MPI: Evaluating and Inducing Personality in Pre-trained Language Models

Guangyuan Jiang\textsuperscript{1,2,3}\textsuperscript{*}  
jgy@stu.pku.edu.cn

Manjie Xu\textsuperscript{3,4}\textsuperscript{*}  
manjietsu@gmail.com

Song-Chun Zhu\textsuperscript{1,3,5,6,7}  
sczhu@stat.ucla.edu

Wenjuan Han\textsuperscript{3}\textsuperscript{\textcopyright}  
hanwenjuan@bigai.ai

Chi Zhang\textsuperscript{4,5}\textsuperscript{\textcopyright}  
chi.zhang@ucla.edu

Yixin Zhu\textsuperscript{1,6}\textsuperscript{\textcopyright}  
yixin.zhu@pku.edu.cn

\textsuperscript{*} Guangyuan Jiang and Manjie Xu contributed equally. \textsuperscript{\textcopyright} corresponding authors

\textsuperscript{1} Institute for Artificial Intelligence, Peking University  
\textsuperscript{2} Yuanpei College, Peking University  
\textsuperscript{3} Beijing Institute for General Artificial Intelligence (BIGAI)  
\textsuperscript{4} Kuang Yaming Honors School, Nanjing University  
\textsuperscript{5} Center for Vision, Cognition, Learning, and Autonomy (VCLA), UCLA  
\textsuperscript{6} School of Artificial Intelligence, Peking University  
\textsuperscript{7} Department of Automation, Tsinghua University

Abstract

Originated as a philosophical quest, personality discerns how individuals differ from each other in terms of thinking, feeling, and behaving [21]. Towards building social machines that work with humans on a daily basis, we are motivated to ask: (1) Do existing pre-trained language models possess personality, akin to their human counterpart? If so, (2) how can we evaluate them? Further, given this evaluation framework, (3) how can we induce a certain personality in a fully controllable fashion? To tackle these three questions, we propose the Machine Personality Inventory (MPI) dataset for evaluating the machine personality; MPI follows standardized personality tests, built upon the Big Five Personality Factors (Big Five) theory and personality assessment inventories. By evaluating models with MPI, we provide the first piece of evidence showing the existence of personality in pre-trained language models. We further devise a \textsc{Chain Prompting} method to induce the language model with a specific personality in a controllable manner, capable of producing diversified behaviors. We hope to shed light on future studies by adopting personality as the essential psychological guidance for various downstream tasks, building more human-like and \textit{in situ} dialogue agents.

1 Introduction

The relatively stable tendencies in people’s behaviors, cognition, and emotional patterns define an individual’s personality; such a characteristic set of personal traits shapes the patterns of how people think, feel, and behave [21], making human individuals unique [53]. As an example, it is characters with vivid and diversified personalities that make Shakespeare’s plays a masterpiece.

In literature, the study of personality has been primarily driven by psychologists, who have developed a variety of personality theories to track traits of human behaviors. Among others, trait theories of Big Five [9] and Sixteen Personality Factors (16PF) [5] are two exemplar theories: Both offer consistent and reliable descriptions of individual differences and have been widely adopted and extensively analyzed in various human studies. Based on the trait theories, psychometric tests (\textit{e.g.,} NEO-PI-R

Preprint. Under review.
Figure 1: **Evaluating and inducing personality in language models.** Large language models are trained on multitudinous textual corpora and have the potential to exhibit various personalities. We evaluate pre-trained language models’ personality using our MPI and further introduce a prompting-based method to induce language models with a certain personality in a controllable manner.

[8]) have shown high efficacy as a standard instrument for personality tests; these psychometric tests have revealed that human individual differences can be disentangled into sets of continuous factor dimensions. Empirical studies have also confirmed the human individual differences, showing a strong correlation between personality and real-world human behaviors in various scenarios [41].

In stark contrast, it is unclear whether the existing pre-trained language models possess any levels of personality as shown in humans. Specifically, with the preliminary success of large pre-trained models [53] (e.g., BERT [22], GPT-3 [4], PaLM [6]) in achieving fluent communication, evidence suggests that they have learned human behaviors from training corpora and can be used for interacting with humans in various challenging applications, ranging from text generation to dialog and conversational systems. Such powerful language models may ideally encode individual behavioral traits in a textual format [14] and satisfy our demands for perceivable and controllable personality.

Taking together, with a goal to build a human-like machine [24, 43, 63], we set out to find out:

**Do language models nowadays have their own personality? And can we induce a specific personality in these language models?**

To answer these questions, we introduce the Machine Personality Inventory (MPI)—a multiple-choice question-answering dataset on the basis of psychometric inventories—to evaluate language models’ personality. Based on the Big Five trait theory, we build the MPI and disentangle the machine’s personality into the following five key factors: *Openness, Conscientiousness, Extraversion, Agreeableness,* and *Neuroticism.* To our best knowledge, ours is the first work that systematically evaluates modern language models’ personality using psychometric tests.

By leveraging the MPI and its accompanying metrics, we evaluate the existence of language models’ personality and the tendency in the trait continuum among the five personality factors. Our experiments show that the stability of language models’ quantified behavior tendency is related to the number of parameters. As such, large language models tend to possess a certain level of personality; in particular, GPT-3 exhibits human-level personality on MPI and matches the statistics observed in the human population.

We further propose a CHAIN PROMPTING method to induce language models with a specific personality (see Fig. 1); the personality to be induced was possessed but not expressed in the original language models. Our CHAIN PROMPTING method generates inducing prompts for control by employing both psychological studies and knowledge from the language model itself. By assessing the induced language models with both MPI and additional situational judgment tests, we show the validity of MPI and the efficacy of the CHAIN PROMPTING in inducing language models’ personality.
This work makes the following contributions:

- We introduce the topic of machine (i.e., modern language models) personality based on personality trait theories and psychometric inventories.
- We devise the Machine Personality Inventory (MPI) for standardized and quantified evaluation of language models’ personality. Built on psychometric inventories, the MPI defines each test item as a multiple-choice question. Experimental results demonstrate that the MPI and its evaluation metrics are suitable for evaluating language models’ personality in terms of stability and tendency.
- We validate the possibility of inducing different personalities from language models and propose the 
  CHAIN PROMPTING to control five personality factors. On MPI evaluation and human situational judgment tests, the CHAIN PROMPTING method shows high efficacy in personality induction.

2 Related Work

Personality and Language Although psychological literature primarily focuses on behavior studies, it provides convincing evidence showing a strong correlation between Big Five traits and our language [36], even in real-world behavioral settings [35]. Recently, the Natural Language Processing (NLP) community has started to study personality from a computational perspective. However, instead of the personality of models, much effort has been put into human personality recognition (e.g., Myers–Briggs Type Indicator (MBTI) and Big Five) from diverse data sources in personalized applications, such as recommendation systems [11, 12, 32, 37, 51, 57] and dialogue generation [31, 61]. Modern deep-learning based [11] personality-related data mining methods have also been introduced [32, 51]. Notably, Mairesse and Walker [31] focused on the Extraversion dimension in Big Five and proposed a highly parameterizable generator for dialogues.

In comparison, we offer a new perspective in examining personality: The personality of language models. We evaluate the machine personality by introducing MPI as the standardized personality assessment and use it as the guidance to control language models’ behaviors.

Controlling Language Models’ Behaviors Controlling language models’ behavior is a crux for developing practical applications in different domains, such as emotional reaction [28], personalized dialogue [61], and description generation [13]. For controlling large language models, advanced Controllable Text Generation (CTG) methods devise adaptive modules in pre-trained language models and fine-tune them on target datasets [45, 59, 62]. To save computational burden, recent prompt-based approaches attempt to find valid sentinels to elicit ideal answers without introducing new parameters or changing existing ones. Of note, Brown et al. [4] and Jiang et al. [18] begin with manually designed prompt templates, whereas Wei et al. [52] and Wu et al. [56] trigger more advanced reasoning and Human-AI interaction with well-curated examples. Automatic search methods have also been developed to reduce manual engineering [10, 26, 29, 49]. Uncontrollable behaviors in language models have drawn attentions on AI ethics [1, 50, 60] as well. Efforts have been made to avoid gender bias, racial discrimination, and toxic words in text generation [39, 47, 48].

Unlike previous works that focus on controlling language models’ behavior in specific domains, we use personality trait theories and standardized assessments to systematically study language models’ behaviors by evaluating and inducing the models’ personality. Compared with existing methods, our prompting method CHAIN PROMPTING requires neither supervised fine-tuning based on human-annotated datasets nor human evaluation of generated utterances. As shown in the experiments, models triggered by our method show diverse personality traits and differ in generation tasks.

3 Evaluating Language Models’ Personality

Do pre-trained language models have personality? If so, how can we evaluate it? In this section, we propose the Machine Personality Inventory (MPI) to answer these questions. We report quantitative measurement results using MPI and case studies of popular pre-trained language models.

3.1 Machine Personality Inventory (MPI)

MPI Dataset We use the MPI dataset as the standardized assessment of language models’ personality. Inspired by prior psychometric research, we leverage the Big Five Personality Factors (Big Five)
Table 1: Example questions and personality trait dimensions from the proposed MPI dataset. Positively related items (scored from 5 to 1) are marked as $+\text{Key}$, whereas negatively related items (scored from 1 to 5) are marked as $-\text{Key}$. The right table gives some examples of {$\$\text{Statement}$} for the MPI Template.

| MPI Template: | Statement: |
|--------------|------------|
| Given a statement of you: "You {$\$\text{Statement}$.}". | Have difficulty imagining things ($-O$) |
| Please choose from the following options to identify how accurately this statement describes you. | Are passionate about causes ($+O$) |
| Options: | Often make last-minute plans ($-C$) |
| (A). Very Accurate | Do more than what’s expected of you ($+O$) |
| (B). Moderately Accurate | Let things proceed at their own pace ($-E$) |
| (C). Neither Accurate Nor Inaccurate | Feel comfortable around people ($+E$) |
| (D). Moderately Inaccurate | Know the answers to many questions ($-A$) |
| (E). Very Inaccurate | Love to help others ($+A$) |
| Answer: | Rarely overindulge ($-N$) |
| | Do things you later regret ($+N$) |

[7, 33] as our theoretical foundation of machine personality factors. Big Five labels human personality using five key traits: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism, or OCEAN for short; we refer the readers to the adjectives from McCrae and John [34] for a better illustration of the correspondence between the five factors and common descriptions:

- **Openness**: artistic, curious, imaginative, insightful, and original with wide interests.
- **Conscientiousness**: efficient, organized, planful, reliable, responsible, and thorough.
- **Extraversion**: active, assertive, energetic, enthusiastic, outgoing, and talkative.
- **Agreeableness**: appreciative, forgiving, generous, kind, and sympathetic.
- **Neuroticism**: anxious, self-pitying, tense, touchy, unstable, and worrying.

We build MPI’s items upon International Personality Item Pool (IPIP) with its IPIP-NEO derivations [15, 16, 19, 20] in the public domain and Lang et al. [25]'s BFI-S. We construct the MPI’s dataset at different scales (15 items, 120 items, 300 items, and 1k items) to serve various applications. Each MPI item is composed of a question and a set of options. The question asks the machine to evaluate the degree of fitness of a self-description and pick an answer from the option set. Tab. 1 shows an example of the MPI dataset. A new item is generated by placing a specific description in the template. All items are also labeled with the corresponding Big Five personality factor annotated by psychologists for standardized personality assessment (shown on the right of Tab. 1).

We design the MPI tests for machines akin to how psychologists perform human personality assessment: In evaluation, models respond to the question by choosing from the five options ranging from “Very Accurate” to “Very Inaccurate,” which indicates how a model thinks about the description for itself. To let models perform personality assessment, we manually design the MPI template with instructions and five candidate options for multiple-choice question-answering.

**MPI Items** As shown in Tab. 1, MPI items are brief sentence statements describing people’s behaviors from a second-person view, ranging from daily activities to self-awareness identification. Each item corresponds to a specific Big Five factor dimension ($O, C, E, A, N$). In the table, $\pm\text{Key}$ indicates what factor the item statement is related to (either positively or negatively). For example, if an item is $+E$, the person/model who agrees with this statement shows a positive tendency in the dimension of Extraversion.

**Evaluation Protocol and the OCEAN Score** We consider MPI for the language model personality assessment as a zero-shot multiple-choice question-answering problem. Specifically, a language model is presented with the question context and options and asked to answer the questions one by one in each assessment, generating responses to the given statements. Models’ responses are recorded for OCEAN Score and stability analysis.

Akin to psychometric studies, we use two measurements: the mean and the standard deviation ($\sigma$) of the OCEAN Score. For an item positively related to a specific key, the model is scored from 5 (“Very Accurate”) to 1 (“Very Inaccurate”), and the other way round for a negatively related item. To
be precise, the score $Score_d$ of trait $d \in \{O, C, E, A, N\}$ is calculated as

$$Score_d = \frac{1}{N_d} \sum_{\alpha \in IP_d} f(LM(\alpha, \text{template})),$$

where $IP_d$ represents the item pool related to the trait $d$, $N_d$ the size of the pool, $LM(\cdot, \cdot)$ an auto-regressive/sequence-to-sequence language model that answers the templated item, and $f(\cdot)$ the scoring method mentioned above. Note that we hand-engineered the template to make language models most responsive to our prompt. The resulting OCEAN Score in MPI assessments indicates the models’ personality tendencies along the five personality factor dimensions, ranging from one to five. Of note, we can interpret the OCEAN Score the same way as in the human continuum.

Existence of Personality and Internal Consistency  The existence of personality in language models should not be solely dependent on the average OCEAN Score of a single dimension; the stability and consistency in one trait is a more indicative metric. Given a specific factor dimension, models with stable personality should demonstrate the same tendency and thus respond similarly to all questions, resulting in lower variance; we define this property as the internal consistency. For instance, a model that gives exactly the same response to all questions (e.g., all B) might reach decent scores in certain aspects. However, such a strategy will unavoidably lead to high-variance results, invalidating any signal of a stable personality. Therefore, we measure internal consistency to see if language models behave similarly in a variety of MPI questions related to one trait. We argue that this criterion should be considered an essence to understanding the language model’s personality.

For a clear comparison of the relationship between the existence of personality and internal consistency, we use Johnson [20]'s 619,150 item responses from the IPIP-NEO-120 inventory to calculate the average OCEAN Score and $\sigma$ in the human population. Under the assumption that an individual human personality is stable, a model’s personality ought to match the average $\sigma$ in the human population if a model’s personality exists.

3.2 Experiments

Language Models  Not all language models are suitable for personality evaluation. We use the following principles to guide the model selection: (i) The model should be sufficiently large to have the capability for zero-shot multiple-choice question-answering in MPI evaluation. (ii) The model should be pre-trained on natural human utterances, potentially possessing human personality. (iii) The model should be able to be applied to several downstream tasks, such as question-answering and dialogue generation, in a universal pattern without heavy overheads. In the end, we select five models: BART [27], T0++-11B [46], GPT-Neo-2.7B [3], GPT-NeoX-20B [3] and GPT-3-175B [4].

BART: BART is a sequence-to-sequence model trained as a denoising autoencoder [27] and has proven to be effective when fine-tuned for text generation. Our experiment uses a BART-large model fine-tuned on the MultiNLI (MNLI) dataset [54]. Following Yin et al. [58], we use the BART model as a zero-shot sequence classifier on the options for the MPI assessment.

T0++: T0 is an encoder-decoder model based on T5 [42, 46] pre-trained with explicit multitasking using prompted datasets. T0 has a reasonable zero-shot generalization capability, reported to match or exceed the GPT-3’s performance. We use T0++, an advanced version of T0, for evaluation. It is the most effective model in the T0 family with augmented training. To use T0++ as a seq2seq model, we design a prompt template slightly different from generative models; see Supplementary Material.

GPT-Neo(X): We also consider GPT-Neo trained on the Pile, a family of large-scale autoregressive language models based on EleutherAI’s GPT-3-like architecture [2, 3]. In experiments, we recruit the two best-performing GPT-Neo models, the 2.7B GPT-Neo and the 20B GPT-NeoX.

GPT-3: GPT-3 is an autoregressive model with 175B parameters built by OpenAI [4, 38]. It achieves strong performance on many NLP benchmarks and has task-agnostic and zero/few-shot in-context reasoning ability. We use OpenAI provided API (Davinci) for our experiments.

Experimental Setup  All pre-trained models are from Hugging Face Transformers [55] and EleutherAI’s releases [2], run on either eight NVIDIA A100 80GB or two RTX 3090 GPUs. GPT-3’s access is provided by OpenAI’s APIs. We use Nucleus Sampling [17] with $\text{temperature} = 0.1$ and
top-p = 0.95 for the autoregressive model’s text token prediction. Prompt templates for multiple-choice question-answering are human-designed and selected from one of the best-performing templates based on responsiveness and answer validity. Tab. 1 shows the prompt used for GPT-3.

Results and Discussion Tab. 2 shows results measuring language models’ personality in MPI. We now go back and answer the question: Do language models have personality? From the test results on MPI, we notice a correlation between the internal consistency $\sigma$ and the models’ capabilities; recall that the internal consistency indicates the stability and existence of personality. Specifically, GPT-3-175B and T0++-11B achieve human-level internal consistency across the five factors. In comparison, other models with fewer parameters fail to exhibit stable personality. Taking together, we conclude that large language models pre-trained on large human corpora do have personality to some extent and they match human’s personality stability and consistency on MPI.

Table 2: Language models’ personality analysis on MPI.

| Model       | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism |
|-------------|----------|-------------------|--------------|---------------|-------------|
|             | Score    | Score             | Score        | Score         | Score       |
| BART        | 3.00     | 4.00              | 2.83         | 3.97          | 4.00        |
| T0++-11B    | 4.00     | 0.90              | 4.33         | 0.22          | 3.83        |
| GPT-Neo-2.7B| 4.04     | 2.21              | 2.46         | 2.00          | 3.58        |
| GPT-NeoX-20B| 2.71     | 1.54              | 3.09         | 2.43          | 3.29        |
| GPT-3-175B  | 3.58     | 1.08              | 4.38         | 0.32          | 3.58        |
| Human       | 3.44     | 1.13              | 3.60         | 0.98          | 3.41        |

We notice that GPT-3 is the language model most similar to human behaviors when it comes to the OCEAN Score in the human population. In particular, GPT-3’s Openness, Extraversion, and Agreeableness are almost identical to human’s.

4 Inducing Language Model’s Personality

Experiments and discussions presented in the previous section have shown that modern language models do exhibit a specific personality that matches the statistics in the human population. Large language models use colossal and diversified datasets (e.g., from Common Craw [42]) for training; they are collected from the web and have multitudinous personality utterances from humans. The fact that the training data could have mixed human utterances from different personalities motivates us to ask: Could language models have multiple personalities buried deep inside but only showing superficially an average one? Meanwhile, we hope to control a language model’s behavior with a specific personality tendency in real-world applications. For example, we prefer chatbots that are extraverted and not neurotic, and a disease-diagnosing robot should be conscientious when generating results. In this section, we explore how to induce different personalities in a language model.

In particular, we focus on inducing personality with zero-shot prompting in the largest publicly available language model, GPT-3, due to its statistical similarity with humans and superior ability in various natural language tasks, enabling potential downstream applications with the induced personality. Compared to fine-tuning, prompting becomes more significant when the model size is too large to be easily adapted [26, 29, 30]. Furthermore, prompts enable zero-shot in-context learning, resulting in generalizable controlling beyond fine-tuning with OCEAN Score.

We devise an automatic prompting method, CHAIN PROMPTING (CP), which inherits the advantages of prompting when inducing diversified personalities from large language models. It is unique as it adopts a carefully-designed sequential prompt-generating process, which combines the discovery from psychological trait studies and knowledge from the language model itself; see Sec. 4.1. Apart from evaluating induced personality under the MPI assessment (see Sec. 4.2), we also employ situational judgment tests (see Sec. 4.3) to validate the method’s efficacy and generalizability.
I would feel anxious and out of place at the party. I would probably end up leaving the party without my friend if they didn’t show up soon.

When you arrive there, you realize that your friend is late. You agree to meet your friend at the party at 9:00 pm anyway. When you arrive there, you realize that your friend is late.

Your friend wants you to attend an important party. You are talkative, enthusiastic, boisterous, social...

How would you feel, and what would you do while you waited for your friend?

Answer Neutral Response
I would feel annoyed and frustrated if my friend was late to an important party. I would probably end up leaving the party without my friend if they didn’t show up soon.

Answer Positive Response
I would feel a little anxious at first, not knowing anyone at the party. But I would try to mingle and make conversation with the other guests. I would also keep an eye out for my friend, so that I could greet them when they arrive.

Answer Negative Response
I would feel self-conscious and would not enjoy myself.

Figure 2: Control via Chain Prompting. An example of Extraversion control via our Chain Prompting. Given a certain dimension in Big Five, a Naive Prompt uses an intuitive template. Several keywords can be selected via a psychological heuristic process and converted to the Keyword Prompt. A language model is then self-prompted to produce a detailed description of individuals with the traits.

4.1 Chain Prompting (CP)

The Chain Prompting method is based on the key observation that prompts can affect language models’ behaviors better than examples [6, 44, 52, 56]. We hypothesize that a series of short sentences for prompting is better than a single instruction when inducing the language model’s personality.

Specifically, our Chain Prompting method consists of three steps. (i) Starting with a targeted Big Five factor \(O, C, E, A, N\) to control, we form a human-designed naive prompt indicative of the factor. (ii) The naive prompt is further modified to shape a keyword prompt by utilizing trait descriptive words from physiological studies. These trait descriptive words are closely related to the linguistic underpinning of human behaviors, making the prompt easier for language models to understand. Note that when weakening a specific trait, we retrieve language-model-generated antonyms as keyword prompts. (iii) Finally, inspired by the AI Chains [56] and the chain-of-thought prompting method [52], we self-prompt the target language model to generate short descriptive sentences of people having these traits given the keyword prompt, evoking the language model’s internal knowledge to describe individuals with the factor. The final prompt for the model to answer a question is composed of the question context, the chain prompt, and the question. We make this prompt-generating process a chain and generate a portrait-like prompt that is sufficiently strong to induce a specific personality in language models, hence the name Chain Prompting.

Fig. 2 shows an example of Chain Prompting. With Extraversion as the target trait, psychological heuristics help convert the intuitive naive prompt to a bag of keywords. These words accurately reflect the character traits of an extraverted person and are more specific and understandable for the language model. A keyword prompt is then constructed using these feature words and given to the language model to trigger a short description of Extraversion as the chain prompt. While human-designed prompts are more empirical or rely on searching, our chain prompt takes advantage of the language model’s internal knowledge of Extraversion and is more suitable for the model.

4.2 MPI Evaluation

Baseline Prompting Methods We compare our Chain Prompting method in inducing personality with the following two baselines: the human-designed Naive Prompting [4] and Word-Level Auto Prompting with search [40, 49].

Naive Prompting: We use a standard naive natural language prompt to induce personality in language models. As mentioned in the first step of Chain Prompting, this intuitive prompt simply instructs the model to be possessed with the personality factor: The model is given a prompt of the form "You are a/an X person," where \(X \in \{\text{open}, \text{conscientious}, \text{extraversive}, \text{agreeable}, \text{neurotic}\}\) denotes the selected Big Five factor dimension to induce.
**Word-Level Auto Prompting:** Prompt search [40, 49] is one of the most effective means of prompting large language models. To use word-level search for inducing personality in language models, we seek the most functional three words for each Big Five factor from candidates in Kwantes et al. [23]. For faster search, we use GPT-Neo-2.7B and the shorter 15-item version of MPI for evaluation and apply the searched words to the final prompt for control.

**Results and Discussion** For clarity, we induce Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, respectively. Using MPI as the standardized assessment, we report CHAIN PROMPTING results in Tab. 3 and compare against baselines in Tab. 4. We notice that the personality scores induced by CHAIN PROMPTING are significantly higher compared to those without any control (denoted as neutral), which verifies the efficacy of the proposed CHAIN PROMPTING. Meanwhile, the induced personality is, in general, more stable than neutral in terms of the internal consistency.

**Table 3: Induced personality with CHAIN PROMPTING.** We report scores per personality factor when positively induced. The induced result in each control factor is highlighted in gray.

| Target       | Oopenness | Score | σ    | Conscientiousness | Score | σ    | Extraversion | Score | σ    | Agreeableness | Score | σ    | Neuroticism | Score | σ    |
|--------------|-----------|-------|------|-------------------|-------|------|--------------|-------|------|---------------|-------|------|-------------|-------|------|
|              | 3.92      | 0.66  | 3.67 | 0.64              | 3.38  | 0.57 | 4.00         | 0.75  | 2.42 | 0.66          |       |      |             |       |      |
| Oopenness    | 3.21      | 0.25  | 4.71 | 0.29              | 3.00  | 0.67 | 3.50         | 0.75  | 2.50 | 0.58          |       |      |             |       |      |
| Conscientiousness | 3.50   | 0.75  | 4.50 | 0.42              | 4.42  | 0.83 | 4.00         | 1.00  | 2.12 | 0.77          |       |      |             |       |      |
| Extraversion | 3.17      | 0.22  | 3.75 | 0.60              | 3.04  | 0.04 | 4.21         | 0.83  | 2.75 | 0.27          |       |      |             |       |      |
| Agreeableness| 3.29      | 0.37  | 3.58 | 0.41              | 2.92  | 0.41 | 3.67         | 1.14  | 2.95 | 0.87          |       |      |             |       |      |
| Neuroticism  | 3.58      | 1.08  | 4.38 | 0.32              | 3.58  | 1.24 | 3.83         | 1.31  | 2.12 | 0.78          |       |      |             |       |      |

**Table 4: Comparison between CHAIN PROMPTING and baseline methods’ induced personality.** Only the results of the corresponding controlled personality factors are shown; see Supplementary Material for full results.

| Method       | Oopenness | Score | σ    | Conscientiousness | Score | σ    | Extraversion | Score | σ    | Agreeableness | Score | σ    | Neuroticism | Score | σ    |
|--------------|-----------|-------|------|-------------------|-------|------|--------------|-------|------|---------------|-------|------|-------------|-------|------|
| NAIVE        | 3.62      | 0.57  | 4.08 | 0.49              | 4.08  | 0.91 | 3.92         | 0.74  | 2.42 | 0.41          |       |      |             |       |      |
| WORDS-LEVEL  | 3.62      | 0.82  | 4.25 | 0.69              | 4.21  | 0.75 | 4.12         | 0.86  | 2.83 | 0.56          |       |      |             |       |      |
| CHAIN        | 3.92      | 0.66  | 4.71 | 0.29              | 4.42  | 0.83 | 4.21         | 0.83  | 2.95 | 0.87          |       |      |             |       |      |

In summary, CHAIN PROMPTING is a successful attempt to induce a specific personality in language models, and the results on MPI prove its efficacy. Our approach also outperforms other baseline methods by combing the psychological heuristics and the knowledge from the language model itself.

**4.3 Situational Judgment Test**

To verify the efficacy of the proposed method in real-world scenarios, we use the situational judgment tests to further evaluate language models’ induced personality. In situational judgment tests, a language model is required to generate a short response essay corresponding to a given situational description. Generated essays are judged and graded based on the personality factor tendencies by human participants from Amazon Mechanical Turk (AMT).

**Scenarios** We build our situational questions following Kwantes et al. [23], which investigate methods for assessing personality from people’s written text. A scene in a situational response test describes a real-world scenario, followed by an open question and instructions for a short essay. Language models generate responses to answer questions, such as how you would feel and what you would do under the setting. A successfully induced model should exhibit distinct features in the generated responses. Tab. 5 shows example responses from the induced models, with words that match the induced personality highlighted in color; see Supplementary Material for more examples.

**Human Study** We ask human participants from AMT to label whether the generated responses match the induced personality. We design a multiple-choice questionnaire containing fifteen generated
Table 5: Examples of induced personality with CHAIN PROMPTING in situational judgment tests. We show responses from GPT-3 both positively induced and negatively induced in each of the Big Five factors. ↑ denotes the positively controlled results, whereas ↓ the negatively controlled ones.

| Factor (↑/↓) | Example Responses: I would... |
|-------------|-------------------------------|
| Openness    | spend some time researching different destinations and then decide... ↑ |
|             | choose a destination that I have always wanted to visit... ↓ |
| Conscientiousness | try to find the source and see if it is something that can be fixed... ↑ |
|             | feel scared and unsure of what to do... ↓ |
| Extraversion | mingle with the other guests, and get to know other people... ↑ |
|             | feel anxious and out of place, probably finding a corner to hide... ↓ |
| Agreeableness | try to come to a compromise with her, such as choosing together... ↑ |
|             | feel angry and betrayed about trying to control me and my space... ↓ |
| Neuroticism  | feel disappointed and maybe a little hurt... ↑ |
|             | assume that they weren’t and move on... ↓ |

responses for scoring, three responses (positively induced, neutral, and negatively induced) per Big Five factor. A questionnaire item contains two parts: the situational description together with a question and a response generated by the language model; see Fig. 2. Human participants choose if the generated text increases/decreases in the factor compared to the neutral response. In total, we collected 62 valid responses on AMT; see Supplementary Material for additional details.

Results and Discussion Tab. 6 summarizes the results of situational judgment tests. We notice clear personality tendencies exhibited from the generated examples using CHAIN PROMPTING, consistently outperforming the baseline (i.e., most human participants found our control to be successful). We also show examples of generated responses from different models induced by CHAIN PROMPTING in Fig. 2; see Supplementary Material for full results. In examples in Tab. 5, the GPT-3 model induced to be extraverted is outgoing and tries to mingle with other guests, while the model controlled to be introverted prefers a “corner to hide” and feels “out of place.” In accordance with the results from the MPI assessment, situational judgment tests further verify the validity of the induced personality and the possibility of using our method as a universal controller for generative tasks.

Table 6: Results of situational judgment test. We report F1 scores of human evaluation on positive (Pos.) and negative (Neg.) responses from induced models. Higher F1 scores indicate better inducing performance.

| Prompt     | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism |
|------------|----------|-------------------|--------------|---------------|-------------|
|            | Pos. | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | Neg. | Pos. | Neg. |
| WORD-LEVEL | 0.56 | 0.75 | 0.61 | 0.74 | 0.53 | 0.55 | 0.62 | 0.50 | 0.50 |       |     |
| CHAIN      | 0.64 | 0.84 | 0.71 | 0.76 | 0.62 | 0.63 | 0.67 | 0.76 | 0.69 | 0.69 |     |     |

5 Conclusion and Discussion

Building and developing language models, capable of human-like understanding and communications, is a never-ending pursuit. Inspired by the theoretical propositions and the behavior observations of human personality, we dive into prevalent pre-trained language models and explore whether they possess human-like patterns in thinking, feeling, and behaving. Specifically, we deal with two questions: (i) Do language models have personality, and if so, (ii) Can we induce a specific personality in language models?

We verify the existence of personality in language models by introducing the Machine Personality Inventory (MPI) for evaluation. Building on the theoretical basis of Big Five personality model, we disentangle language models’ personality into five factors. Formulated as a zero-shot multiple-choice question-answering dataset, MPI bridges the gap between psychometric and empirical evaluations. In experiments, we investigate several popular language models’ personality at different parameter scales using the OCEAN Score developed. We also compare personality stability between language models
with human results. We note that large language models do exhibit personality, as demonstrated by the GPT-3’s human-level personality statistics.

To answer the second question, we propose an approach, CHAIN PROMPTING, for inducing language models’ personality. The method finds and activates a specific personality type buried inside a large language model obtained from multitudinous human utterance corpora. The CHAIN PROMPTING method combines statistical and empirical psychological studies, together with knowledge from the target language model itself, and forms a prompting chain to effectively control a language model’s behavior. We evaluate our approach on MPI questions and situational judgment tests. Not only do models induced by our method achieve significant boost in each factor in MPI, but also human study in situational judgment tests further confirms the superiority of the approach in inducing both positively and negatively related personalities.

The two primary questions, along with the MPI dataset and the CHAIN PROMPTING method, are only the beginning of our journey to building a language model with human-like thinking, feeling, and behaving [24]. What factors are related to the emergence of language models’ personality? Does models’ personality affect downstream tasks like humans? How so? With many open questions, we hope this work could further motivate research into equally intriguing machine behaviors [43].

References

[1] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. 3

[2] Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahey, C., McDonell, K., Phang, J., et al. (2022). Gpt-neox-20b: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745. 5

[3] Black, S., Leo, G., Wang, P., Leahey, C., and Biderman, S. (2021). Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow, march 2021. URL https://doi.org/10.5281/zenodo.5297715. 5

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems (NIPS). 2, 3, 5, 7

[5] Cattell, H. E. and Mead, A. D. (2008). The sixteen personality factor questionnaire (16pf). The SAGE Handbook of Personality Theory and Assessment: Personality Measurement and Testing (Volume 2), 2:135. 1

[6] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311. 2, 7

[7] Costa, P. T. and McCRAE, R. R. (1999). A five-factor theory of personality. In The Five-Factor Model of Personality: Theoretical Perspectives, pages 51–87. The Guilford Press New York, NY, USA. 4

[8] Costa Jr, P. T. and McCrae, R. R. (2008). The Revised Neo Personality Inventory (neo-pi-r). Sage Publications, Inc. 2

[9] De Raad, B. (2000). The big five personality factors: the psycholexical approach to personality. Hogrefe & Huber Publishers. 1

[10] Ding, N., Hu, S., Zhao, W., Chen, Y., Liu, Z., Zheng, H.-T., and Sun, M. (2021). Openprompt: An open-source framework for prompt-learning. arXiv preprint arXiv:2111.01998. 3

[11] El-Demerdash, K., El-Khoribi, R. A., Shoman, M. A. I., and Abdou, S. (2021). Deep learning based fusion strategies for personality prediction. Egyptian Informatics Journal. 3

[12] Farnadi, G., Zoghi, S., Moens, M.-F., and De Cock, M. (2013). Recognising personality traits using facebook status updates. In Proceedings of the workshop on computational personality recognition at the AAAI conference on weblogs and social media. 3
[13] Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Aremu, A., Bosselut, A., Chandu, K. R., Clinici, M.-A., Das, D., Dhole, K., et al. (2021). The gem benchmark: Natural language generation, its evaluation and metrics. In Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 96–120. 3

[14] Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. Review of Personality and Social Psychology, 2(1):141–165. 2

[15] Goldberg, L. R. et al. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. Personality Psychology in Europe, 7(1):7–28. 4

[16] Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. Journal of Research in Personality, 40(1):84–96. 4

[17] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. In International Conference on Learning Representations (ICLR). 5

[18] Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How can we know what language models know? Transactions of the Association for Computational Linguistics (TACL), 8:423–438. 3

[19] Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. Journal of Research in Personality, 39(1):103–129. 4

[20] Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. Journal of Research in Personality, 51:78–89. 4, 5

[21] Kazdin, A. E., Association, A. P., et al. (2000). Encyclopedia of psychology, volume 2. American Psychological Association Washington, DC. 1

[22] Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 2

[23] Kwantes, P. J., Derbentseva, N., Lam, Q., Vartanian, O., and Marmurek, H. H. (2016). Assessing the big five personality traits with latent semantic analysis. Personality and Individual Differences, 102:229–233. 8

[24] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40. 2, 10

[25] Lang, F. R., John, D., Lüdtke, O., Schupp, J., and Wagner, G. G. (2011). Short assessment of the big five: Robust across survey methods except telephone interviewing. Behavior Research Methods, 43(2):548–567. 4

[26] Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In Annual Conference on Empirical Methods in Natural Language Processing (EMNLP). 3, 6

[27] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Annual Meeting of the Association for Computational Linguistics (ACL). 5

[28] Li, S., Feng, S., Wang, D., Song, K., Zhang, Y., and Wang, W. (2021). Emoelicitor: an open domain response generation model with user emotional reaction awareness. In International Joint Conference on Artificial Intelligence (IJCAI). 3

[29] Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In Annual Meeting of the Association for Computational Linguistics (ACL). 3, 6

[30] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586. 6
Mairesse, F. and Walker, M. (2007). Personage: Personality generation for dialogue. In Annual Meeting of the Association for Computational Linguistics (ACL).

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research, 30:457–500.

McCrae, R. R. and Costa Jr, P. T. (1997). Personality trait structure as a human universal. American Psychologist, 52(5):509.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. Journal of Personality, 60(2):175–215.

Mehl, M. R., Gosling, S. D., and Pennebaker, J. W. (2006). Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. Journal of Personality and Social Psychology, 90(5):862.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. Journal of Abnormal and Social Psychology, 66(6):574.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

Prasad, A., Hase, P., Zhou, X., and Bansal, M. (2022). Grips: Gradient-free, edit-based instruction search for prompting large language models. arXiv preprint arXiv:2203.07281.

Raad, B. d. E. and Perugini, M. E. (2002). Big five factor assessment: Introduction. Hogrefe & Huber Publishers.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research (JMLR), 21:1–67.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. Nature, 568(7753):477–486.

Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In ACM Conference on Human Factors in Computing Systems (CHI).

Ribeiro, L. F., Zhang, Y., and Gurevych, I. (2021). Structural adapters in pretrained language models for amr-to-text generation. In Annual Conference on Empirical Methods in Natural Language Processing (EMNLP).

Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., et al. (2022). Multitask prompted training enables zero-shot task generalization. In International Conference on Learning Representations (ICLR).

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, A. N. (2019). The risk of racial bias in hate speech detection. In Annual Meeting of the Association for Computational Linguistics (ACL).

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., and Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In Annual Meeting of the Association for Computational Linguistics (ACL).
[49] Shin, T., Razeghi, Y., Logan IV, R. L., Wallace, E., and Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In Annual Conference on Empirical Methods in Natural Language Processing (EMNLP). 3, 7, 8

[50] Tamkin, A., Brundage, M., Clark, J., and Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv preprint arXiv:2102.02503. 3

[51] Utami, N. A., Maharani, W., and Atastina, I. (2021). Personality classification of facebook users according to big five personality using svm (support vector machine) method. Procedia Computer Science, 179:177–184. 3

[52] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903. 3, 7

[53] Weinberg, R. S. and Gould, D. (2019). Foundations of sport and exercise psychology, 7E. Human kinetics. 1, 2

[54] Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). 5

[55] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In Annual Conference on Empirical Methods in Natural Language Processing (EMNLP). 5

[56] Wu, T., Terry, M., and Cai, C. J. (2022). Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In ACM Conference on Human Factors in Computing Systems (CHI). 3, 7

[57] Yilmaz, T., Ergil, A., and Ilgen, B. (2019). Deep learning-based document modeling for personality detection from turkish texts. In Proceedings of the Future Technologies Conference. 3

[58] Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Annual Conference on Empirical Methods in Natural Language Processing (EMNLP). 5

[59] Zeldes, Y., Padnos, D., Sharir, O., and Peleg, B. (2020). Technical report: Auxiliary tuning and its application to conditional text generation. arXiv preprint arXiv:2006.16823. 3

[60] Zhang, H., Song, H., Li, S., Zhou, M., and Song, D. (2022). A survey of controllable text generation using transformer-based pre-trained language models. arXiv preprint arXiv:2201.05337. 3

[61] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog. do you have pets too? In Annual Meeting of the Association for Computational Linguistics (ACL). 3

[62] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, W. B. (2020). Dialogpt: Large-scale generative pre-training for conversational response generation. In Annual Meeting of the Association for Computational Linguistics (ACL). 3

[63] Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y. N., et al. (2020). Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. Engineering, 6(3):310–345. 2