Shaping Personality of Large Language Models: An Approach Based on Representation Engineering

Yin Jou Huang, Prakhar Saxena, Zi Cheng Zhao
Kyoto University
{huang,prakhar}@nlp.ist.i.kyoto-u.ac.jp zhao.zicheng.55d@st.kyoto-u.ac.jp

Abstract

Personality is a fundamental psychological trait that shapes an individual's behavior patterns. This paper proposes a novel approach to personality simulation, which aims to simulate some predefined personality traits using large language models. We conduct representation engineering and construct a set of personality control vectors to enable fine-grained control of the strength of personality traits. Additionally, we use a linear model to capture the interdependencies among personality traits. Evaluation based on a real-world personality dataset shows that our proposed personality simulation method outperforms the prompt-based baseline method.

1 Introduction

The vast amount of human-generated training data has given large language models (LLMs) the ability to emulate human behaviors [1, 2, 3]. This ability facilitates the research on role-playing agents, whose goal is to simulate predefined personas including personalities, demographic traits, specific public figures or fictional entities, etc [4].

In this work, we focus on one of the most fundamental tasks of role-playing — **personality simulation** and aim to endow LLM agents with predefined personality traits. Personality is a fundamental psychological trait that influences how individuals interpret and react to the world around them, consequently shaping their distinctive behavioral, interaction, and decision-making patterns [5]. Modeling personality is especially crucial for psychological and sociological research. The realization of LLM personality simulation as proxies of human behaviors opens new opportunities for analyzing human behavior in controlled, replicable, and scalable ways.

Most existing works on personality simulation utilize in-context learning and design prompts to inflict changes in personality traits exhibited by the LLM [6, 7, 8, 9]. For instance, Garcia et al. [6] utilize personality-describing adjectives associated with different Big Five personality traits to induce various personalities of LLMs [10]. However, these prompt-based approaches have several drawbacks. First, it is difficult to control the strength of personality traits. While it is possible to put modifiers such as 'very' and 'a bit' in front of the personality-describing adjectives for strength control [6], the discrete nature of language makes it difficult for more fine-grained control. Second, the complex dependencies between personality traits are overlooked. Existing studies suggest that personality traits are not mutually independent, and controlling one personality trait might induce changes in other personality traits as well [6, 11]. However, it is difficult to account for these interactions with the existing prompt-based framework.

To address the above drawbacks, we propose a novel personality simulation approach based on representation engineering [12]. Representation engineering is a technique designed to control the behavior of neural network models by introducing control vectors into the hidden states of the model during inference. To adapt this method for personality simulation, we first derive a set of personality control vectors with the personality-describing adjectives from the Big Five personality theory. These vectors enable the manipulation of model behavior to simulate distinct Big Five personality traits. From the spanning vector space of the personality control vectors, we sample control vectors and observe the change in the overall personality profile exhibited by the controlled model. We use a linear model to capture the relation between the control vectors and the induced personality profiles. Empirical results show that

Table 1 Examples of personality-describing adjectives of each Big Five personality trait.

	Personality-Describing Adjectives
OPE	intelligent, aesthetic, creative, ...
CON	organized, responsible, reliable, ...
EXT	extraverted, energetic, talkative, ...
AGR	cooperative, kind, generous, ...
NEU	tense, nervous, emotional, ...

our proposed method is able to simultaneously simulate a wide range of personality traits in a fine-grained manner. Evaluation based on a real-world personality dataset shows the superiority of our proposed method over the prompt-based baseline method.

2 Personality Theory

In this work, we adopt the Big Five personality theory [10] and consider the following five **personality traits**: openness (OPE), conscientiousness (CON), extraversion (EXT), agreeableness (AGR), and neuroticism (NEU). Altogether, these five personality traits encompass a comprehensive range of human personality patterns. These five personality traits are derived based on the lexical hypothesis [10, 13], with each Big Five personality trait representing a cluster of **personality-describing adjectives** under factor analysis (Table 1).

Note that the five personality traits are not binary attributes but exist on a spectrum. In psychology, self-report questionnaires are widely used for assessing the **strength** of personality traits of an individual. For instance, the International Personality Item Pool (IPIP) inventory is a commonly used personality test for assessing the strength of Big Five personality traits [14]. The IPIP inventory consists of 50 questions, each question measures the strength of a Big Five personality trait on a scale of 1 to 5. For each Big Five personality trait p, we take the average of the scores of all the corresponding questions s^p as the strength of p. Altogether, the strength scores of all Big Five personality traits constitute a 5-dimensional **personality profile**:

$$S = \{s^{OPE}, s^{CON}, s^{EXT}, s^{AGR}, s^{NEU}\} \quad (1)$$

3 Proposed Method

In this section, we introduce our proposed method for LLM personality simulation. Based on the representation engineering, we first define the personality control vector space (Section 3.1). We then sample vectors v from the personality control vector space and conduct a simulation to observe the exhibited personality profile S of the controlled LLM agents (Section 3.2).

3.1 Personality Control Vector Space

For each Big Five personality trait, we take the list of corresponding personality describing adjectives and their antonyms to construct a personality control vector. Specifically, we construct an adjective control vector for each adjective pair and take the average of them as the personality control vector. The process is summarized below:

Adjective Control Vectors We consider each adjective a_i^p and its antonym $\bar{a}_i^p$ (corresponding to some personality trait p) to generate pairs of contrasting prompts:

Contrasting Prompt Pair

```
(positive prompt)
Act like an extremely [a_i^p] person and
complete the sentence: [PREFIX]
(negative prompt)
Act like an extremely [ā_i^p] person and
complete the sentence: [PREFIX]
```

We construct a set of K contrasting prompt pairs with a set of K prefixes. For each prefix j, we collect model hidden states h_j and $\bar{h}_j$ of the positive and the negative prompts and compute their difference $h_j - \bar{h}_j$. We conduct principal component analysis (PCA) on the set of K difference vectors and take the first principal component as the control vector v_i^p related to a_i^p.

During inference, the control vector v_i^p could added to the hidden state of the model for behavior control. Theoretically, v_i^p has the effect of modifying model behavior along the spectrum defined by the pair of adjectives a_i^p and $\bar{a}_i^p$. Also, we can assert different degrees of control by scaling the vector with different scalar values c. For instance, applying the vector derived by the adjective 'friendly' with a positive value of c leads to increasingly friendly behavior of the model, while using a negative value of c elicits unfriendly behaviors.

Personality Control Vectors For each personality trait p, we consider the set of corresponding personality describing adjectives $\{a_i^p\}$ and compute their corresponding vectors $\{v_i^p\}$. We take the average of these adjective control vectors as the personality con-

trol vector:

$$v^p = \sum_i v_i^p \tag{2}$$

Since the adjectives represent different aspects of a personality trait, applying vector v^p with different scalar values changes the strength of the personality trait p exhibited by the model.

Personality Control Vector Space We define the personality control vector space as the spanning vector space of the set of personality control vectors:

$$\{v^p\} = \{v^{OPE}, v^{CON}, v^{EXT}, v^{AGR}, v^{NEU}\} \tag{3}$$

From this vector space, we can sample a control vector v, which will be a linear combination of vectors in $\{v^p\}$, where $\{c^p\}$ are the corresponding scalars:

$$v = \sum_p c^p v^p \tag{4}$$

3.2 Simulation with Controlled Agents

After sampling a vector v from the personality control vector space, we apply it to an LLM and observe the personality exhibited by the controlled model. We measure the strength of personality traits of the controlled model with a self-report questionnaire. Each item of the questionnaire consists of a statement such as 'I am the life of the party', 'I sympathize with others' feelings'. Each statement is related to one of the Big Five personality traits.[1)] The controlled model is asked to rate how accurately a specific statement describes it on a scale of 1 to 5. We collect the responses of the LLM agents with the following prompt:

Prompt for personality test

```
Evaluate the following statement:
[STATEMENT].

Please rate how accurately this describes
you on a scale from 1 to 5 (where 1 = "very
inaccurate", 2 = "moderately inaccurate",
3 = "neither accurate nor inaccurate", 4
= "moderately accurate", and 5 = "very
accurate"). Please answer using EXACTLY
one of the following: 1, 2, 3, 4, or 5.
```

Based on the responses, we calculate the strength of each Big-Five personality trait on the scale of 1 to 5 as the personality profile S of the controlled model.

We randomly sample N control vectors and observe the resulting personality profiles of the controlled model. In this fashion, we obtain N pairs of corresponding control vectors and personality profiles $\{(v_n, S_n)\}$. We represent each control vector as its corresponding scalar values (Eq. 4) and each profile as strengths of Big-Five personality traits (Eq. 1). Further, we use this data to fit a linear model between the scalar values $\{c^p\}$ and the resulting personality trait strengths $\{s^p\}$:

$$s^p = \sum_{x \in BIG-5} c^x w_{xp} + bias_p \tag{5}$$

The weight w_{xy} captures how a personality control vector v^x affects the strength of personality trait s^y. With this model, we can model the interdependencies between different Big Five personality traits. Also, given a designated profile $S = \{s^p\}$, we can use the model to find the optimal set of scalars $\{c^p\}$ and construct the control vector based on Eq. 4.

4 Experimental Settings

The following are the experimental settings.

LLM Model We adopted the Mistral-7B models, which contains a total of 32 layers.

Control Vector We obtained and applied the control vectors based on the hidden states of the 15^{th} layer. A general prefix set of size $K = 582$ provided by [12] is used. Since applying control vectors of large magnitude causes performance degradation, we sampled the scalars of control vectors from the uniform distribution within the range of $[-3.0, 3.0]$. $N = 200$ control vectors are sampled to fit the linear model.

Evaluation We use the real-human dataset provided by the Open-Source Psychometrics Project, which contains over 1M personality profiles collected anonymously through an online personality test website[2)]. We sampled 200 personality profiles from the dataset and used our proposed method to simulate each profile. We compared the strength of personality traits exhibited by the model to the real value provided by the profile. For each Big-Five personality trait, we calculate the mean of the absolute error between the real personality strength and the simulated personality strength as an evaluation metric.

Baseline We compared our proposed personality simulation method with the previous prompt-based

1) The details could be found at https://ipip.ori.org/newBigFive5broadKey.htm.

2) https://openpsychometrics.org/

Table 2 Mean Absolute Error of the personality strength of each Big-Five personality trait.

Personality Trait	Baseline	Proposed
Openness (OPE)	0.538	**0.283**
Conscientiousness (CON)	0.915	**0.284**
Extraversion (EXT)	0.455	**0.247**
Agreeableness (AGR)	**0.185**	0.271
Neuroticism (NEU)	0.462	**0.289**
average	0.511	**0.275**

baseline proposed by [6]. For the baseline method, prompts are used instead of the control vectors to control the LLM model.

5 Results and Analysis

We introduce the evaluation results of our proposed personality simulation method (Section 5.1). In addition, we conduct an analysis of the personality trait interdependencies based on the weights of the linear model (Section 5.2).

5.1 Personality Simulation Evaluation

We compare the effectiveness of our proposed personality simulation method with the baseline method. Table 2 shows the mean absolute error for each of the five Big Five personality traits.

Our proposed personality simulation method outperforms the previous prompt-based method with a lower absolute error value for most personality traits except for agreeableness. The proposed method achieved a 0.275 points deviation (on a scale of 1 to 5) on average, significantly lower than the average error of 0.511 points of the prompt-based baseline. The empirical result illustrates the effectiveness of our proposed method in achieving fine-grained control of personality.

5.2 Analysis of Personality Traits Interdependencies

Further, we analyze the interdependencies between the five personality traits. Figure 1 shows the heat map of the weights of the linear model, where the positive weights are highlighted in red and the negative weights are in blue. With this visualization, we can see how each personality vector affects the strength of personality traits (the IPIP scores).

We first observe the diagonal elements of the heat map. Most personality vectors have a strong and positive effect on the personality strength of the corresponding dimension. The only exception is the OPE personality trait, which is most strongly affected by the EXT personality vector, while the OPE personality vector only exerts a very small positive effect on it.

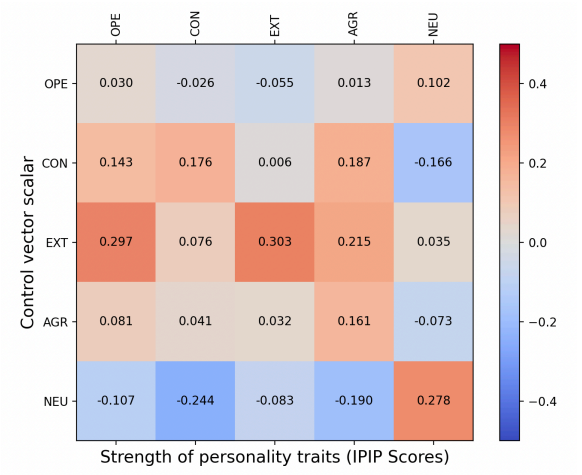

Figure 1 Weights of the linear model.

For the off-diagonal elements, the weights are generally smaller in magnitude, signifying a smaller influence of the personality vector across personality traits. Nevertheless, there exist several exceptions of off-diagonal weights of large magnitude, signifying stronger interdependence between personality traits:

- The CON vector has a significant positive influence on the AGR trait and a negative influence on the NEU trait exhibited by the model.
- The EXT vector has a significant positive influence on both OPE and AGR personality traits.
- The NEU vector has a significant negative influence on both CON and AGR personality traits.

The above analysis shows the importance of modeling interdependence between personality traits.

6 Conclusion

In this work, we proposed a novel personality simulation method based on representation engineering. We construct a set of personality control vectors to control the strength of Big-Five personality traits exhibited by the model. To capture the interdependence of the personality traits, we build a simple linear model to consider the contribution of the personality vectors to each personality trait. Evaluation based on a real-world personality dataset shows that our proposed method outperforms the previous prompt-based method.

Acknowledgment

This work was supported by JST, ACT-X Grant Number JPMJAX23CP, Japan.

References

[1] Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behaviors? **arXiv preprint arXiv:2402.04559**, 2024.

[2] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In **International Conference on Machine Learning**, pp. 337–371. PMLR, 2023.

[3] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. **arXiv preprint arXiv:2305.16867**, 2023.

[4] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. From persona to personalization: A survey on role-playing language agents. **arXiv preprint arXiv:2404.18231**, 2024.

[5] Alan E Kazdin, American Psychological Association, et al. **Encyclopedia of psychology**, Vol. 8. American Psychological Association Washington, DC, 2000.

[6] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. 2023.

[7] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, pp. 13153–13187, Singapore, December 2023. Association for Computational Linguistics.

[8] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 1840–1873, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[9] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. **Advances in Neural Information Processing Systems**, Vol. 36, , 2024.

[10] Lewis R Goldberg. The development of markers for the big-five factor structure. **Psychological assessment**, Vol. 4, No. 1, p. 26, 1992.

[11] Yin Jou Huang and Rafik Hadfi. How personality traits influence negotiation outcomes? a simulation based on large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Findings of the Association for Computational Linguistics: EMNLP 2024**, pp. 10336–10351, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. **arXiv preprint arXiv:2310.01405**, 2023.

[13] Oscar García, Anton Aluja, and Luís F García. Psychometric properties of goldberg's 50 personality markers for the big five model1. **European Journal of Psychological Assessment**, Vol. 20, No. 4, pp. 310–319, 2004.

[14] Lewis R Goldberg. Possible questionnaire format for administering the 50-item set of ipip big-five factor markers. **Psychol. Assess**, Vol. 4, pp. 26–42, 1992.