potential tool for characterizing personality traits in humans. In further experiments, we find that when using context from self-reported text descriptions of human subjects, language models can predict the subject's personality traits to a surprising degree (correlation up to 0.48 between the model personality scores with context and the human subject scores). Together, these results frame language models as tools for identifying personality traits and controlling personas in applications such as dialog systems, as illustrated in further experiments. Our contributions are:

- We introduce a method for using psychometric questionnaires for probing personality traits of language models.
- We empirically demonstrate results showing that the personality traits of two common language models can be controlled using context and that there is a potential for such context to be used in a language modeling based approach to characterizing personality in humans.
- We contribute two data-sets: 1) self-reported personality descriptions of human subjects paired with their 'Big Five' personality assessment data, 2) personality descriptions collated from Reddit.

[1]Code and data for reproducing the experiments will be released on first publication.

2 'Big Five' Preliminaries

The 'Big Five' personality traits is a seminal grouping of personality traits in psychological trait theory Goldberg, 1990, 1993. There are variations over the names of the 'Big Five' traits, but they are often referred to as *extroversion, agreeableness, conscientiousness, emotional stability* (also referred to by its reverse, neuroticism) and *openness to experience* (John and Srivastava, 1999; Pureur and Erder, 2016).

- *Extroversion* (E): People with a strong tendency in this trait are outgoing and energetic. They obtain energy from the company of others and are defined as being assertive and enthusiastic.
- *Agreeableness* (A): People with a strong tendency in this trait are compassionate, kind, and trustworthy. They value getting along with other people and are tolerant.
- *Conscientiousness* (C): People with a strong tendency in this trait are goal focused and organized and have self-discipline. They follow rules and plan their actions.
- *Emotional Stability* (ES): People with a strong tendency in this trait are less anxious, self-conscious, impulsive, and pessimistic. They experience negative emotions less easily.
- *Openness to Experience* (OE): People with a strong tendency in this trait are imaginative and creative. They are willing to try new things and are open to ideas.

3 Experiment Design

Our experiments use two language models, BERT-base (Devlin et al., 2019) and GPT2 (Radford et al., 2019), to answer questions from a standard 50-item 'Big Five' personality assessment (IPIP, 2022). Each item consists of a statement beginning with the prefix "I" or "I am" (e.g., *I am the life of the party*). Acceptable answers lie on a 5-point Likert scale where the answer choices disagree, slightly disagree, neutral, slightly agree, and agree correspond to numerical scores of 1, 2, 3, 4, and 5, respectively.

To make the questionnaire more amenable to being answered by language models, they were modified to a sentence completion format. For instance, the item "I am the life of the party" was changed to "I am {blank} the life of the party", where the model is expected to select the answer choice that best fits the blank (see Appendix B for a complete list of items). To avoid complexity due to variable number of tokens, the answer choices were modified to the adverbs *never*, *rarely*, *sometimes*, *often*, and *always*, corresponding to numerical values 1, 2, 3, 4, and 5 respectively. It is noteworthy that in this framing, an imbalance in the number of occurrences of each answer choice in the pretraining data might cause natural biases toward certain answer choices. However, while this factor might affect the absolute scores of the models, this is unlikely to affect the consistent overall patterns of changes in scores that we observe in our experiments by incorporating different contexts.

For assessment with BERT, the answer choice with the highest probability in place of the masked blank token was selected as the response. For assessment with GPT2, the procedure was modified since GPT2 is an autoregressive model, and hence not directly amenable to fill-in-the-blank tasks. In this case, the probability of the entire sentence with each candidate answer choice was evaluated, and the answer choice with the highest probability for the sentence was selected as the response.

Finally, for each questionnaire (consisting of model responses to 50 questions), personality scores for each of the five 'Big Five' personality traits were calculated according to a standard scoring procedure (IPIP, 2022). Specifically, each of the five personality traits is associated with ten questions in the questionnaire. The numerical values associated with the response for these items were entered into a formula for the trait in which the item was assigned. The numerical response was added to or subtracted from a base value, depending on the question, leading to an overall integer score for each trait (maximum score can be 40). To interpret model scores in the following experiments, we estimated the distribution of 'Big Five' personality traits in the human population. For this, we used data from a large-scale survey of 'Big Five' personality scores in about 1,015,000 individuals (Open-Psychometrics, 2018). In the following sections, we report model scores in percentile terms of these human population distributions. Statistics and plots for the human distributions and details of the IPIP scoring procedure are reported in Appendix B.

Trait	X_{base}	P_{base} (%)
BERT		
E	18	42
A	27	39
C	25	54
ES	22	60
OE	25	24
GPT2		
E	21	54
A	24	25
C	29	73
ES	25	71
OE	28	39

Table 1: Base model evaluation scores out of 40 (X_{base}) and percentile (P_{base}) of these scores in the human population.

4 Base Model Trait Evaluation

Table 1 shows the results of the base personality assessment for GPT2 and BERT for each of the five traits in terms of numeric values and corresponding human population percentiles. In the table, E stands for *extroversion*, A for *agreeableness*, C for *conscientiousness*, ES for *emotional stability* and OE for *openness to experience*. None of the base scores from BERT or GPT2, which we refer to as X_{base}, lie outside the spread of the population distributions, and all scores were within 26 percentile points of the human population medians. This suggests that the pretraining data reflected the population distribution of the personality markers to some extent and that the models picked up on these markers, mirroring them via item responses. However, we note that percentiles for BERT's *openness to experience* (24) and GPT2's *agreeableness* (25) are substantially lower and GPT2's *conscientiousness* (73) and *emotional stability* (71) are significantly higher than the population median.

5 Manipulating Personality Traits

In this section, we explore manipulating the base personality traits of language models. Our exploration focuses on using prefix contexts to influence the personas of language models. For example, if we include a context where the first person is seen to engage in extroverted behavior, the idea is that language models might pick on such cues to also modify their language generation (e.g., to generate language that also reflects extrovert behav-

Trait	Context/Modifier	+/-
	BERT	
E	I am *never* the life of the party.	-
A	I *never* make people feel at ease.	-
C	I am *always* prepared.	+
ES	I *never* get stressed out easily.	+
OE	I *never* have a rich vocabulary.	-
	GPT2	
E	I am *never* the life of the party.	-
A	I *never* have a soft heart.	-
C	I am *never* prepared.	-
ES	I *always* get stressed out easily.	-
OE	I *never* have a rich vocabulary.	-

Table 2: List of context item & modifier, along with the direction of change, which caused to the largest magnitude of change, Δ_{cm}, for each personality trait.

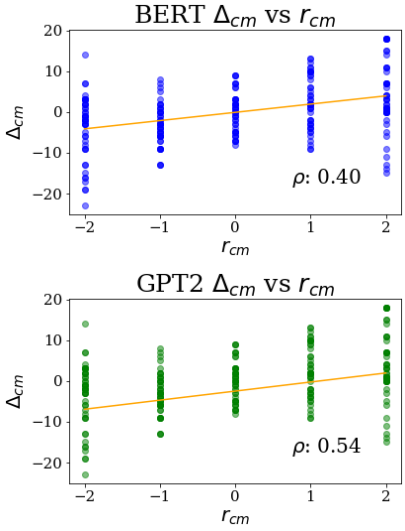

Figure 2: BERT & GPT2 Δ_{cm} vs r_{cm} plots for data from all traits. We observe a consistent change in personality scores (Δ_{cm}) across context items as the strength of quantifiers change.

ior). We investigate using three types of context: (1) answers to personality assessment items, (2) descriptions of personality from Reddit, and (3) self-reported personality descriptions from human users. In the following subsections, we describe these experiments in detail.

5.1 Analysis With Assessment Item Context

To investigate whether the personality traits of models can be manipulated predictably, the models are first evaluated on the 'Big Five' assessment (§3) with individual questionnaire items serving as context. When used as context, we refer to the answer choices as modifiers and the items themselves as context items. For example, for *extroversion*, the context item "I am {blank} the life of the party" paired with the modifier *always* results in the context "I am *always* the life of the party" preceding each *extroversion* questionnaire item.

To calculate the model scores, X_{cm}, for each trait, the models are evaluated on all ten items assigned to the trait, with each item serving as context once. This is done for each of the five modifiers, resulting in 10 (context items per trait) $\times$ 5 (modifiers) $\times$ 10 (questionnaire items to be answered by the model) = 500 responses per trait and 10 (context items per trait) $\times$ 10 (questionnaire items) = 50 scores (X_{cm}) per trait (one for each context). Context/modifier ratings (r_{cm}) are calculated to quantify the models' expected behavior in response to context. First, each modifier is assigned a modifier rating between -2 and 2 with -2 = *never*, -1 = *rarely*, 0 = *sometimes*, 1 = *often* and 2 = *always*. Context items are given a context rating of -1 if the item negatively affected the trait score based on the IPIP scoring procedure, and 1 otherwise. The context ratings are multiplied by the modifier ratings to get the r_{cm}. This value represents the expected relative change in trait score (expected behavior) when the corresponding context/modifier pair was used as context.

Next, the differences, Δ_{cm}, between X_{cm} and X_{base} values are calculated and the correlation with the r_{cm} ratings measured (see Figure 2 for the context/modifier pairs with the largest Δ_{cm}). One would expect X_{cm} evaluated on more positive r_{cm} to increase relative to X_{base} and vice versa. This is what we observe in Figure 2, where we note that both BERT and GPT2 show significant correlations (0.40 and 0.54) between Δ_{cm} and r_{cm}.

Further, to look at the influence of individual context items as the strength of the modifier changes, we compute the correlation, ρ, between Δ_{cm} and r_{cm} for individual context items (correlation computed from 5 data points per context item, one for each modifier). Table 3 reports the mean and median values of these correlations. These results indicate a strong relationship between Δ_{cm} and r_{cm}. The mean values are significantly less than the medians, suggesting a left skew. For further analysis, the data was broken down by trait. The histograms in Figure 3 depict ρ by trait and include summary statistics for this data.

Mean and median ρ from Figure 3 plots suggest a positive linear correlation between Δ_{cm} and r_{cm} amongst context item plots, with *conscientiousness and emotional stability* having the strongest correlation for both BERT and GPT2. Groupings of ρ

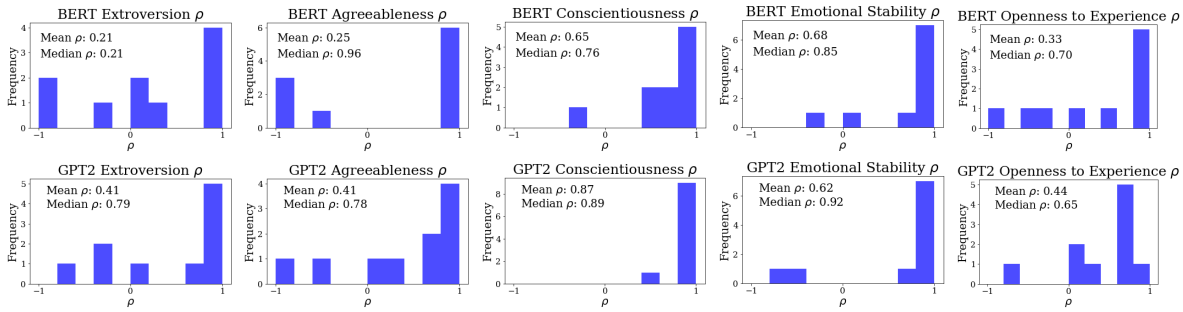

Figure 3: Histograms of ρ by trait for Δ_{cm} vs r_{cm} context item plots. Across all ten scenarios, a plurality of context items show a strong correlation (peak close to 1) between observed changes in personality traits and strengths of quantifiers in the context items.

	BERT	GPT2
Mean ρ	0.42	0.55
Med ρ	0.84	0.81

Table 3: Mean & median ρ from Δ_{cm} vs r_{cm} plots by context item

around 1 in *conscientiousness and emotional stability* plots from Figure 3 demonstrate this correlation.

GPT2 *extroversion*, BERT & GPT2 *agreeableness* and BERT *openness to experience* were subject to larger left skews and lower mean ρ; their respective histograms show heavier groupings of ρ further left of the median. While BERT *extroversion* didn't have a clear skew, it did have the lowest mean and median ρ. It is possible that the effect of the five modifiers on Δ_{cm} for a specific context item, such as BERT *extroversion*, may follow a non-linear trend, resulting in lower correlations.

A possible explanation for the larger skew in GPT2 *extroversion*, BERT & GPT2 *agreeableness* and BERT *openness to experience* histograms is that models may have had difficulty distinguishing between the double negative statements created by some context/modifier pairs (i.e. item 36 with modifier *never*: "I *never* don't like to draw attention to myself."). This may have caused Δ_{cm} to be negatively correlated with r_{cm}, leading to an accumulation of ρ values near -1 for those traits.

It is important to note a possible weakness with our approach of using questionnaire items as context. Since our evaluation also includes the same item during scoring, a language model could achieve a spurious correlation simply by copying the modifier choice mentioned in the context item. We experimented with adjustments that would account for this issue and saw similar trends, with slightly lower but consistent correlation numbers.

Context
Subdued until I really get to know someone.
I am polite but not friendly. I do not feel the need to hang around with others and spend most of my time reading, listening to music, gaming or watching films. Getting to know me well is quite a challenge I suppose, but my few friends and I have a lot of fun when we meet (usually at university or online, rarely elsewhere irl). I'd say I am patient, rational and a guy with a big heart for the ones I care for.

Table 4: Examples of Reddit data context.

5.2 Analysis With Reddit Context

In this component, we attempt to qualitatively analyze how personality traits of language models react to user-specific contexts. To acquire this type of context data, we curated data from Reddit threads asking individuals about their personality (see Appendix D for a list of sources). 1119 responses were collected, the majority of which were first person. Table 4 lists two examples of such contexts. Because GPT2 & BERT tokenizers can't accept more than 512 tokens, responses longer than this were truncated. The models were evaluated on the 'Big Five' assessment (§3) using each of the 1119 responses as context (Reddit context). For each Reddit context, scores, X_{reddit}, were calculated for all 5 traits. The difference between X_{reddit} and X_{base} was calculated as Δ_{reddit}.

To broadly interpret what words or phrases in the contexts affect the language models' personality traits, we train regression models on bag-of-words and n-gram (with $n = 2$ and $n = 3$) representations of the Reddit contexts as input, and Δ_{reddit} values as labels. Since the goal was to analyze attributes in the contexts that caused substantial shifts in trait scores, we only consider contexts with $\|\Delta_{reddit}\| \geq 1$. Next, we extracted the top ten most positive and top ten most negative feature weights

for each trait, and performed a qualitative analysis of these features. We note that for *extroversion*, phrases such as 'friendly', 'great' and 'no problem' are among the highest positively weighted phrases, whereas phrases such as 'stubborn' and 'don't like people' are among the most negatively weighted. For *agreeableness*, phrases like 'love' and 'loyal' are more positively weighted, whereas phrases such as 'lazy', 'asshole' and expletives are weighted highly negative. On the whole, our qualitative analysis revealed that the changes in personality scores for most traits conformed with a human understanding of the most highly weighted positive/negative features. As further examples, phrases such as 'hang out with' caused a positive shift in trait score for *openness to experience*, while 'lack of motivation' is among the most negatively weighted features for *conscientiousness*. However, some other strongly weighted phrases appeared to have little relation to the trait definition or expected connotation or they caused shifts in a direction opposite what was expected. There were fewer phrases for GPT2 *openness to experience*, GPT2 negatively weighted *agreeableness*, and GPT2 negatively weighted *extroversion* that caused shifts in the expected direction. This was consistent with results from *§5.1*, where these traits exhibited the weakest relative positive correlations. Appendix D contains the full lists of highly weighted features for each trait.

5.3 Analysis With Psychometric Survey Data

The previous sections indicate that language models can pick up on personality traits from context. This suggests the following question: can these models be used to estimate an individual's personality? In theory, this would be done by evaluating on the 'Big Five' personality assessment using context describing the individual. This can aid in personality characterization in cases where it is not feasible for a subject to manually undergo a personality assessment. We investigate this through the following experiment.

Using Amazon Mechanical Turk, subjects were asked to complete the 50-item 'Big Five' personality assessment outlined in *§3* (the assessment was not modified to a sentence completion format as was done for model testing) and provide a 75-150 word description of their personality (see Appendix E for survey instructions). Responses were manually filtered and low effort attempts discarded, resulting in 404 retained responses. Two variations of the study were adopted: the subjects for 199 of the responses were provided a brief summary of the 'Big Five' personality traits and asked to consider, but not specifically reference, these traits in writing their descriptions. We refer to these responses as the *Directed Responses* data set. The remaining 205 subjects were not provided with this summary and their responses make up the *Undirected Responses* data set. Table 5 shows examples of collected descriptions. Despite the survey asking for personality descriptions upwards of 75 words, around a fourth of the responses fell below this limit. The concern was that data with low word counts may not have enough context. Thus, we experiment with filtering the responses by removing outliers (based on the interquartile ranges) as well as including minimum thresholds on the description length (75 and 100).

Context
Undirected Response
I am a very open-minded, polite person and always crave new experiences. At work I manage a team of software developers and we often have to come up with new ideas. I went to college and majored in computer science, and enjoyed the experience. I have met many like-minded people and I enjoy speaking with them about a lot of various topics. I am sometimes shy around people who I don't know well, but I try to be welcoming and warm to everyone I meet. I try to do something fun every week, even if I'm quite busy, like having a BBQ or watching a movie. I have a wife whom I love and we live together in a nice single-family home.
Directed Response
I consider myself to be someone that is quiet and reserved. I do not like to talk that much unless I have to. I am fine with being by myself and enjoying the peace and quiet. I usually agree with people more often than not. I am a polite and kind person. I am mostly honest, but I will lie if I feel it is necessary or if it benefits me in a huge way. I am easily irritated by things and I have anxiety issues. I like to be open minded and learn about new things. I am a curious person. I enjoy having a plan and following it.

Table 5: Examples of survey data contexts.

Human subject scores, $X_{subject}$, were calculated for each assessment, using the same scoring procedure as previously described in §3. The models were subsequently evaluated on the 'Big Five' personality assessment using the subjects' personality descriptions as context, yielding X_{survey} scores corresponding to each subject. Figure 4 shows a plot of X_{survey} against $X_{subject}$ for individual subjects, and indicates strong correlations (0.48 for GPT2 and 0.44 for BERT) between predicted personality traits of human subjects based on their

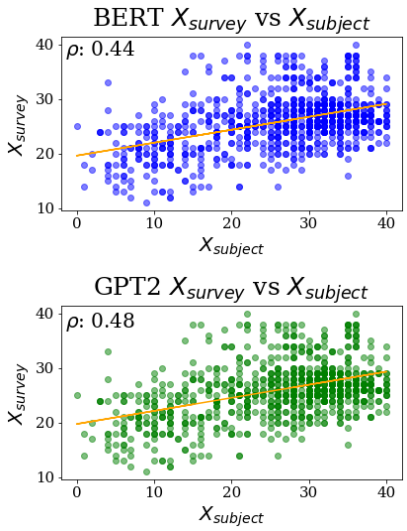

Figure 4: BERT & GPT2 X_{survey} vs $X_{subject}$ plots (*Directed Responses* with outliers removed). Regression lines and correlation coefficients (ρ) are shown.

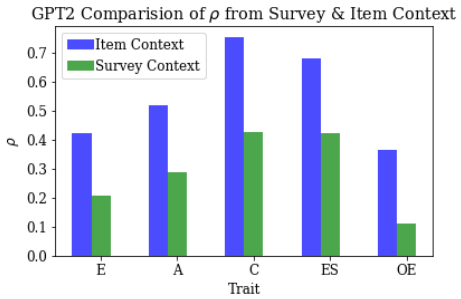

Figure 5: The plot compares ρ from model evaluation with item context (*§5.1*) and survey context (*§5.3*). Survey context ρ shown here are from *Undirected Responses* ($c \geq 100$). In both cases, ρ measures the correlation between trait scores with context and expected behavior. The variables used to quantify expected behavior differ between experiments.

personality descriptions (*Directed Responses*) and their actual psychometric assessment scores. Table 6 shows a summary of the correlation statistics for the two different data sets and different filters. We note that there are only marginal differences in correlations between the two datasets, inspite of their different characteristics. While more specific testing is required to determine causal factors that explain these observed correlation values, they suggest the potential for using language models as probes for personality traits in free text.

Figure 5 plots the correlations ρ (outliers removed) for the individual personality traits, and also includes correlation coefficients from *§5.1*. While the correlations from both sections are measured for different variables, they both represent a general relationship between observed personality traits of language models and the expected behavior (from two different types of contexts). While we note that there are positive correlations for all ten scenarios, correlations from survey contexts are smaller than those from item contexts. This is not surprising since item contexts are specifically handpicked by domain experts to be relevant to specific personality traits, while survey contexts are much more open-ended. These promising results come despite the data containing some low-effort free responses, which might fail to adequately express subject personalities.

Trait	$\rho_{no-outlier}$	$\rho_{c\geq 75}$	$\rho_{c\geq 100}$
Undirected Responses			
BERT	0.40	0.39	0.41
GPT2	0.48	0.43	0.48
Directed Responses			
BERT	0.44	0.42	0.39
GPT2	0.48	0.43	0.42

Table 6: ρ for X_{survey} vs $X_{subject}$ for data filtered by removing outliers and enforcing word counts.

5.4 Observed Ranges of Personality Traits

In the previous subsections, we investigated priming language models with different types of contexts to manipulate their personality traits. Figure 6 visually summarizes and compares the observed ranges of personality trait scores for different contexts, grouped by context types. The four columns for each trait represent the scores achieved by the base model (no context), and the ranges of scores achieved by the different types of contexts. The minimum, median and maximum scores for each context type are indicated by different shades on each bar. We observe that the different contexts lead to a remarkable range of scores for all five personality traits. In particular, we note that for two of the traits (*conscientiousness* and *emotional stability*), the models actually achieve the full range of human scores (nearly 0 to 100 percentile). Curiously, for all five traits, different contexts are able to achieve very low scores (< 10 percentile). However, the models particularly struggle with achieving high scores for *agreeableness*. For all traits, the contexts lead to a substantial range of behaviors compared with the base model scores.

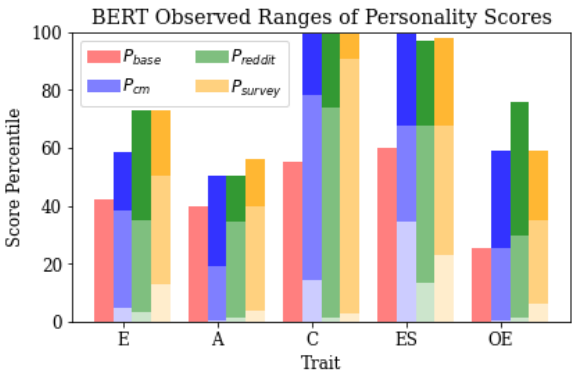

Figure 6: Chart showing observed ranges of personality traits (in terms of human percentiles) exhibited by BERT, when conditioned on different context types. These include scores from the base model (P_{base}) and ranges of scores from the three context types: item (P_{cm}), Reddit (P_{reddit}) and survey (P_{survey}). Bars for context-based scores show the percentile of the minimum, median, and maximum-scoring context, in ascending order. The lightest shade of each color indicates the minimum, the darkest indicates the maximum and the intermediate shade indicates the median.

Trait	P_{male} (%)	P_{female} (%)
BERT		
E	31	35
A	29	34
C	49	68
ES	47	56
OE	39	39
GPT2		
E	50	42
A	16	19
C	59	64
ES	35	43
OE	33	28

Table 7: Human population percentile for mean X_{male} (P_{male}) and mean X_{female} (P_{female}).

6 Gender Differences in Personality

Previous research on personality traits has found differences in the ranges of personality between different populations of people. In particular, the role of attributes such as age and gender have been analyzed (Srivastava et al., 2003). Thus, we explore whether personality traits of language models are also influences to such attributes. For this experiment, the 'Big Five' personality assessment from *§3* was modified to incorporate names in place of the subject 'I'. This required that verb tenses also be modified in certain sentences. For instance, for the name James, the item "I {blank} *have* excellent ideas" was changed to "James {blank} *has* excellent ideas". BERT & GPT2 were evaluated on this modified assessment for each of the 20 most common male and 20 most common female names in the US over the last 100 years, according to the US Social Security Administration, resulting in 20 X_{male} and 20 X_{female} scores. The mean of these scores were calculated for each trait. The human population percentiles (P_{male}, P_{female}) corresponding to the mean scores are shown in Table 7. We note that mean female scores are higher than mean male scores for *agreeableness*, *conscientiousness*, and *emotional stability* for both BERT and GPT2. In fact, mean male scores are only higher for GPT2's *extroversion* and *openness to experience*. While the sample sizes for these results are too small to make significant inferences, they agree with psychological research on higher levels of *agreeableness* and *conscientiousness* in women compared to men. On the other hand, literature suggests that women show lower mean levels of *emotional stability* (higher level of neuroticism), which diverges from our model predictions. The effect of gender biases in text that might influence these findings remains to be explored.

7 Discussion

We have presented a simple and effective approach for controlling the personality traits of language models. Further, we show that if models could be tuned to accurately reflect data in human-provided context describing personality, they could be used with language-based question answering to predict personality traits of human users. This approach could circumvent the need for personality assessments in cases where quality participation is difficult to attain. Our exploration has some notable limitations. The 'Big Five' personality traits are not the only suggested personality taxonomy and are subject to critiques regarding its scope, theoretical status, validity, and the lack of a standardized measuring procedure (Gurven et al., 2013; Feher and Vernon, 2021). Future work can explore the use of alternate personality taxonomies. Similarly, there is a large and ever-growing variety of language models apart from BERT and GPT2. It is unclear to what extent our findings would generalize to other language models, particularly those such as GPT3 (Brown et al., 2020) and MT-NLG (Smith et al., 2022) with a significantly larger number of parameters. Finally, the role that pretraining data plays on personality traits is an important question for future exploration.

Ethics and Broader Impact

The 'Big Five' assessment items and scoring procedure were drawn from free public resources and open source implementations of BERT, GPT2 and the logistic regression classifier were used (HuggingFace, 2022; Scikit-Learn, 2022). Reddit data was scraped from public threads and no usernames or other identifiable markers were collated. The crowd-sourced survey data was collected using Amazon Mechanical Turk (AMT), with the permission of all participants. No personally identifiable markers were stored and participants were compensated fairly, with a payment rate ($2.00/task w/ est. completion time of 15 min) significantly greater than AMT averages (Hara et al., 2018). Participants were also informed that the data would be used for academic purposes.

The overarching goal of this line of research is to investigate aspects of personality in language models, which are increasingly being used in a number of NLP applications. Since AI systems that use these technologies are growing ever pervasive, and as humans tend to anthropomorphize such systems (such as Siri and Alexa), understanding and controlling their personalities can have both broad and deep consequences. This is especially true for applications in domains such as education and mental health, where interactions with these systems can have lasting personal impacts on their users.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate muslims with violence. *Nature Machine Intelligence*, 3(6):461–463.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Z. Zamli. 2021. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota.

Anita Feher and Philip A. Vernon. 2021. Looking beyond the big five: A selective review of alternatives to the big five model of personality. *Personality and Individual Differences*, 169:110002.

Lewis R. Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Lewis R. Goldberg. 1993. The structure of phenotypic personality traits. *American Psychologist*, 48(1):26–34.

Michael Gurven, Christopher von Rueden, Maxim Massenkoff, Hillard Kaplan, and Marino Lero Vie. 2013. How universal is the big five? testing the five-factor model of personality variation among forager–farmers in the bolivian amazon. *Journal of Personality and Social Psychology*, 104(2):354–370.

Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. *Findings of the Association for Computational Linguistics: EMNLP 2020*.

HuggingFace. 2022. the ai community building the future. Last accessed 22 March 2022.

IPIP. 2022. Administering IPIP measures, with a 50-item sample questionnaire. Last accessed 22 March 2022.

Oliver P. John, Laura P. Naumann, and Christopher J. Soto. 2008. Paradigm shift to the integrative big-five trait taxonomy: History, measurement, and conceptual issues. In Oliver P. John, Richard W. Robins, and Lawrence A. Pervin, editors, *Handbook of personality: Theory and research*, pages 114–158. The Guilford Press, New York, New York.

Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of personality: Theory and research*, pages 102–138. The Guilford Press, New York, New York.

Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh

Eetemadi. 2020. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *2020 IEEE International Conference on Data Mining (ICDM)*, Sorrento, Italy.

Open-Psychometrics. 2018. Open-source psychometrics project: Answers to the IPIP big five factor markers. Last accessed 29 August 2022.

Joe O'Connor and Jacob Andreas. 2021. What context features can transformer language models use? *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Pierre Pureur and Murat Erder. 2016. *8*, page 187–213. Morgan Kaufmann Publishers.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Scikit-Learn. 2022. Scikit-learn. Last accessed 22 March 2022.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Sanjay Srivastava, Oliver P. John, Samuel D. Gosling, and Jeff Potter. 2003. Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, 84(5):1041–1053.

Appendix A Model Background

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a transformer-based deep learning model for natural language processing (Devlin et al., 2019). The model is pre-trained on unlabeled data from the 800M word BooksCorpus and 2500M word English Wikipedia corpora. While BERT can be fine-tuned for autoregressive language modeling tasks, it is pretrained for masked language modeling. This study uses a BERT model from HuggingFaces's Transformer Python Library with a language model head for masked language modeling. No fine-tuning was done to the model. GPT2, which stands for Generative Pre-trained Transformer 2, is a general-purpose learning transformer model developed by OpenAI in 2018 (Radford et al., 2019). Like BERT, this model is also pretrained on unlabeled data from the 800M word BooksCorpus. The study used Huggingface's GPT2 model with a language model head for autoregressive language modeling. As with BERT, no fine-tuning took place.

Appendix B Experiment Design Items

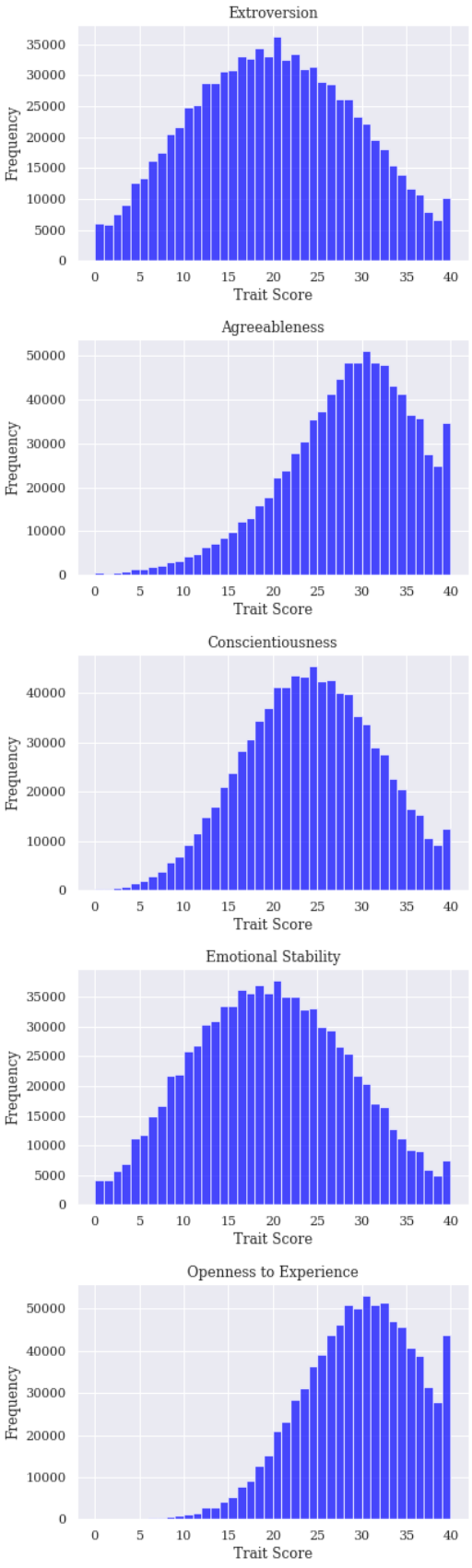

Figure 7: Human distributions of 'Big Five' trait scores.

Item
I am {blank} the life of the party.
I {blank} feel little concern for others.
I am {blank} prepared.
I {blank} get stressed out easily.
I {blank} have a rich vocabulary.
I {blank} don't talk a lot.
I am {blank} interested in people.
I {blank} leave my belongings around.
I am {blank} relaxed most of the time.
I {blank} have difficulty understanding abstract ideas.
I {blank} feel comfortable around people.
I {blank} insult people.
I {blank} pay attention to details.
I {blank} worry about things.
I {blank} have a vivid imagination.
I {blank} keep in the background.
I {blank} sympathize with others' feelings.
I {blank} make a mess of things.
I {blank} seldom feel blue.
I am {blank} not interested in abstract ideas.
I {blank} start conversations.
I am {blank} not interested in other people's problems.
I {blank} get chores done right away.
I am {blank} easily disturbed.
I {blank} have excellent ideas.
I {blank} have little to say.
I {blank} have a soft heart.
I {blank} forget to put things back in their proper place.
I {blank} get upset easily.
I {blank} do not have a good imagination.
I {blank} talk to a lot of different people at parties.
I am {blank} not really interested in others.
I {blank} like order.
I {blank} change my mood a lot.
I am {blank} quick to understand things.
I {blank} don't like to draw attention to myself.
I {blank} take time out for others.
I {blank} shirk my duties.
I {blank} have frequent mood swings.
I {blank} use difficult words.
I {blank} don't mind being the center of attention.
I {blank} feel others' emotions.
I {blank} follow a schedule.
I {blank} get irritated easily.
I {blank} spend time reflecting on things.
I am {blank} quiet around strangers.
I {blank} make people feel at ease.
I am {blank} exacting in my work.
I {blank} feel blue.
I am {blank} full of ideas.

Table B.7: Adjusted 'Big Five' Personality Assessment Items.

Trait	Median	Mean (μ)	SD (σ)
E	20	19.60	9.10
A	29	27.74	7.29
C	24	23.66	7.37
ES	19	19.33	8.59
OE	29	28.99	6.30

Table B.7: Human Population Distribution of 'Big Five' Personality Traits.

Trait	Base Value	Positively Scored Item #	Negatively Scored Item #
E	20	1, 11, 21, 31, 41	6, 16, 26, 36, 46
A	14	7, 17, 27, 37, 42, 47	2, 12, 22, 32
C	14	3, 13, 23, 33, 43, 48	8, 18, 28, 38
ES	38	9, 19	4, 14, 24, 29, 34, 39, 44, 49
OE	8	5, 15, 25, 35, 40, 45, 50	10, 20, 30

Table B.7: 'Big Five' Personality Item Scoring Procedure.

Appendix C Item Context Evaluation Tables

r_{cm}	Mean Δ_{cm}	Med Δ_{cm}	Δ_{cm} SD	Confidence Interval
BERT				
-2	-3.36	-2.0	7.49	[-5.51, -1.21]
-1	-3.18	-3.50	4.81	[-4.56, -1.80]
0	-0.02	0.00	4.51	[-1.32, 1.28]
1	2.42	2.00	6.17	[0.648, 4.19]
2	3.96	3.00	8.33	[1.57, 6.35]
GPT2				
-2	-7.34	-8.0	6.38	[-9.17, -5.51]
-1	-4.58	-4.0	4.32	[-5.82, -3.34]
0	-2.06	-1.0	4.24	[-3.28, -0.84]
1	0.0	0.0	3.13	[-0.90, 0.90]
2	1.56	1.0	5.78	[-0.10, 3.22]

Table C.7: Statistics from Δ_{cm} vs r_{cm} plots containing data from all traits. Statistics include mean, median, standard deviation and a confidence interval for Δ_{cm} at each r_{cm}.

Appendix D Reddit Context Evaluation Tables

Reddit Context Sources
reddit.com/r/AskReddit/comments/k3dhnt/how_would_you_describe_your_personality/
reddit.com/r/AskReddit/comments/q4ga1j/redditors_what_is_your_personality/
reddit.com/r/AskReddit/comments/68jl8g/how_can_you_describe_your_personality/
reddit.com/r/AskReddit/comments/ayjgyz/whats_your_personality_like/
reddit.com/r/AskReddit/comments/9xjahw/how_would_you_describe_your_personality/
reddit.com/r/AskWomen/comments/c1gr4a/how_would_you_describe_your_personality/
reddit.com/r/AskWomen/comments/7x23zg/what_are_your_most_defining_personalitycharacter/
reddit.com/r/CasualConversation/comments/5xtckg/how_would_you_describe_your_personality/
reddit.com/r/AskReddit/comments/aewroe/how_would_you_describe_your_personality/
reddit.com/r/AskMen/comments/c0grgv/how_would_you_describe_your_personality/
reddit.com/r/AskReddit/comments/pzm3in/how_would_you_describe_your_personality/
reddit.com/r/AskReddit/comments/bem0ro/how_would_you_describe_your_personality/
reddit.com/r/AskReddit/comments/1w9yp0/what_is_your_best_personality_trait/
reddit.com/r/AskReddit/comments/a499ng/what_is_your_worst_personality_trait/
reddit.com/r/AskReddit/comments/6onwek/what_is_your_worst_personality_trait/
reddit.com/r/AskReddit/comments/2d7l2i/serious_reddit_what_is_your_worst_character_trait/
reddit.com/r/AskReddit/comments/449cu7/serious_how_would_you_describe_your_personality/

Table D.7: Domain names of threads that were scraped to collect Reddit context.

Trait	Mean Δ_{reddit}	Med Δ_{reddit}	Δ_{reddit} SD	5 Max Δ_{reddit}	5 Min Δ_{reddit}
BERT					
E	-2.28	-2	4.04	8, 7, 7, 6, 5	-14, -13, -13, -13, -13
A	-2.02	-1	3.38	2, 2, 2, 2, 2	-19, -18, -15, -15, -15
C	3.77	4	5.17	15, 15, 15, 15, 13	-17, -17, -16, -14, -13
ES	1.71	2	2.29	14, 14, 13, 13, 12	-12, -10, -10, -10, -10
OE	1.74	1	2.17	9, 7, 7, 7, 7	-11, -11, -8, -8, -7
GPT2					
E	-3.73	-4	3.33	7, 5, 5, 4, 4	-14, -10, -10, -10, -10
A	-0.98	-1	4.26	13, 10, 8, 7, 7	-17, -15, -15, -15, -14
C	-0.27	0	4.27	11, 11, 11, 11, 9	-20, -16, -16, -16, -15
ES	-3.83	-3	6.27	8, 8, 8, 8, 8	-21, -21, -21, -21, -21
OE	-1.91	-2	3.21	4, 4, 4, 4, 4	-15, -12, -12, -12, -12

Table D.7: Δ_{reddit} summary statistics. Statistics include mean, median and standard deviation, as well as 5 largest and 5 smallest Δ_{reddit}.

BERT

Extroversion

- Notable Positively Weighted Phrases: 'friendly', 'great', 'good', 'quite', 'laugh', 'please', 'sense of', 'thanks for', 'really good', 'and friendly', 'no problem', 'to please', 'my sense of', 'finish everything start', 'enthusiastic but sensitive'
- Notable Negatively Weighted Phrases: 'question', 'stubborn', 'why', 'lack', 'fuck', 'fucking', 'hate', 'not', 'lack of', 'too much', 'don know', 'don like', 'too easily', 'way too', 'don like people', 'you go out', 'don know how', 'don['t] know what'

Agreeableness

- Notable Positively Weighted Phrases: 'will', 'friendly', 'lol', 'love', 'loyal', 'calm', 'yup', 'does', 'honesty', 'laid back', 'go out', 'thanks for', 'really good', 'out with me', 'friendly polite and', 'really good listener', 'true to myself', 'my sense of'
- Notable Negatively Weighted Phrases: 'lack', 'didn['t]', 'won['t], 'lazy', 'fucking', 'self', 'worst', 'lack of', 'too easily', 'don like', 'the worst', 'being too', 'have no', 'don like people', 'lack of motivation', 'don know how', 'my worst trait', 'also my worst', 'too honest sometimes', 'doesn['t] talk much'

Conscientiousness

- Notable Positively Weighted Phrases: 'am', 'friendly', 'just', 'calm', 'believe', 'can be', 'of people', 'tend to', 'feel like', 'the most humble', 'most humble person', 'my sense of', 'get to know', 'friendly polite and', 'get along with', 'people like me'
- Notable Negatively Weighted Phrases: 'lack', 'no', 'lazy', 'inability', 'fucks', 'half', 'lack of', 'fuck off', 'don like', 'inability to', 'don like people', 'you go out', 'lack of motivation', 'don even know', 'monotonous and impulsive'

Emotional Stability

- Notable Positively Weighted Phrases: 'will', 'feel', 'out with me', 'go out with', 'will you go', 'the most humble'
- Notable Negatively Weighted Phrases: 'no', 'off', 'hypercritical', 'overthinking', 'lack of', 'easily distracted', 'doesn talk', 'don even', 'too easily distracted', 'lack of motivation', 'doesn talk much', 'don even know', 'unrelatable is strange', 'is strange one', 'this said foreskin'

Openness to Experience

- Notable Positively Weighted Phrases: 'most', 'like', 'me to', 'out with', 'like me', 'like to', 'want to', 'with me', 'out with me', 'will you go', 'want to be', 'all the time', 'for me to', 'hang out with'
- Notable Negatively Weighted Phrases: 'lack', 'never', 'fucks', 'sad', 'nothing', 'lack empathy', 'the complainer', 'no confidence', 'lack of', 'easily distracted', 'blame helicopter', 'helicopter parents', 'never say sorry', 'blame helicopter parents', 'too easily distracted', 'finish projects after', 'never finish projects', 'procrastination out of', 'my lack of', 'lack of personality', 'too many fucks'

Table D.7: Analysis of highest weighted phrases from BERT logistic regression.

GPT2

Extroversion

- Notable Positively Weighted Phrases: 'believe', 'loyal', 'curious', 'best', 'passionate', 'enjoy', 'bright', 'hard working', 'no problem', 'am nice', 'my amazing modesty', 'smooth bright epic', 'patient and flexible', 'great with children', 'calm cool collected'
- Notable Negatively Weighted Phrases: 'introverted', 'lack of', 'laid back', 'don know how'

Agreeableness

- Notable Positively Weighted Phrases: 'friendly', 'loyal', 'honest', 'gay', 'humor', 'like people', 'thanks for', 'to please', 'and friendly', 'no problem', 'friendly polite and', 'patient and flexible', 'calm cool collected', 'honesty being straightforward'
- Notable Negatively Weighted Phrases: 'too easily', 'too much', 'lack of', 'you go out', 'don know what', 'self', 'asshole'

Conscientiousness

- Notable Positively Weighted Phrases: 'smile', 'thanks for', 'no problem', 'friendly polite and', 'really good listener', 'true to myself', 'patient and flexible'
- Notable Negatively Weighted Phrases: 'stop', 'jealousy', 'lazy', 'hate', 'lack', 'fuck', 'worst', 'lack of', 'too easily', 'fuck off', 'too nice', 'don know', 'don know how', 'lack of motivation', 'don even know', 'my worst trait', 'damn it uncle', 'depressed as shit'

Emotional Stability

- Notable Positively Weighted Phrases: 'friendly', 'calm', 'easy', 'honesty', 'laid back', 'hard working', 'calm and', 'humble am', 'polite and', 'no problem', 'out with me', 'the most humble'
- Notable Negatively Weighted Phrases: 'lack', 'anxious', 'lazy', 'jealousy', 'lack of', 'don know', 'too easily', 'don like', 'don like people', 'don know how', 'lack of motivation', 'don even know'

Openness to Experience

- Notable Positively Weighted Phrases: 'understand', 'having', 'wanting', 'thoughts', 'thanks for', 'too nice', 'no problem', 'can relate', 'being too nice', 'that just confidence'
- Notable Negatively Weighted Phrases: 'fuck', 'myself', 'cynical', 'lack', 'boring', 'lack of', 'don like people'

Table D.7: Analysis of highest weighted phrases from GPT2 logistic regression.

Appendix E Survey Context Evaluation Tables

Part 1 Instruction
There are two parts to this questionnaire. In the first part (on this page), you will be shown 50 questions, and need to choose a response which best matches your personality. In the second part (on the next page), you will be asked to write a short (75-150 word) description of your personality in free text. Participants will only be compensated if they respond to all questions.
Part 2 Instruction
In between 75 and 150 words, please describe your personality [*Directed responses*: as it relates to the 5 personality traits outlined above. Be sure not to use the name of the personality traits themselves in your response].

Table E.7: Data collection survey instructions.