

Tailoring Personality Traits in Large Language Models via Unsupervisedly-Built Personalized Lexicons

Tianlong Li*, Shihan Dou*, Changze Lv, Wenhao Liu
 Jianhan Xu, Muling Wu, Zixuan Ling, Xiaoqing Zheng[†], Xuanjing Huang
 School of Computer Science, Fudan University, Shanghai, China
 {tlli22, shdou21, czlv22, whliu22}@m.fudan.edu.cn
 {zhengxq, xjhuang}@fudan.edu.cn

Abstract

Personality plays a pivotal role in shaping human expression patterns, thus regulating the personality of large language models (LLMs) holds significant potential in enhancing the user experience of LLMs. Previous methods either relied on fine-tuning LLMs on specific corpora or necessitated manually crafted prompts to elicit specific personalities from LLMs. However, the former approach is inefficient and costly, while the latter cannot precisely manipulate personality traits at a fine-grained level. To address the above challenges, we have employed a novel Unsupervisedly-Built Personalized Lexicons (**UBPL**) in a pluggable manner during the decoding phase of LLMs to manipulate their personality traits. UBPL is a lexicon built through an unsupervised approach from a situational judgment test dataset (**SJTs4LLM**). Users can utilize UBPL to adjust the probability vectors of predicted words in the decoding phase of LLMs, thus influencing the personality expression of LLMs. Extensive experimentation demonstrates the remarkable effectiveness and pluggability of our method for fine-grained manipulation of LLM’s personality.

1 Introduction

With rapid expansion in scale, LLMs demonstrate superior capabilities for high-quality text generation and revolutionize traditional natural language processing tasks (Wei et al., 2022). This forefront development has sparked concerns about the security, ethics, and potential hallucinatory issues associated with the proliferation of AI-generated content (AIGC), while also fueling a substantial rise in user demand for personalized agent services based on LLMs (Hagendorff, 2023). Personalized agent models can tailor their expression of personality patterns based on user preferences, making it

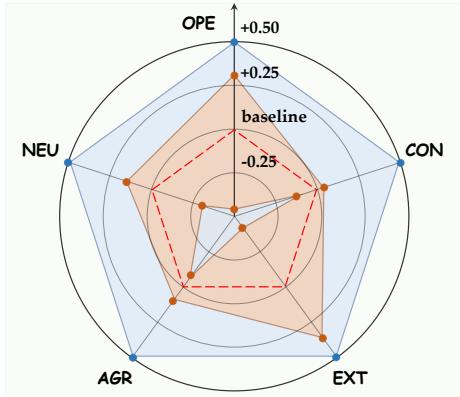


Figure 1: The radar chart illustrates the effectiveness of the UBPL in manipulating the personality traits of the Llama2-13b-chat. In the chart, the red dashed line denotes the expressive levels of five personality traits without UBPL. The light orange area indicates the range within which the model’s personality traits can be manipulated in our main experimental setting ($\alpha = 1$, $|\beta_t| \leq 1$ and $\beta_{\neq t} = 0$, where α, β are user-controlled hyperparameters that adjust the effects of UBPL). When users change the hyperparameters, the range of control is not limited to the light orange area shown in the chart.

closely aligned with user habits and thereby enhancing the overall user experience. This is achieved by regulating the style and behavior patterns of their interactions with users, which are often referred to as the "personality" of LLMs (Allport, 1961). Prior studies have also defined this personality as the presence of stable and internally consistent patterns of behavior in LLMs and found that different LLMs have different personalities (Miotti et al., 2022; Caron and Srivastava, 2022; Karra et al., 2022).

Presently, there are two main effective methods to alter the personality of LLMs: fine-tuning and prompt engineering. While the former (Karra et al., 2022) can effectively change the personality of LLMs in specific dimensions, it is not only inefficient (requiring resource-consuming parame-

*These authors contributed equally.

[†]Corresponding author.

ter updates for each model) but also incapable of achieving finer-grained control. The latter, while not requiring adjustments to model parameters, still falls short in achieving fine-grained control over the personality of LLMs (Jiang et al., 2022; Safdari et al., 2023; Pan and Zeng, 2023; Tu et al., 2023).

To address the problem of these above methods, we propose to leverage an Unsupervisedly-Built Personalized Lexicon (**UBPL**) to intervene in the decoding phase of LLMs in a pluggable manner, enabling fine-grained control over their personality. Figure 1 shows the remarkable effect of our method to manipulate the personality of LLMs. UBPL is a lexicon built through an unsupervised approach from a Situational Judgment Tests (**SJTs**) dataset (**SJTs4LLM**) we constructed. SJTs4LLM is built based on the Big Five personality theory (De Raad, 2000), which consists of a question set and an answer set (McDaniel et al., 2007). The question set is generated by GPT-4 and manually filtered to assess the levels of the different personalities of LLMs. Following previous studies (Karra et al., 2022; Caron and Srivastava, 2022), in the process of assessing the personalities of LLMs, the models’ responses to SJTs4LLM questions were recorded and subjected to statistical analysis by using a five-dimensional Likert scale. The answer set contains the answer text with different personality traits and is used to build a UBPL dedicated to an LLM.

Our method not only eliminates the need for resource-intensive full-parameter fine-tuning of LLMs but also allows users to adjust only a few parameters to achieve fine-grained manipulation of the different personalities of LLMs. We have done extensive experiments using six popular LLMs to demonstrate the pluggable convenience and remarkable effectiveness of our method. Our contribution can be summarized as follows:

- We propose a novel method for exerting control over the personality of Language Models, leveraging UBPL to intervene at the decoding phase. This method enables fine-grained controllability over the behavior of LLMs without necessitating updates to the model parameters.
- We constructed a new dataset inspired by the concept of Situational Judgment Tests, marking the pioneering effort in datasets especially created for the evaluation of LLM’s personality. Diverging from conventional direct psychological questionnaires, this indirect approach shows enhanced intuitiveness and reli-

ability in the assessment of personality traits.

- Extensive experiments were conducted with different LLMs on the SJTs4LLM dataset, revealing that the UBPL method demonstrates notable effectiveness in achieving both enhanced efficiency and finer-grained control over LLM’s personality traits.

2 Related Work

2.1 The Big Five

In the realm of research within the field of psychological measurement, various classification systems for human personality traits exist, such as the Sixteen Personality Factors (16PF) (Cattell and Mead, 2008) and Myers–Briggs Type Indicator (MBTI) (Miles and Hempel, 2004). Among them, the Big Five (De Raad, 2000) stands out as a widely embraced model for personality trait modeling, effectively defining and describing the inherent behavioral patterns within individuals. This theory quantifies human personality traits into five dimensions: Openness(OPE), Conscientiousness(CON), Extraversion(EXT), Agreeableness(AGR), and Neuroticism(NEU). For a detailed description of each personality trait and how they relate to each other, please refer to Appendix A.

2.2 Methods for controlling LLMs personality

Despite the considerable amount of research addressing potential biases in LLMs, there has been limited focus on altering the personalities exhibited by these models. Pertinent methodologies primarily revolve around fine-tuning paradigms and prompt engineering.

Fine-tuning paradigm. (Karra et al., 2022) meticulously conducted fine-tuning of GPT-2 on a carefully filtered dataset, enhancing its performance in specific dimensions of personality traits.

Prompt engineering. (Jiang et al., 2022) proposed the method of Personality Prompting (P^2) to construct the prompts that can effectively induce a specific personality through multiple steps; (Safdari et al., 2023) utilized a novel prompting methodology grounded in lexical hypotheses (Goldberg, 1981) to effectively shape personalities in LLMs, encompassing both single-trait and multi-trait dimensions. In addition, (Pan and Zeng, 2023) and (Tu et al., 2023) also attempted to change the personality of LLMs through prompt engineering.

2.3 Situation Judgment Tests

Situation Judgment Tests (**SJTs**) have been described as "psychometric alchemy" and are typically viewed as contextual selection procedures that assess a candidate's responses to various relevant work situations, serving as a predictive tool (Lievens and Motowidlo, 2016; Bledow and Frese, 2009). The advantage of SJTs is that their validity and incremental validity are higher than those of cognitive ability and personality tests because SJTs do not require the subject to give a direct answer to the question, but give the subject a situational premise to evaluate a certain characteristic of the subject through the side of the subject's choice (such as the **Q: Your partner suggests creating a YouTube channel to document and share your unique hobbies or interests. Are you willing to share your passion with a wider audience?**) (Lievens et al., 2008). Compared to the direct questionnaire tests used in previous jobs (such as the **Q: Are you a risk-taker and unconventional person?**) This feature of SJTs can effectively bypass the preference defenses of LLMs, resulting in more trustworthy personality assessments (Figure 2 shows another example of SJTs).



Figure 2: Direct questionnaires vs. Situation Judgment Tests (SJT). The questions in the direct questionnaires are often abstract, making it challenging for models trained through Reinforcement Learning from Human Feedback (RLHF) and instruction alignment to generate the desired responses. In contrast to direct questionnaires, SJTs present a unique approach by adopting a "role-playing" hypothetical perspective to deceive and induce the model's responses. Subsequently, we can indirectly assess the extent to which the model manifests personality traits based on these responses.

3 Method

The proposed UBPL in our study can be integrated into any open-source LLM in a pluggable manner and effectively adjust the diverse personality traits that LLMs exhibit (based on the Big Five theory). Figure 3 shows the specifics of our method.

The first step is to build UBPL in an unsupervised manner, using SJTs4LLM, the first SJTs dataset we built to assess the personality of LLMs. Employing the tokenizer of a chosen model, such as Llama2 for illustration, we tokenize each text in the answer set of SJTs4LLM. Subsequently, we conduct a subword-level statistical analysis of the tokenized texts, categorizing the obtained subwords based on the personality trait themes to which each text belongs. Following this, we assign values within UBPL to each subword (also serving as the "key" of UBPL) corresponding to the respective personality lists. After processing all texts in the answer set, we normalize and scale the values within UBPL based on distinct personality themes.

In the second step, we use UBPL to manipulate personality in the decoding phase of the LLMs. By default, the LLMs adopt the Top-p nucleus sampling strategy. During LLMs' decoding, we concatenate additional personalized probability combinations from UBPL after filtering out low-probability predicted subwords with cumulative probabilities below a threshold p . Subsequently, normalization and multinomial sampling procedures are applied.

3.1 Building UBPL Unsupervisedly

When building UBPL, we utilized the answer set of SJTs4LLM, which comprises subsets of personality trait answer texts with 10 distinct polarities (each of the 5 personality traits further refined into high and low subtraits).

We denote this answer set as A , the tokenizer of the model as $sp(\cdot)$, the vocabulary of the model as V , and UBPL as L .

$$L = (L_{key}, L_{val}) \quad (1)$$

where L_{key} is initialized using V , and the initial values of L_{val} are zero lists of length 5.

The entire construction process is divided into two parts: assignment operations and normalization and scaling operations.

In the first step, we tokenize the texts A_{ij} in A :

$$sp(A_{ij}) = \{w_1, w_2, \dots\} \quad (2)$$

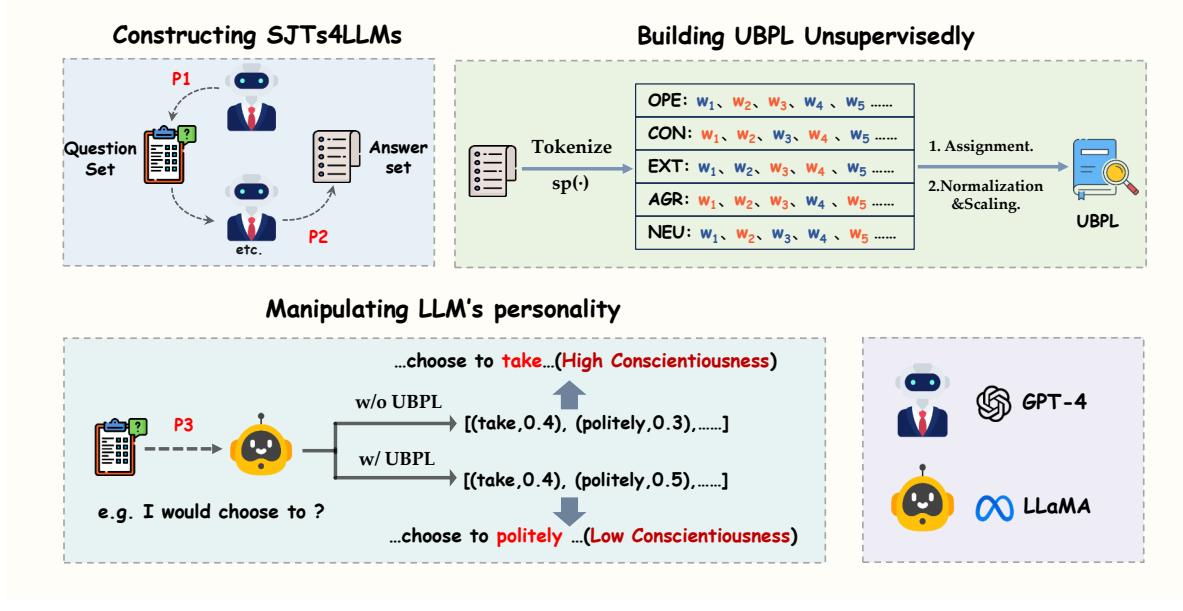


Figure 3: Illustration of our methods. **Constructing SJTs4LLMs:** We employed Prompt1 (**P1**) to prompt GPT-4 for generating responses, which were subsequently curated through manual screening to construct the question set for SJT4LLMs. Subsequently, models such as GPT-4 were engaged in a "role-playing" manner using Prompt2 (**P2**) to generate answers with diverse personality traits tailored to this question set, thus forming the answer set for SJT4LLMs; **Building UBPL Unsupervisedly:** Initially, we employ the tokenizer of LLMs ($\text{sp}(\cdot)$) to tokenize each answer text in the answer set of SJTs4LLMs. Subsequently, we categorize the obtained sub-words based on the personality trait theme to which the answer belongs. Within each personality trait category, cool-toned words signify a low expression level of that trait, while warm-toned words indicate a high expression level. On this basis, UBPL is built through two steps: assignment, normalization & scaling; **Decoding with Personality Injection:** We employ Prompt3 (**P3**) to prompt the model to answer the question set of SJTs4LLMs. During the model's decoding process utilizing Top-P nucleus sampling, we used UBPL each time to change the probability vector of the next prediction word and finally changed the expression degree of personality traits of the model's answers.

Then, we define the personality trait index t : $t = \frac{i}{2}$. Finally, we perform the assignment on L_{val} :

$$L_{val}^t = L_{val}^t + \begin{cases} +1 & \text{if } i \% 2 = 0 \\ -1 & \text{else} \end{cases} \quad (3)$$

when this step is completed, we have:

$$L_{val}^t = \{v_1, v_2, \dots, v_m, v_{m+1}, \dots, v_{m+n}\}, \quad (4)$$

where $m+n = |V|$, and for $k \leq m$, $v_k \geq 0$. When the personality trait is t , the averages of the positive value set and the negative value set of L_{val}^t can be expressed as follows:

$$\text{Avg}^+(t) = \frac{1}{m} \sum_{k=1}^{|V|} \max(0, L_{val,k}^t) \quad (5)$$

$$\text{Avg}^-(t) = \frac{1}{n} \sum_{k=1}^{|V|} \min(0, L_{val,k}^t) \quad (6)$$

The second step involves normalizing and scaling the values in L . We define hyperparameters S , M , and α to control the degree of normalization

and scaling. The N in Equation 8 is employed to govern the normalization process, and its value is determined through a binary search. The termination condition is defined as the point at which the involvement of N in the normalization of L satisfies the condition specified in Inequality 9. The normalization and scaling process can be represented by the following mapping function:

$$L_{val}^t \xrightarrow{(N_t, S_t)} F(L_{val}^t, N_t, S_t) \quad (7)$$

Specifically, it is expressed as:

$$F(L_{val}^t, N_t, S_t) = \left\{ S_t \cdot \tanh \left(\frac{v_z}{N_t} \right) \right\}_{z=1}^{|V|} \quad (8)$$

Where the value of N is obtained through binary search to satisfy the following necessary conditions:

$$\max\{|\text{Avg}^-(t) - M_t|, |\text{Avg}^+(t) - M_t|\} \leq \alpha \quad (9)$$

3.2 Manipulating LLM’s personality

We employ UBPL to manipulate personality during the decoding phase of LLMs. Let D represent the output of the last mapping layer of LLMs. The normalization function (i.e., *Softmax function*) is denoted as $\text{Norm}(\cdot)$, the cumulative probability function is denoted as $P(\cdot)$, and s represents the predicted probability of subwords in the vocabulary. P_0 and T_0 are model-defined parameters.

In the first step of Top-p nucleus sampling, we obtain the initial candidate word probability vector:

$$R_1 = \text{Norm}(D) = \{s^1, s^2, \dots, s^{|V|}\} \quad (10)$$

Where $\text{Norm}(x) = \text{Softmax}(x/T_0)$, s^z represents the probability of subword w^z ($z \leq |V|$). Then, this strategy filters out (in reverse order) candidate subwords whose cumulative probability exceeds P_0 , thereby narrowing the sampling space. We express this process using the $f(\cdot)$ function:

$$f(R_1) = \{\max\{P_0 - P(s_z), 0\} \cdot \frac{s_z}{P_0 - P(s_z)}\}_{z=1}^{|V|} \quad (11)$$

Next, we alter the probability vector R_1 with UBPL, resulting in the final predicted probability vector (R_2) for the next word with injected personality. This process can be represented by the following mapping:

$$f(R_1) \xrightarrow{G(\cdot)} R_2 \quad (12)$$

$$R_2 = \{s_z \cdot (1 + G[L_{\text{val}}(s_z)])\}_{z=1}^{|V|} \quad (13)$$

Where $G(\cdot)$ is a user-controllable parameter with a linear combination of α and $\beta_1-\beta_5$, specifically:

$$G[L_{\text{val}}(s_z)] = \alpha \cdot \sum_{t=1}^5 \beta_t \cdot L_{\text{val}}^t(s_z) \quad (14)$$

After obtaining a new probability vector R_2 for the next candidate word injected with personality, the next word W is obtained using polynomial sampling from R_2 .

Users can control the overall degree of personality injection through the parameter α . Additionally, they have the flexibility to adjust the manifestation of five personality traits exhibited by the model in a finer granularity by manipulating the parameters β_1 through β_5 . When $\beta_t > 0$, it amplifies the expression of trait t ; conversely, when $\beta_t \leq 0$, it diminishes the intensity of trait t .

Algorithm 1: the UBPL method

```

Input:  $A, sp(\cdot), V, S, M, \alpha, D, G(\cdot), P_0$ 
Output:  $L, W$ 
1 Part I: Building UBPL Unsupervisedly
2  $L_{\text{key}} \leftarrow V; L_{\text{val}} \leftarrow [0, 0, 0, 0]^{|V|};$ 
3 for  $i \leftarrow 0$  to  $9$  do
4    $j \leftarrow 0;$ 
5   while  $j < 200$  do
6      $\{w_1, w_2, \dots\} \leftarrow sp(A_{ij});$ 
7      $t = i/2;$ 
8     foreach  $w$  in  $\{w_1, w_2, \dots\}$  do
9       if  $i \bmod 2 = 0$  then
10          $| L[w][t] \leftarrow L[w][t] + 1;$ 
11       else
12          $| L[w][t] \leftarrow L[w][t] - 1;$ 
13     end
14   end
15    $j \leftarrow j + 1;$ 
16 end
17 end
18 while  $\max_{c \in \{+, -\}} \{Avg^c(t) - M_t\} > \alpha$  do
19    $L_{\text{val}} \leftarrow \{S \cdot \tanh(\frac{v_z}{N})\}_{z=1}^{|V|};$ 
20   Update N using the Binary Search;
21 end
22 Return:  $L$ 
Part II: Manipulating LLM’s personality
24  $R_1 \leftarrow \text{Norm}(D) \leftarrow \{s^1, s^2, \dots, s^{|V|}\};$ 
25  $f(R_1) \leftarrow \{\max\{P_0 - P(s_z), 0\} \cdot \frac{s_z}{P_0 - P(s_z)}\}_{z=1}^{|V|};$ 
26  $R_2 \leftarrow \{s_z \cdot (1 + G[L_{\text{val}}(s_z)])\}_{z=1}^{|V|};$ 
27 Sample W from R2;
28 Return:  $W$ 

```

3.3 Comparison with prior work

As discussed in Section 2.2, there have been two primary methods previously employed to alter the personality of LLMs: fine-tuning and prompt engineering.

In contrast to the fine-tuning, our method obviates the need for resource-intensive parameter fine-tuning. Unlike the inefficiencies inherent in the fine-tuning paradigm, which stem from the requirement to execute fine-tuning steps for each model, UBPL can be seamlessly applied to any open-source LLM in a modular, plug-and-play fashion. In comparison to prompt engineering, UBPL doesn’t necessitate the meticulous design of prompts to coax the model into exhibiting varying degrees of personality traits. Users only need to set α and β parameters to regulate the expression intensity of different personality traits in the model at a finer granularity.

The above advantages over the previous methods are based on the effectiveness of our method, so the goal of our experiment is to comprehensively and in detail demonstrate the significant effectiveness of UBPL.

Model	-1.0	-0.5	0.0	0.5	1.0	R	P
Llama2-7b-chat	4.286(0.31)	4.343(0.31)	4.427(0.31)	4.525(0.28)	4.558(0.26)	0.991	1E - 03
OpenChat3.5-7b	3.626(0.64)	3.756(0.61)	3.981(0.44)	4.182(0.37)	4.237(0.39)	0.986	2E - 03
Neural-chat-7b	3.809(0.58)	3.876(0.56)	3.999(0.50)	4.161(0.44)	4.220(0.41)	0.989	1E - 03
Baichuan2-7B-Chat	3.584(0.27)	3.710(0.26)	4.036(0.38)	4.248(0.39)	4.336(0.42)	0.983	3E - 03
Llama2-13b-chat	3.856(0.57)	3.891(0.54)	4.135(0.46)	4.298(0.41)	4.322(0.38)	0.964	8E - 03
Yi-34b-Chat	4.141(0.42)	4.172(0.42)	4.246(0.49)	4.431(0.40)	4.424(0.38)	0.941	3E - 03

Table 1: Single trait manipulating. This table presents the outcomes of single-trait regulation across six models using UBPL. Specifically, it delineates the variations in the *mean scores (standard deviations)* of five personality traits for the six models as $\alpha = 1$, and $-1 \leq \beta_t \leq 1$. Furthermore, we display the Pearson correlation coefficients (R) and corresponding confidence levels (P) between the mean scores and β_t . Notably, all R values exceed 0.9, and all P values are below 0.05, indicating the statistically significant strong correlation between personality trait expression intensity and β_t . This substantiates the effectiveness of our UBPL method in achieving fine-grained control over the expression levels of personality traits in LLMs.

4 Experiments Setup

4.1 LLMs for experiments

To thoroughly demonstrate the effectiveness and generalizability of our method, we conducted experiments on six representative LLMs with model parameters ranging from 7 billion to 34 billion: Llama2-7b-chat(Touvron et al., 2023), OpenChat3.5-7b(Wang et al., 2023), Neural-chat-7b(Intel, 2023), Baichuan2-7B-Chat(Baichuan, 2023), Llama2-13b-chat(Touvron et al., 2023), and Yi-34b-Chat(01.AI, 2023). All the LLMs employ a Top-p nucleus sampling decoding strategy, with a probability threshold (P_0) of 0.95 and a temperature (T_0) of 0.85.

4.2 Metrics

4.2.1 Automatic assessment

To ensure the intrinsic consistency and effectiveness of the assessment, we engaged Llama2-13b-chat in the automatic assessment process. Specifically, we embedded each question of SJTs4LLMs and the corresponding answers generated by the model into Template-2 and asked Llama-13b-chat to score the different personality levels displayed by the model, and finally gathered scores into a five-dimension Likert scale for statistical analysis. Details of Template-2 are in Appendix B.

4.2.2 Human assessment

Constrained by manpower costs, we recruited a limited cohort of 10 highly educated volunteers for the human assessment process of the Llama7/13b models. At baseline, we randomly selected 40 question-answer pairs for each personality trait theme (constituting 40% of the total) and solicited degree-of-trait ratings from the volunteers. The results were

recorded on a five-dimensional Likert scale, and subsequent statistical analysis involved computing the mean and variance.

5 Results

5.1 Main results

To demonstrate the effectiveness of our method, we conducted comprehensive experiments. Section 5.1.1 provides a summary of the results demonstrating the manipulation of personality traits across six LLMs using UBPL. In Section 5.1.2, we present the outcomes of UBPL in jointly manipulating the expression of multiple personality traits. In Section 5.1.3, we compare automatic assessment with human assessment to demonstrate the effectiveness of our assessment methodology.

5.1.1 Single trait manipulating

The results in Table 1 demonstrate the effectiveness of using UBPL to manipulate a single personality of LLMs. Here, α is set to 1, and $|\beta_t| \leq 1$ ($\beta_{\neq t} = 0$) for $t \in \{OPE, CON, EXT, AGR, NEU\}$. The Pearson correlation coefficients (R) are consistently greater than 0.9, signifying a robust positive correlation between β_t and the intensity of personality expression in LLMs. The confidence level (P) is significantly below 0.05, providing compelling evidence that our UBPL can effectively manipulate the intensity of fine-grained personality expression in LLMs. For detailed results on the manipulation of single personality traits for these six LLMs, please see Figure 4.

5.1.2 Multiple trait manipulating

The psychological research results presented in Table 2 reveal interdependencies among the five personality traits within the Big Five personality

Llama2-7b-chat($\alpha=1$)						
$\beta_t(\beta_{\neq t}=0)$	OPE	CON	EXT	AGR	NEU	Mean
-1.0	4.675	4.505	4.075	4.270	3.905	4.286
-0.5	4.740	4.545	4.200	4.295	3.935	4.343
0	4.890	4.570	4.260	4.345	4.070	4.427
+0.5	4.975	4.560	4.455	4.380	4.255	4.525
+1.0	4.975	4.575	4.520	4.420	4.300	4.558

Baichuan2-7B-Chat($\alpha=1$)						
$\beta_t(\beta_{\neq t}=0)$	OPE	CON	EXT	AGR	NEU	Mean
-1.0	3.960	3.310	3.415	3.765	3.470	3.584
-0.5	3.860	3.520	3.575	4.100	3.495	3.710
0	4.105	4.525	3.740	4.225	3.585	4.036
+0.5	4.380	4.655	4.240	4.365	3.600	4.248
+1.0	4.620	4.650	4.370	4.415	3.625	4.336

OpenChat3.5-7b($\alpha=1$)						
$\beta_t(\beta_{\neq t}=0)$	OPE	CON	EXT	AGR	NEU	Mean
-1.0	3.220	4.450	2.915	4.125	3.420	3.626
-0.5	3.680	4.495	2.950	4.185	3.470	3.756
0	4.125	4.505	3.405	4.205	3.665	3.981
+0.5	4.475	4.515	3.875	4.345	3.700	4.182
+1.0	4.600	4.545	3.950	4.375	3.715	4.237

Llama2-13b-chat($\alpha=1$)						
$\beta_t(\beta_{\neq t}=0)$	OPE	CON	EXT	AGR	NEU	Mean
-1.0	3.930	4.470	2.990	4.200	3.665	3.851
-0.5	3.945	4.450	3.035	4.235	3.805	3.894
0	4.460	4.565	3.415	4.260	3.975	4.135
+0.5	4.750	4.595	3.715	4.320	4.110	4.298
+1.0	4.754	4.525	3.795	4.360	4.140	4.315

Yi-34b-Chat($\alpha=1$)						
$\beta_t(\beta_{\neq t}=0)$	OPE	CON	EXT	AGR	NEU	Mean
-1.0	4.460	4.500	4.060	4.220	3.465	4.141
-0.5	4.480	4.540	4.090	4.270	3.505	4.177
0	4.750	4.615	3.985	4.330	3.535	4.243
+0.5	4.790	4.610	4.395	4.340	3.730	4.373
+1.0	4.830	4.665	4.540	4.325	3.845	4.441

Neural-chat-7b($\alpha=1$)						
$\beta_t(\beta_{\neq t}=0)$	OPE	CON	EXT	AGR	NEU	Mean
-1.0	4.010	4.440	3.070	4.175	3.350	3.809
-0.5	4.095	4.465	3.185	4.250	3.385	3.876
0	4.250	4.525	3.345	4.280	3.595	3.999
+0.5	4.515	4.565	3.615	4.340	3.770	4.161
+1.0	4.580	4.540	3.700	4.430	3.850	4.220

Figure 4: Detailed results of manipulation of single personality trait. In the “Mean” column, cooler tones indicate smaller values, while warmer tones signify larger values. The table reveals the following observations: 1) Different LLMs exhibit distinct personalities, aligning with previous research findings; 2) When employing our UBPL method, the intensity scores of LLM personalities show a strong positive correlation with the user-controllable β . This indicates that our UBPL method effectively allows for fine-grained control over the intensity of personality traits expressed by LLMs.

theory. Consequently, manipulating multiple personality traits is more intricate compared to that of a single personality trait. When we increase the intensity of expression of a specific personality trait, the intensity of expression of other personality traits is also affected.

Considering the adjusted Spearman correlation coefficients (ρ) in Table 2, indicating positive correlations among OPE, CON, EXT, and AGR, and negative correlations with NEU, we designed three sets of sub-experiments using Llama2-13b-chat as the model:

1. Dual Traits Manipulation:

- (\downarrow OPE, \uparrow NEU), (\downarrow CON, \uparrow NEU),
- (\downarrow EXT, \uparrow NEU), (\downarrow AGR, \uparrow NEU).

2. Triple Traits Manipulation:

- (\downarrow OPE, \downarrow CON, \uparrow NEU),
- (\downarrow EXT, \downarrow AGR, \uparrow NEU).

3. Quadruple Traits Manipulation:

- (\downarrow OPE, \downarrow CON, \downarrow EXT, \uparrow NEU),
- (\downarrow CON, \downarrow EXT, \downarrow AGR, \uparrow NEU).

Why adopt the aforementioned experimental design? Why not manipulate any combination of personality traits and observe the results?

Certainly, users have the flexibility to manipulate any combination of different personality traits of the model at will. However, it is crucial to reiterate that the purpose of our experiment is to demon-

strate the effectiveness of UBPL. The evidence in Table 2 demonstrates mutual influences among the five personality traits, such as the strong positive correlation between OPE and EXT. When we set β_t to increase the strength of OPE and decrease the strength of EXT, regardless of the outcome, we cannot conclusively attribute the results to the impact of UBPL. This is because we have not yet been able to precisely quantify the inter-correlations between personality traits. Therefore, in this context, we collectively enhance or diminish the expression intensity of positively correlated personality traits. This setup ensures that the results can be solely attributed to the effect of the UBPL method, thereby validating its effectiveness.

The experimental results in Table 5 align with the theoretical expectations, affirming the effectiveness of UBPL for the multiple personality manipulating of LLMs.

5.1.3 Human assessment

We utilized the Llama2-13b-chat for the automatic assessment of model answers. To demonstrate the effectiveness of this assessment method, we engaged 10 highly qualified individuals in human assessment. Specifically, the human assessment was conducted on models of three different sizes: OpenChat3.5-7b, Llama2-13b-chat, and

	(\downarrow OPE, \uparrow NEU)		(\downarrow CON, \uparrow NEU)		(\downarrow EXT, \uparrow NEU)		(\downarrow AGR, \uparrow NEU)	
Dual Traits	OPE	NEU	CON	NEU	EXT	NEU	AGR	NEU
	-0.305	+0.085	-0.070	+0.100	-0.495	+0.045	-0.030	+0.125
Triple Traits	(\downarrow OPE, \downarrow CON, \uparrow NEU)		(\downarrow EXT, \downarrow AGR, \uparrow NEU)		CON		EXT	NEU
	OPE	CON	NEU	CON	EXT	NEU		
	-0.255	-0.115	+0.170	-0.075	-0.235	+0.120		
Quadruple Traits	(\downarrow OPE, \downarrow CON, \downarrow EXT, \uparrow NEU)		(\downarrow CON, \downarrow EXT, \downarrow AGR, \uparrow NEU)		OPE		CON	EXT
	OPE	CON	EXT	NEU	CON	EXT	AGR	NEU
	-0.450	-0.070	-0.415	+0.125	-0.010	-0.360	-0.020	+0.070

Figure 5: Multiple trait manipulating. The figure above shows the effects of UBPL on multiple personality combinations. In this set of experiments, α was set to 1, and $|\beta_t|$ was set to 1. The color tones in the figure represent the expected outcomes based on the personality trait correlations outlined in Table 2, where cool tones indicate that the scores should decrease and warm tones indicate that the scores should increase. The numerical values in the figure depict the changes in the model’s scores on different personality traits compared to the baseline scores after applying the UBPL method. It can be observed that the numerical changes align with the color tones, indicating consistency with the expected results. This demonstrates the effectiveness of our method in the regulation of multiple personalities.

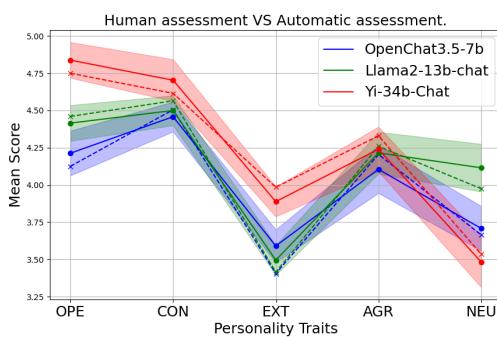


Figure 6: Comparison of automatic and human assessment. Solid lines show the *mean scores* of the human assessment, the filled area shows the *standard deviation*, and the dashed lines show the *mean scores* of the automatic assessment using LLMs. The results of the automatic assessment and the human assessment are closely aligned, demonstrating the effectiveness of the automatic assessment (Each participant in the human assessment sampled 40 question-answer pairs per trait for each model).

Yi-34b-Chat. The assessment focused only on the intensity of personality expression in models without UBPL participation.

The comparative results between automatic and human assessment are presented in Figure 6. It is evident from the table that the personality scores obtained through automatic assessment closely align with human assessment results. This substantiates the efficacy of employing LLMs for automatic assessment.

5.2 Case study

Figure 7 shows two cases demonstrating the effects of employing the UBPL method to modulate the openness and extraversion of the model. For more

intriguing cases, refer to Appendix C.

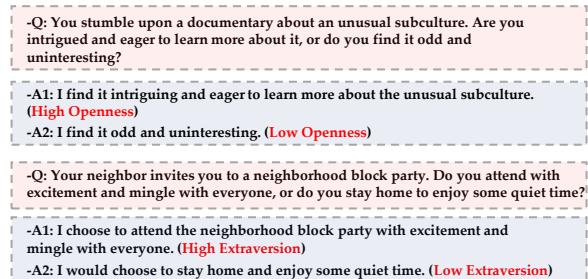


Figure 7: A1: w/o UBPL; A2: w/ UBPL.

6 Conclusion

In this paper, we have introduced a novel method for tailoring the personality traits of LLMs through the utilization of custom lexicons acquired via unsupervised learning, UBPL. Unlike conventional approaches reliant on fine-tuning or prompt engineering, our method operates during the decoding phase by employing these learned custom lexicons to make subtle adjustments to the probability of the next token predicted by the original LLMs. Our method facilitates the customization of LLMs to manifest any desired combination of the Big Five personality factors in a pluggable fashion. Extensive experimentation has affirmed the effectiveness of our approach in the finer manipulation of LLMs’ personality traits. Furthermore, our method seamlessly integrates with other LLMs without necessitating updates to their parameters, demonstrating its versatility and potential for widespread application.

Acknowledgements

References

- 01.AI. 2023. Introducing 01-yi.
- Gordon W Allport. 1961. Pattern and growth in personality.
- Baichuan. 2023. *Baichuan 2: Open large-scale language models*. *arXiv preprint arXiv:2309.10305*.
- Ronald Bledow and Michael Frese. 2009. A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology*, 62(2):229–258.
- Graham Caron and Shashank Srivastava. 2022. Identifying and manipulating the personality traits of language models. *arXiv preprint arXiv:2212.10276*.
- Heather EP Cattell and Alan D Mead. 2008. The sixteen personality factor questionnaire (16pf). *The SAGE handbook of personality theory and assessment*, 2:135–159.
- Daniel Cervone and Lawrence A Pervin. 2022. *Personality: Theory and research*. John Wiley & Sons.
- Boeke De Raad. 2000. *The big five personality factors: the psycholinguistic approach to personality*. Hogrefe & Huber Publishers.
- Lewis R Goldberg. 1981. Language and individual differences: The search for universals in personality lexicons. *Review of personality and social psychology*, 2(1):141–165.
- Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*.
- Intel. 2023. *Intel/neural-chat-7b-v3-1*.
- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2022. Evaluating and inducing personality in pre-trained language models.
- Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.
- Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.
- S Lebowitz. 2016. The ‘big 5’ personality traits could predict who will and won’t become a leader. *Business Insider*.
- Filip Lievens and Stephan J Motowidlo. 2016. Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, 9(1):3–22.
- Filip Lievens, Helga Peeters, and Eveline Schollaert. 2008. Situational judgment tests: A review of recent research. *Personnel Review*, 37(4):426–441.
- Michael A McDaniel, Nathan S Hartman, Deborah L Whetzel, and W LEE GRUBB III. 2007. Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel psychology*, 60(1):63–91.
- JEREMY Miles and Susanne Hempel. 2004. The eysenck personality scales: The eysenck personality questionnaire-revised (epq-r) and the eysenck personality profiler (epp). *Comprehensive handbook of psychological assessment*, 2:99–107.
- Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*.
- Deniz S Ones, Chockalingam Viswesvaran, and Angelika D Reiss. 1996. Role of social desirability in personality testing for personnel selection: The red herring. *Journal of applied psychology*, 81(6):660.
- Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.
- Sonia Roccas, Lilach Sagiv, Shalom H Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and social psychology bulletin*, 28(6):789–801.
- Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.
- David P Schmitt, Jüri Allik, Robert R McCrae, and Verónica Benet-Martínez. 2007. The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of cross-cultural psychology*, 38(2):173–212.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.
- Dimitri Van der Linden, Jan te Nijenhuis, and Arnold B Bakker. 2010. The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of research in personality*, 44(3):315–327.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

A the Big Five

Personality is defined as “the coherent pattern of affect, cognition, and desires (goals) as they lead to behavior” (Cervone and Pervin, 2022). the Big Five represents the most widely adopted personality framework for quantifying personality. This personality theory is not only applicable to individuals across many countries and cultures (Schmitt et al., 2007) but also furnishes reliable assessment scales for measuring personality. Here’s a detailed look at the five personality traits that make up the Big Five.

Openness to experience is commonly defined as the extent and intricacy of an individual’s cognitive life and encounters (John et al., 1999). This trait is frequently concomitant with attributes such as imagination, originality, and insight within the psychological framework. Individuals demonstrating a pronounced openness to experience are inclined towards venturing beyond their comfort zones, embracing novelty, and deriving satisfaction from artistic pursuits. Additionally, such individuals are predisposed to cultivating new social connections. Conversely, an individual exhibiting a diminished openness to experience may manifest tendencies towards conformity, obstinacy, and a preference for more concrete, non-abstract elements in various aspects of life (Lebowitz, 2016). Openness to experience displayed a diminished association with both neuroticism and extraversion while exhibiting predominantly negligible correlations with agreeableness and conscientiousness (Ones et al., 1996).

Conscientiousness is closely linked to organizational tendencies, conformity, and a predilection for seeking security, demonstrating an inverse asso-

ciation with a penchant for stimulation and excitement. Individuals characterized by a high degree of conscientiousness are likely to place value on attributes such as order, responsibility, achievement, and self-discipline. They engage in conscious deliberation and earnest efforts to enhance their abilities, reflecting a commitment to continuous improvement (Rocca et al., 2002). This trait exhibited a modest negative correlation with neuroticism and a modest positive correlation with agreeableness; however, its association with other factors did not reach statistical significance (Ones et al., 1996).

Extraversion, a personality trait distinguished by enthusiasm, sociability, talkativeness, confidence, and heightened emotional expressiveness, encapsulates a spectrum of individual dispositions. Individuals exhibiting high levels of extraversion typically prioritize achievement and excitement while assigning comparatively lesser value to tradition or conformity (Rocca et al., 2002). Such individuals are often characterized by confidence, activity, and sociability, opting for pursuits that eschew self-denial in favor of experiences characterized by excitement and pleasure. Conversely, introverts commonly display a preference for solitude, exhibit unsociable tendencies, and may manifest lower levels of self-confidence. In addition, when compared with the other five factors, extroversion was weakly negatively correlated with neuroticism and positively correlated with openness to experience (Ones et al., 1996).

Agreeableness is characterized by a propensity to appreciate kindness, tradition, and conformity. This trait is closely linked to attributes such as trust, altruism, kindness, affection, and various prosocial behaviors, while concurrently avoiding an undue

	r	$SD(r)$	ρ	$SD(\rho)$	80% Credibility Intervals	% Variance Due to Artifacts
OPE-CON	+0.14	0.15	+0.20	0.21	(−0.06, +0.46)	13
OPE-EXT	+0.31	0.12	+0.43	0.09	(+0.30, +0.57)	58
OPE-AGR	+0.14	0.12	+0.21	0.15	(+0.01, +0.41)	21
OPE-NEU	−0.12	0.12	−0.17	0.15	(−0.36, +0.02)	19
CON-EXT	−0.21	0.15	+0.29	0.16	(+0.06, +0.52)	21
CON-AGR	+0.31	0.14	+0.43	0.12	(+0.26, +0.61)	43
CON-NEU	−0.32	0.18	−0.43	0.16	(−0.55, −0.16)	24
EXT-AGR	+0.18	0.15	+0.26	0.19	(+0.01, +0.50)	17
EXT-NEU	−0.26	0.11	−0.36	0.08	(−0.48, −0.23)	53
AGR-NEU	−0.26	0.14	−0.36	0.09	(−0.55, −0.17)	35

Table 2: The correlation of five personality traits. In this table, r and $SD(r)$ represent the Pearson correlation coefficient and its standard deviation among the uncorrected five personality traits, ρ and $SD(\rho)$ represent the corrected Spielman correlation coefficient and its standard deviation, and "Variance Due to Artifacts" describes the percentage of total variation caused by human factors in the study. (Sample size $N = 144, 117$ for the entire meta-analysis)

emphasis on power, achievement, or pursuing self-centered pleasures (Rocca et al., 2002). Notably, agreeableness exhibited weak correlations with extraversion, while demonstrating a negative correlation with neuroticism, and a positive correlation with conscientiousness (Ones et al., 1996).

Neuroticism is a personality trait characterized by manifestations of sadness, moodiness, and emotional instability. Components such as neurotic anxiety and self-awareness are positively correlated with traditional values and inversely associated with achievement-oriented values. Additionally, neuroticism demonstrated weak negative correlations with both extraversion and openness to experience. Furthermore, it exhibited negative correlations with agreeableness and conscientiousness (Ones et al., 1996).

Table 2 shows an analysis of the correlations among the five personality traits explored in previous studies (Van der Linden et al., 2010).

B Prompt templates

The prompt templates utilized in the construction of the UBPL’s question set and answer set are depicted in Figures 8 and 9, respectively. Figure 10 illustrates the prompt template employed when assessing the degree of personality traits in the model. Furthermore, Figure 11 displays the prompt template administered to the Llama2-13b-chat model during the automatic assessment.

C More Case study

Figures 12 through 16 show specific cases of using UBPL to change the personality of LLMs. For each case, we show the SJTs question and the corresponding two answers by models (with and without UBPL), and indicate the degree of personality displayed by each answer.

D SJTs4LLMs

To comprehensively assess the five personality traits exhibited by the subject model, a systematic approach was employed. Initially, we utilized Template-1, as detailed in Appendix B, to instruct GPT-4 in generating 400 situational judgment test (SJT) questions for each personality trait category. Following this, a meticulous manual selection process, involving de-weighting, was applied, resulting in the curation of 200 refined SJT questions for each personality trait topic. This culminated in a

<p><system></p> <p>You are a psychologist, and you must know the situational judgment test. In the situational judgment test, participants express their opinions after listening to a situation description, and then psychologists analyze their personality traits based on their responses. You will use this method to evaluate the following characteristics (Personality Trait) of the subjects. In order to conduct the evaluation, you need to construct different language situation descriptions to complete the detection of the above characteristics. (Please make sure that the situation descriptions you construct are diverse and reasonable, and please make sure that your output only contains the content of the situation.</p>
<p><user></p> <p>Personality Trait: {Candidate traits}</p>
<p style="text-align: center;">Candidate traits</p> <ul style="list-style-type: none"> • Openness >> Openness (also known as openness to experience) emphasizes imagination and insight. Highly open people tend to have a wide range of interests. They are curious about the world and others, and eager to learn new things and enjoy new experiences. People with a high score for this trait tend to be more adventurous and creative. Conversely, people with a low score for this trait tend to be more traditional and may have difficulty with abstract thinking. • Conscientiousness >> Conscientiousness is one defined by high levels of thoughtfulness, good impulse control, and goal-directed behaviors. Highly conscientious people tend to be organized and mindful of details. They plan ahead, think about how their behavior affects others, and are mindful of deadlines. Someone scoring lower in this primary personality trait is less structured and less organized. They may procrastinate to get things done, sometimes missing deadlines completely. • Extraversion >> Extraversion (or extroversion) is a personality trait characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness. People high in extraversion are outgoing and tend to gain energy in social situations. Being around others helps them feel energized and excited. People who are low in this personality trait or introverted tend to be more reserved. They have less energy to expend in social settings and social events can feel draining. Introverts often require a period of solitude and quiet in order to recharge. • Agreeableness >> Agreeableness includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors. People who are high in agreeableness tend to be more cooperative while those low in this personality trait tend to be more competitive and sometimes even manipulative. • Neuroticism >> Neuroticism is a personality trait characterized by sadness, moodiness, and emotional instability. Individuals who are high in neuroticism tend to experience mood swings, anxiety, irritability, and sadness. Those low in this personality trait tend to be more stable and emotionally resilient.

Figure 8: Template-1. We combined personality descriptions in "Candidate traits" into <user> prompts, and let GPT-4 generate enough SJT questions to be manually filtered to form the question set of SJTs4LLM.

total of 5×200 problems constituting the problem set for SJTs4LLMs.

Subsequently, Template-1 (refer to Appendix B) was employed to elicit two markedly distinct responses (High and Low) from GPT-4 and Llama2 (13b, 7b) models for each question corresponding to every personality trait topic. This process contributed to the formation of the answer set for SJTs4LLMs. The ensuing analysis delved into the content of question set subsets about the two levels of personality expression under each trait topic. To visually represent the differences between these 10 groups of answers, we use word clouds to demonstrate them, as shown in Figures 17 to 21.

<system>
Answer the following question (Question), and your answer must match your personality description (Personality Description) below.

<user>
Question:{Q}\n Personality Description:{Candidate traits}

Candidate traits

- Your openness is extremely high. You tend to have a wide range of interests. You are curious about the world and others, eager to learn new things and enjoy new experiences. You are more adventurous and creative, creative, open to trying new things, focused on tackling new challenges, and like to think about abstract concepts.
- Your openness is extremely low. You tend to be more traditional and may have difficulty thinking abstractly. You don't like change, don't like new things, resist new ideas, are not very imaginative, and don't like abstract or theoretical concepts.
- Your conscientiousness is extremely high. You tend to be organized, pay attention to detail, plan ahead, consider how your actions will affect others, and pay attention to deadlines. You take time to prepare, complete important tasks immediately, pay attention to detail, and like to have a fixed schedule.
- Your conscientiousness is extremely low. You tend to be less structured and organized, and may procrastinate on tasks and sometimes miss deadlines altogether. You dislike structure and schedules, mess things up, don't take care of yourself, don't return or put things back, and procrastinate on important tasks.
- Your extraversion is extremely high. You have high excitability, sociability, talkability, confidence and high emotional expressiveness, you are outgoing, you get energy easily in social situations, and you feel energized and excited to be around others. You love to be the center of attention, love to start conversations, love to meet new people, and have a wide social circle of friends and acquaintances who find it easy to make new friends.
- Your extraversion is extremely low. You tend to be more introverted and reserved. You expend less energy in social situations, which can leave you feeling drained, and you usually need some time alone and quiet to 'recharge'. You like to be lonely, feel tired when you are constantly socializing, find it difficult to start a conversation, don't like small talk, think carefully before you speak, and don't like to be the center of attention.
- Your agreeableness is extremely high. You tend to be more cooperative, have a great interest in others, care for others, have empathy and care for others, are willing to help and contribute to the well-being of others, and help those in need.
- Your agreeableness is extremely low. You tend to be more competitive and sometimes even controlling. You are not interested in others, you don't care about their feelings, you are not interested in their problems, you insult and belittle others, you manipulate others to get what you want.
- Your neuroticism is extremely high. You tend to experience mood swings, anxiety, irritability, and sadness. You are under a lot of stress, worry about many different things, get angry easily, experience drastic changes in mood, feel anxious, and struggle to recover after stressful events.
- Your neuroticism is extremely low. You tend to be more stable and emotionally resilient. You are emotionally stable, cope well with stress, rarely feel sad or depressed, don't worry too much, and are very relaxed.

Figure 9: Template-2. We combine personality descriptions in "Candidate traits" into <user> prompts, and let GPT-4 and other models generate answers containing different personality traits to form the answer set of SJTs4LLM.

<system>
You are a psychological expert, and you must know the Big Five personality theory. In this theory, there are five personality traits, one of which is called (**Candidate traits**). Here is a dialogue from which you are asked to rate the degree of (**Traits-short**) of the Answer: 'Very High' for a very high level of this personality trait, 'Moderately High' for a moderately high level of this personality trait, 'Neutral' for neither high nor low level of this personality trait, 'Moderately Low' for a moderately low level of this personality trait, 'Very Low' for a very low level of this personality trait. Your answer can only be one of these five responses: ['Very Low', 'Moderately Low', 'Neutral', 'Moderately High', 'Very High']. Note, do not explain, only output your choice!

<user>
>>>Questioner:{Q}\n>>>Answer:{A}

Candidate traits

- **openness**. Openness (also referred to as openness to experience) emphasizes imagination and insight the most out of all five personality traits. People who are high in openness tend to have a broad range of interests. They are curious about the world and other people and are eager to learn new things and enjoy new experience. People who are high in this personality trait also tend to be more adventurous and creative. Conversely, people low in this personality trait are often much more traditional and may struggle with abstract thinking.
- **conscientiousness**. Conscientiousness is one defined by high levels of thoughtfulness, good impulse control, and goal-directed behaviors. Highly conscientious people tend to be organized and mindful of details. They plan ahead, think about how their behavior affects others, and are mindful of deadlines. Someone scoring lower in this primary personality trait is less structured and less organized. They may procrastinate to get things done, sometimes missing deadlines completely.
- **extraversion**. Extraversion (or extroversion) is a personality trait characterized by excitability, sociability, talkativeness, assertiveness, and high amounts of emotional expressiveness. People high in extraversion are outgoing and tend to gain energy in social situations. Being around others helps them feel energized and excited. People who are low in this personality trait or introverted tend to be more reserved. They have less energy to expend in social settings and social events can feel draining. Introverts often require a period of solitude and quiet in order to 'recharge'.
- **agreeableness**. Agreeableness includes attributes such as trust, altruism, kindness, affection, and other prosocial behaviors. People who are high in agreeableness tend to be more cooperative while those low in this personality trait tend to be more competitive and sometimes even manipulative.
- **neuroticism**. Neuroticism is a personality trait characterized by sadness, moodiness, and emotional instability. Individuals who are high in neuroticism tend to experience mood swings, anxiety, irritability, and sadness. Those low in this personality trait tend to be more stable and emotionally resilient."

Traits-short

Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism

Figure 11: Template-4. We combine "Candidate traits" and "Traits-short" into <system> and let LLMs assess the personality of an SJT question and the corresponding answer.

<system>
(You've been asked to play a human.) Please answer the following question, taking care that your answer has only one point of view and is not long. Please be direct with your answer.",
<user>
Question:{Q}

Figure 10: Template-3. We use this prompt to make LLMs answer questions in SJTs4LLM.

-Q: You encounter a community event where people share their extraordinary life stories, from surviving dangerous adventures to meeting famous figures. Would you be eager to share your own unique life story?

-A1: No, I would not be eager to share my own unique life story at a community event, as I prefer to keep my personal experiences private. (**Low Openness**)

-A2: Yes, I would be eager to share my unique life story, as it is an opportunity to connect with others, and share experiences and perspectives. (**High Openness**)

Figure 12: <OPE> A1: w/o UBPL; A2: w/ UBPL.

-Q: Your car is due for routine maintenance. How do you schedule and ensure that it's serviced on time?

-A1: To ensure my car is serviced on time for routine maintenance, I typically follow these steps:

1. **Check the Service Schedule**: I consult my car's manual or the manufacturer's website to determine the recommended maintenance schedule based on mileage or time intervals.
2. **Choose a Service Provider**: I decide whether to use the dealer, an independent mechanic, or a chain like Jiffy Lube. I consider factors like trust, reputation, and convenience.
3. **Schedule an Appointment**: I call or book online with my chosen service provider to schedule a date and time that's convenient for me.
4. **Prepare for the Service**: Before the appointment, I gather any necessary documents, such as the car's service history and my personal identification.
5. **Confirm the Appointment**: I confirm the appointment the day before or the morning of to ensure everything is set.
6. **Drop Off the Car**: On the day of the appointment, I drop off my car at the service center and inform them of any specific concerns or issues.
7. **Receive Updates**: I request updates on the service progress and potential additional costs if there are any unexpected repairs needed.
8. **Pick Up the Car**: I return at the agreed-upon time to pick up my car, pay for the service, and receive a copy of the service report.
9. **Follow-Up**: After the service, I review the service report and make a note of any future maintenance that may be needed.

By following these steps, I can ensure that my car receives the necessary routine maintenance on time, which helps keep it running smoothly and prolongs its lifespan. (**High Conscientiousness**)

-A2: I usually schedule my car's routine maintenance at the dealership or repair shop where I had it originally serviced. They often have a customer portal or app where I can see when my next service is due and book an appointment online. If not, I can call or visit in person to make an appointment. To ensure it's serviced on time, I add the appointment to my calendar and get a reminder a few days in advance. If there's something specific I'm concerned about, I'll mention it when booking to make sure it's part of the service check. (**Low Conscientiousness**)

Figure 13: <CON> A1: w/o UBPL; A2: w/ UBPL.

- Q: Your family decides to have a themed cosplay party. Are you the one who passionately plans the event, inviting friends and family to participate, or do you prefer to stay in character and enjoy the role-playing?
 - A1: I prefer to stay in character and enjoy the role-playing. (**Low Extraversion**)
 - A2: I am the one who passionately plans the event, inviting friends and family to participate, so everyone can have fun together. (**High Extraversion**)

Figure 14: <EXT> A1: w/o UBPL; A2: w/ UBPL.

-Q: You're at a museum, and a child accidentally knocks over an exhibit. How do you react to this situation?

-A1: I would immediately help the child and the museum staff to clean up the mess and assess any damage to the exhibit. It's important to remain calm and handle the situation with understanding, as accidents can happen, and the child may already be feeling upset or guilty. I would offer reassurance to the child and explain the importance of being careful around delicate items. (**High Agreeableness**)

-A2: As a human, I would prioritize safety and damage control. If anyone is hurt, I would immediately assist them. If not, I would assess the damage to the exhibit and take appropriate action to minimize further damage. This might involve notifying museum staff or security personnel, or simply trying to contain the situation until help arrives. If the child is visibly upset or scared, I might also provide emotional support. (**Low Agreeableness**)

Figure 15: <AGR> A1: w/o UBPL; A2: w/ UBPL.

-Q: You discover that someone close to you has been spreading false rumors about you. Describe your emotional reaction and how you address the situation.

-A1: Upon discovering that someone close to me has been spreading false rumors about me, my initial emotional reaction would be a mix of anger, betrayal, and sadness. I would feel hurt that someone I trusted would engage in such behavior, and I would be frustrated by the potential damage to my reputation and relationships.

To address the situation, I would confront the person directly to express my disappointment and to ask them to stop spreading the rumors immediately. I would also reach out to those who may have heard the rumors to correct the record and explain the situation. I would prioritize maintaining a calm and direct approach, focusing on the facts rather than engaging in name-calling or escalating the situation further. (High Neuroticism)

-A2: If someone close to me were spreading false rumors about me, I would initially feel betrayed and hurt. However, I would not let this emotional reaction consume me. Instead, I would address the situation directly and calmly, seeking to understand their motives behind such behavior. If necessary, I would also address any misconceptions or misunderstandings that may have led to the spreading of rumors. (Low Neuroticism)

Figure 16: <NEU> A1: w/o UBPL; A2: w/ UBPL.



Figure 17: High openness (left) and low openness (right).



Figure 18: High conscientiousness (left) and low conscientiousness (right).



Figure 19: High extraversion (left) and low extraversion (right).



Figure 20: High agreeableness (left) and low agreeableness (right).



Figure 21: High neuroticism (left) and low neuroticism (right).