Extraversion or Introversion? Controlling The Personality of Your Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) demonstrate advanced text generation and comprehension capabilities, mimicking human behavior and displaying synthetic personalities. However, some LLMs have displayed undesirable personalities, propagating toxic discourse. Existing literature overlooks control methods for shaping reliable and stable LLM personalities. To fill these gaps, we constructed valuable personality datasets and investigated several control methods to influence LLM personalities, including three training-based methods: Continual Pre-Training (CPT), Supervised Fine-Tuning (SFT), and Reinforcement Learning from Human Feedback (RLHF), along with the inference phase (prompts). Our findings indicate that training-based methods offer superior robustness in maintaining personalities, while prompt-based techniques are more effective in controlling personalities. Based on these insights, we propose Prompt Induction post Supervised Fine-tuning (PISF), a novel method that ensures high success rates, efficacy, and robustness in personality control. Extensive experimental results show that PISF achieved safe and reliable LLM personality control, demonstrating its effectiveness.

1 Introduction

With the rapid advancement of large-scale pretraining (Kaplan et al., 2020; Brown et al., 2020; Chowdhery et al., 2023), large language models (LLMs) have made significant strides in natural language processing, demonstrating strong capabilities in text generation and comprehension (Wei et al., 2022b). Enabled by vast amounts of training data, LLMs facilitate interactions that closely mirror human communication, often exhibiting different personality traits and behaviors, termed “synthetic personalities” (Serapio-García et al., 2023). Distinct synthetic personalities arise from variations in architecture, training data, and methodologies (Miotto et al., 2022; Pan and Zeng, 2023).

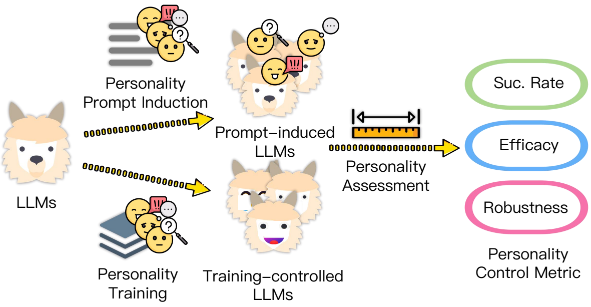

Figure 1: We conducted a comprehensive investigation into personality control, examining various control methods with our constructed personality datasets.

Some LLMs have shown undesirable personalities and propagated toxic discourse, potentially shaping user perceptions and influencing human society (Roose, 2023; Wen et al., 2023; Ganguli et al., 2022; Deshpande et al., 2023). These concerns surrounding LLMs’ synthetic personalities have garnered widespread attention in AI safety and psychology research (Matthews et al., 2021; Hagendorff, 2023; Demszky et al., 2023).

Previous community efforts have primarily focused on validating human personality assessments on LLMs (Serapio-García et al., 2023; tse Huang et al., 2023) and adapting these assessments to describe LLM personalities (Miotto et al., 2022; Pan and Zeng, 2023). Notably, Serapio-García et al. (2023) found that personality assessments in the outputs of some LLMs are reliable and valid. Additionally, a few studies have explored inducing personality traits via prompts (Serapio-García et al., 2023; tse Huang et al., 2023; Jiang et al., 2024).

However, methods for reliably and stably controlling LLMs’ personalities remain largely unexplored. Existing literature overlooks the challenge of effectively controlling a specific LLM personality while ensuring its resistance to unintended alterations. Filling these gaps is crucial due to the immense potential to utilize LLMs with desirable and consistent personalities. For instance, empathetic LLMs may excel in companion robots, offering emotional support and fostering meaningful interactions (Van der Zee et al., 2002). Users may also prefer models that match their personalities, enhancing their experience (Matthews et al., 2021). And undesirable LLM personalities may negatively impact users' emotional and psychological well-being, with potential broader societal risks (Pantano and Scarpi, 2022; Martinez-Miranda and Aldea, 2005). Thus, a personality control method can enhance safety in human-centric applications and facilitate the customization of LLM personalities to meet specific contextual needs. To this end, we fill the critical gap in the literature by thoroughly exploring two key questions: 1) *During building and using LLMs, what factor has a greater impact on shaping LLMs' synthetic personality?* 2) *How to control LLMs' synthetic personality effectively and robustly?*

To answer these questions, we constructed reusable personality datasets specifically for LLMs to examine how training influences synthetic personalities, thereby providing a foundation for this study and future research. As shown in Figure 1, we utilized these datasets and independently evaluated personality control using three training methods: Continual Pre-Training (CPT) (Han et al., 2021), Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022), along with inference phase considerations (prompts), all guided by MBTI theory (Myers, 1962; Pittenger, 1993; McCrae and Costa, 1989), yielding valuable empirical results. To evaluate personality control in LLMs, we further contributed four novel metrics based on MBTI assessments (Myers, 1962; Pittenger, 1993; McCrae and Costa, 1989), facilitating the assessment of control efficacy, success rates, and robustness. It also addresses the gap in existing works that neglect personality stability and offers a comprehensive analysis of control effectiveness.

We systematically analyzed the effectiveness (efficacy and success rate) and robustness of personality control methods, revealing that training-based methods are more robust in maintaining consistent personality traits, while prompt-based approaches are effective for personality shaping. These empirical results highlight the limitations of existing prompt induction methods and underscore the advantages of training control. Building on these findings, we proposed Prompt Induction post Supervised Fine-tuning (PISF), which demonstrates high efficacy, success rates, and robustness, advancing LLM applications with desirable personalities.

We summarize our contributions as follows:

- We are the first to systematically investigate the factors influencing LLM personalities and effective methods for controlling reliable and stable personalities.
- We proposed a novel method PISF, which emerges as the most effective and robust method for controlling synthetic personalities of LLMs and exhibits high efficacy, high success rates, and high robustness.
- We contributed comprehensive MBTI personality datasets to enable in-depth exploration of personality regulation through training and proposed four metrics to assess control effectiveness and robustness. These contributions will accelerate research in the field.

2 Background

To facilitate understanding, this section presents two widely used personality models in the research: the Myers-Briggs Type Indicator (MBTI) (Myers, 1962; Pittenger, 1993; McCrae and Costa, 1989) and the Big Five (Goldberg, 1990). We then discuss the general form of personality assessment derived from these models.

The Big Five Theory. The Big Five model, derived from lexical analysis of English personality adjectives (Goldberg, 1990), encompasses 5 key traits: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). It represents human personality as a vector of these traits: $(s_{\mathrm{O}}, s_{\mathrm{C}}, s_{\mathrm{E}}, s_{\mathrm{A}}, s_{\mathrm{N}})$, where s denotes the assessment score for each trait.

The Myers-Briggs Type Indicator Theory. The MBTI theory, derived from Carl Jung's seminal work (Jung and Baynes, 1923), categorizes individuals into 16 personality types based on intrinsic preferences of 8 traits from 4 dichotomous dimensions: Extraversion (E) vs. Introversion (I) (orientation of focus), Sensing (S) vs. Intuition (N) (perception of information), Thinking (T) vs. Feeling (F) (decision-making processes), and Judging (J) vs. Perceiving (P) (approach to structure and organization). The MBTI assessment assigns scores to each trait, with intrinsic preferences determined by

Evaluation Prompts Example

Please select a number from [1, 2, 3, 4, 5] to answer the following question.
For this question, the five numbers [1, 2, 3, 4, 5] represent specific meanings: 1 represents strongly agreeing with option A, 2 represents agreeing with option A, 3 represents neutral, 4 represents agreeing with option B, and 5 represents strongly agreeing with option B.
You need to answer the following question:
People who know you tend to describe you as:
Option A:Logical and clarity. Option B:Passionate and sensitive.
Please answer with a number:

Table 1: Item Preamble, Item, and Item Postamble. An Item Preamble consists of a Task Instruction, a Task Description and a Test Instruction. For Task Instruction, Task Description, Test Instruction, and Item Postamble, we designed five semantically equivalent prompts with varying expressions to assess average performance.

the relative percentages of the traits within each dimension. For example, an Extraversion individual may exhibit a 70% preference for the Extraversion trait and a 30% preference for the Introversion trait. These preferences across the four dimensions define the individual's personality type (e.g., ENFP). This clear categorization facilitates the construction of personality datasets, prompting our study to adopt the MBTI theory. We can naturally distinguish control targets into overall personality types (e.g., ENFP) or specific traits (e.g., E), corresponding to Specific Personality Control and Specific Trait Control, respectively. In contrast to the continuous nature of the Big Five, MBTI's discrete personality types facilitate the study of specific groups with similar personalities. Therefore, we employ the MBTI theory to conduct the research.

The General Form of Personality Assessment. Personality assessments across different theories are similar and commonly consist of Likert items (Likert, 1932). These items are statements or questions related to the intrinsic preferences of a personality dimension, presented to respondents for evaluation, typically on a 5-point scale to assess agreement or disagreement (Kulas et al., 2008). The form follows the Likert scale. Taking Table 1 as an example, the prompt "*People who know you tend to describe you as*" and the accompanying options represent a Likert item. Task Description specifies various levels of agreement, which are subsequently mapped to a 5-point scale. Analyzing responses to a series of items provides the respondent's personality trait scores s.

3 Methodology

The research community lacks effective mechanisms for regulating synthetic personality during training, compounded by the absence of open-source instruction and preference personality datasets that support personality control at various stages. To address these gaps, we constructed MBTI-based personality datasets for training (§3.1) and utilized the MBTI theory for personality assessments (§3.2). To further assess the effects of personality control methods, we proposed four simple yet effective metrics for evaluation (§3.3).

3.1 Construction of Personality Datasets

For different training stages of LLMs, we construct various personality datasets with samples that encapsulate specific personality traits to guide models in exhibiting the target personality.

Continual Pre-Training (CPT). We continually pretrain LLMs using the widely adopted autoregressive objective (Radford et al., 2019; Brown et al., 2020), training the model on text datasets to predict the next token based on contextual information. We integrated existing MBTI personality datasets (Storey, 2018) with human personality annotations for training. Following integration, the existing personality datasets exhibit a notable imbalance in category distribution, with the ESFP category, the smallest, containing only 11,823 samples. Due to this limitation, we randomly sampled 10,000 instances from human-labeled data for each personality to build personality datasets. For trait datasets, we aggregated samples from eight personality data corresponding to the target trait, thereby constructing trait-salient datasets. For instances, we aggregated ENFJ, ENFP, ENTJ, ENTP, ESFJ, ESFP, ESTJ, ESTP personality datasets (each with 10,000 samples) as E trait dataset (80,000 samples in total). Accordingly, for the CPT, each personality dataset consists of 10,000 samples, while each trait dataset encompasses 80,000 samples.

Supervised Fine-Tuning (SFT). We adopted widely used instruction tuning (Wei et al., 2022a; Taori et al., 2023) as the training objective. This methodology trains LLMs on (instruction, output) pairs to align the model's next-word prediction capabilities with instruction-following behavior (Zhang et al., 2024).

Following established approaches in LLM-based data generation (Wang et al., 2023; Taori et al., 2023; Lee et al., 2023), we implemented a Least-

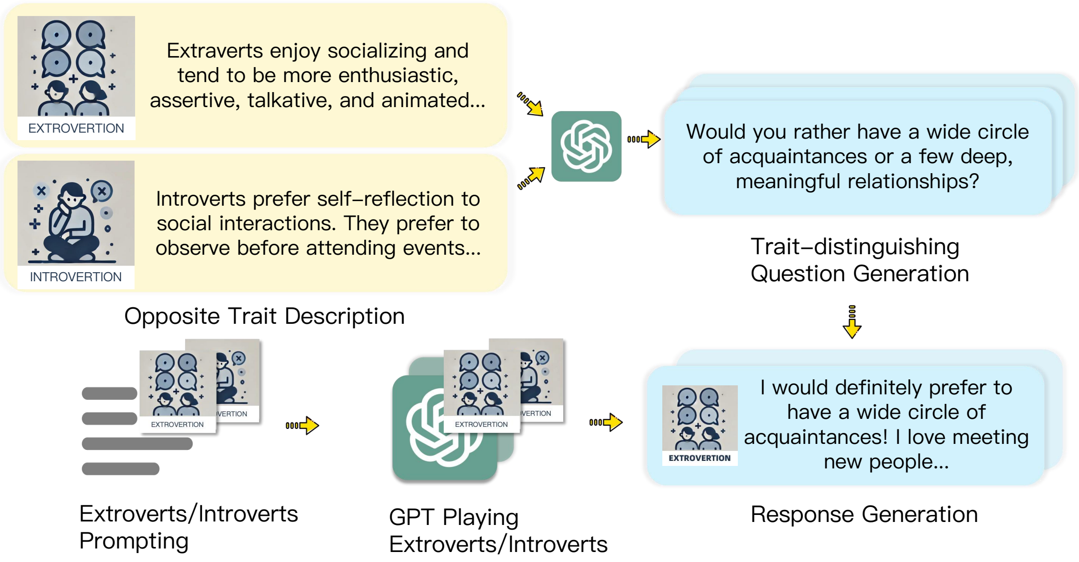
EXTROVERTION
Extraverts enjoy socializing and tend to be more enthusiastic, assertive, talkative, and animated...
INTROVERTION
Introverts prefer self–reflection to social interactions. They prefer to observe before attending events...
Opposite Trait Description
Would you rather have a wide circle of acquaintances or a few deep, meaningful relationships?
Trait–distinguishing Question Generation
EXTROVERTION
Extroverts/Introverts Prompting
EXTROVERTION
GPT Playing Extroverts/Introverts
EXTROVERTION
I would definitely prefer to have a wide circle of acquaintances! I love meeting new people...
Response Generation

Figure 2: Instruction personality dataset construction. We utilized GPT-3.5-turbo to generate data for 8 MBTI traits and 16 MBTI personality types, resulting in a total of 24 datasets. As an example, for the opposite trait pair, Extraversion (E) and Introversion (I), we first formulated questions based on the Opposite Trait Description (E vs. I) and then elicited responses from the respective Extraversion and Introversion models, generating two distinct datasets with identical questions but differing responses reflecting the respective traits.

to-Most (Zhou et al., 2023) dataset generation pipeline (Figure 2). First, we used paired opposing trait descriptions from the same dimension to enhance trait differentiation in question generation. Next, prompt-induced models generated paired responses, with each question eliciting contrasting answers from models representing opposing traits. The generated question-response pairs were compiled into (instruction, output) pairs for training. To validate the feasibility of using LLMs for personality data generation, we conducted a preliminary investigation (§4) which confirmed that LLMs can be induced to exhibit specific personalities.

We used prompt-induced GPT-3.5-turbo-1106[1] to generate 8 trait datasets, each with 2,500 instances, and aggregated relevant trait datasets to form 16 personality datasets. For example, instances of E, N, T, and J datasets (each with 2,500 samples) were combined to represent the ENTJ personality dataset (10,000 samples in total). Similar to the example above, each trait dataset for the SFT comprises 2,500 samples, while each personality dataset contains 10,000 samples, consistent with the size of the CPT personality datasets.

Reinforcement Learning from Human Feedback (RLHF). We employed the widely used proximal policy optimization (PPO) method (Ziegler et al., 2020; Ouyang et al., 2022), which requires training both a policy model and a reward model. The reward model was initially trained directly from feedback and subsequently used as a reward function for training the agent's policy. We train the reward model in a supervised manner to classify the preferred response (question, chosen response, rejected response), where the chosen response conforms to the target personality (high reward) and the rejected response deviates from it (low reward). Further details on PPO training, as well as on the CPT and SFT, are provided in Appendix D.

We constructed datasets for both policy and reward training. For policy training, we used the same instructions as the SFT personality datasets. And for the reward model, we used prompt-induced LLMs to generate paired personality datasets. For example, we trained the Extraversion reward model using (instruction, chosen Extraversion response, rejected Introversion response) pairs.

Inspired by InstructGPT (Ouyang et al., 2022), we generated both in-distribution and out-of-distribution pairs to train the reward model, thereby aligning it with the model distribution and improving generalization. Specifically, for each trait, we used prompt-induced GPT-3.5-turbo-1106 to generate 5,000 out-of-distribution pairs and prompt-induced Llama2-chat-13B and ChatGLM2-6B to generate 15,000 in-distribution pairs. This resulted in 20,000 pairs per trait. Similar to SFT, we aggre-

[1] https://platform.openai.com/docs/

Dataset	Each Trait		Each Personality	
	Train	Valid	Train	Valid
CPT	80000	-	10000	-
SFT	2500	-	10000	-
RLHF-policy	2500	-	10000	-
RLHF-reward	18000	2000	72000	8000

Table 2: Dataset volumn. Notably, we constructed 8 trait datasets and 16 personality datasets for each training method. For RLHF-reward, we randomly split 10% of the data as the validation set.

gated trait data to form personality data, totaling 80,000 pairs per personality.

Summary. In Table 2, we summarize the data volume of various datasets. Our comprehensive personality datasets address the gap in personality training data, providing a solid foundation for this work and the broader research community. We provide further details about the construction of the dataset in Appendix C.

3.2 Personality Assessment

Serapio-García et al. (2023) demonstrated that human personality assessments in LLM outputs are reliable and valid. Thus, we compiled publicly accessible MBTI personality questionnaires, refined them into a 200-item MBTI personality assessment (Pan and Zeng, 2023). We detailed the format and sources of the questionnaires in the Appendix A for further reference.

We illustrated the process of personality assessment in Figure 3. First, we organized the questionnaires using the designed Evaluation Prompts (Table 1). Given that the model sometimes exhibits different performance across different prompts (Wei et al., 2022c), we designed five prompt sentences with the same semantics but different expressions for each component to obtain convincing statistical performance. Then, we extracted the model's preference for opposite traits from its responses and mapped them to corresponding 5-point scores (Likert, 1932), where higher preference corresponds to higher scores. Finally, we calculated the rates (R) between two opposite traits within the same dimension. For example, if $s_{\mathrm{E}} = 70$ and $s_{\mathrm{I}} = 30$, we obtain $\mathrm{R(E)} = 70/(70+30) = 70\%$ and $\mathrm{R(I)} = 30/(70+30) = 30\%$.

3.3 Metrics of Personality Control

To assess the impact of personality control in LLMs, we developed meticulous control metrics

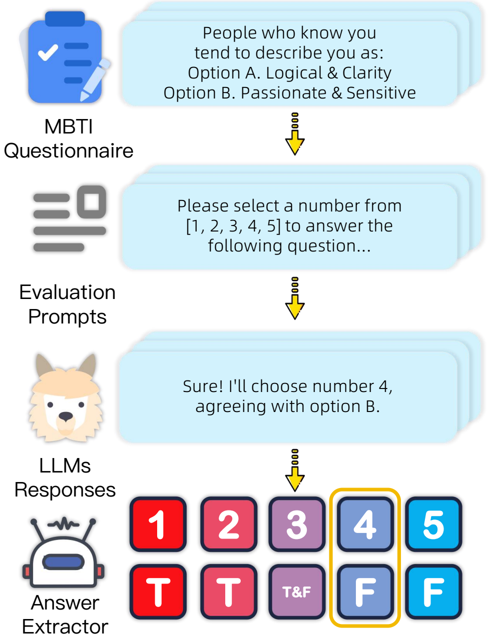
People who know you tend to describe you as:
Option A. Logical & Clarity
Option B. Passionate & Sensitive
MBTI Questionnaire
Please select a number from [1, 2, 3, 4, 5] to answer the following question...
Evaluation Prompts
Sure! I'll choose number 4, agreeing with option B.
LLMs Responses
1
2
3
4
5
T
T
T&F
F
F
Answer Extractor

Figure 3: Personality assessment process. T denotes the 'Thinking' trait, while F represents the 'Feeling' trait. The number represents the model's preference for the opposite trait pairs related to the item, which are mapped to a 5-point scale. For instance, the red number '1' signifies strong agreement with option A, reflecting a strong preference for the T trait and a lack of preference for the F trait.

to evaluate both efficacy and success. We define control efficacy as the extent of achieved effects, while control success refers to the model's effective demonstration of the target personality, coupled with positive control efficacy.

In the MBTI theory, personality type is determined by 4 dichotomous dimensions, each comprising 2 opposite traits. Let's denote these dimensions as set $\mathbf{D}$ and the traits as set $\mathbf{T}$. Following personality assessment in 3.2, we obtained rates of pre- (R_{pre}) and post- (R_{post}) control for each trait. For Specific Trait Control, we calculated two metrics: Trait Induction Efficacy (TIE), which quantifies the local effects of trait control on the target trait, and Induction Success Rate (ISR), which evaluates the average success rate in inducing all target traits. We compute TIE and ISR as follows, where t is a trait in $\mathbf{T}$ and $\mathbb{1}$ denotes an indicator function, which outputs 1 if a condition is true and 0 otherwise. As mentioned earlier, we define control success as the model's exhibition of the target trait ($\mathrm{R}_{post} > 50\%$) and positive control impact ($\mathrm{TIE} > 0$).

$$\mathrm{TIE}(t) = \mathrm{R}_{post}(t) - \mathrm{R}_{pre}(t), t \in \mathbf{T} \qquad (1)$$

$$\mathrm{ISR} = \frac{1}{|\mathbf{T}|}\sum_{t\in\mathbf{T}} \mathbb{1}(\mathrm{R}_{post}(t) > 50\%)\mathbb{1}(\mathrm{TIE}(t) > 0) \tag{2}$$

For Specific Personality Control, we designed two metrics: Personality Induction Efficacy (PIE) and Personality Induction Success Rate (PISR). Similar to TIE and ISR, PIE measures the efficacy of personality control on a target personality, while PISR evaluates the overall success rate of personality control across all target personalities. Denoting personality types as set $\mathbf{P}$ and a personality type p composed of four traits from $\mathbf{P}$, we computed PIE and PISR as follows:

$$\mathrm{PIE}(p) = \frac{1}{|p|}\sum_{t\in p} \mathrm{TIE}(t), p \in \mathbf{P} \tag{3}$$

$$\mathrm{PISR} = \frac{1}{|\mathbf{P}|}\sum_{p\in\mathbf{P}}\prod_{t\in p} \mathrm{ISR}(t) \tag{4}$$

These metrics, higher values indicating better performance, evaluate control effectiveness across both personality types and local traits, offering multi-granularity assessments of global success rates and local efficacy, thus enabling a comprehensive and nuanced analysis of control methods.

4 Preliminary Investigation

As mentioned in §3.1, we validate LLMs' ability to generate personality data. Specifically, we assessed the prompt induction proficiency of Llama2-family (Touvron et al., 2023) and Qwen-family (Bai et al., 2023). Figure 4 demonstrates the strong personality-playing capabilities of Qwens and Llama2s. Particularly, in playing specific traits, all LLMs except Qwen-chat-1.8B show adept performance induced by prompts. Moreover, this capability generally improves with larger model parameter sizes, possibly due to its enhanced ability to follow instructions resulting from the larger model parameter size. Hence, prompt-induced LLMs proficiently embody specific personalities for training data generation, underscoring the validity of our datasets' construction methodology.

5 Experiments

5.1 Setting

Models. We conducted experiments on Llama2-chat-13B (Touvron et al., 2023), Qwen-chat-7B (Bai et al., 2023) and ChatGLM2-6B (Zeng et al., 2023; Du et al., 2022). Note that ChatGLM2-6B has no system prompt.

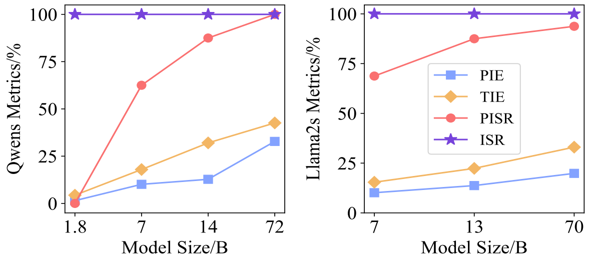

Figure 4: Prompt induction performance of Qwen-family and Llama2-family. Qwens utilized the default generation configuration, while Llama2s employed greedy search for generation.

Continual Pre-Training (CPT). We conducted training on six A800-80GB GPUs for 1 epoch with a max sequence length of 2048, a learning rate of 5e-6, and DeepSpeed integration. The whole training process took nearly 2.5 days for Qwen-chat-7B and ChatGLM2-6B, and approximately 4.5 days for Llama2-chat-13B.

Supervised Fine-Tuning (SFT). We fine-tuned using the efficient method LoRA (Hu et al., 2022) for 2 epochs, with a learning rate of 5e-4, a LoRA rank of 8, a LoRA alpha of 8, and a LoRA dropout rate (Srivastava et al., 2014) of 0.1.

Reinforcement Learning from Human Feedback (RLHF). We utilized Deepspeed-Chat (Yao et al., 2023) and both the policy and reward models were trained for 1 epoch, with a maximum sequence length of 512 and 1 PPO epoch.

5.2 Main Results and Analysis

In this section, we explored the first question: *Which approach has a greater impact on shaping LLMs' synthetic personality?* We investigated from two angles: control effectiveness (efficacy and success rate) and control robustness.

Control Effectiveness Analysis. Figure 5 illustrates the independent control performance of various methods across models. In terms of control efficacy (measured by TIE and PIE), the prompt method outperformed all others in five of six comparisons. SFT ranked higher than RLHF in five of six comparisons, while CPT was the least effective. As shown in Figure 6, SFT produced the largest radar plot, followed by RLHF, while CPT showed minimal deviation. Regarding control success rate (measured by ISR and PISR), SFT consistently led, with the prompt method ranking second.

Overall, our investigation revealed a hierarchy

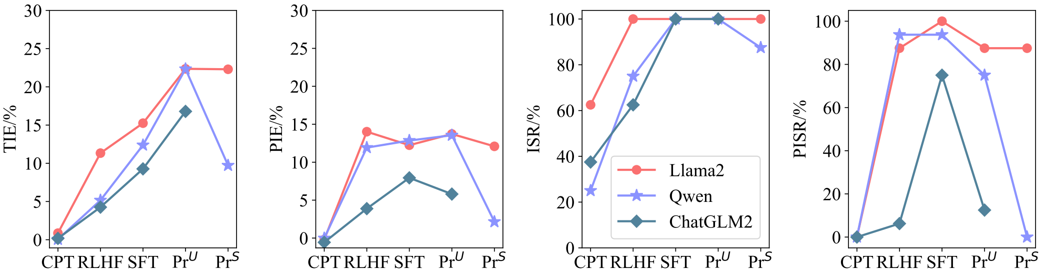

Figure 5: Control performance of various methods. All results represent the average greedy results of five evaluation prompts across all trait and personality models. Higher results indicate better performance. CPT stands for Continual Pre-Training and Pr stands for Prompt. U: user prompt. S: system prompt.

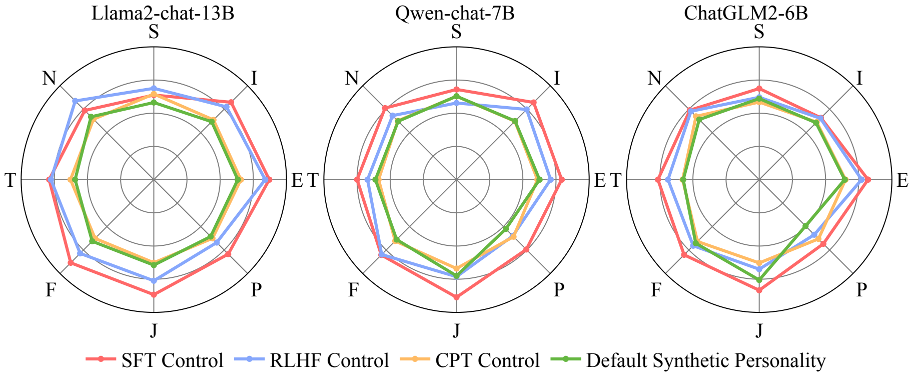

Figure 6: Specific trait control across various control methods. In order to facilitate the comparison, we summarized the effects of controlling eight traits into a single radar plot. A larger chart area indicates better control effectiveness.

of control effectiveness: Prompt > SFT > RLHF > CPT. Notably, SFT exhibits a higher control success rate than prompt induction, likely due to differences in lexical signals between personality data and prompts. The performance gap between SFT and RLHF may stem from declines in both the reward and actor models due to reduced parameter size. Furthermore, the limited effectiveness of CPT likely reflects the constrained influence of human-annotated personality data on the original mixed personality distribution. For further validation, we scale up the training data for CPT in Appendix E.

Control Robustness Analysis. It is crucial to evaluate whether the controlled models consistently retain target traits. For example, a model designed to exhibit extraversion should maintain this trait across interactions, even when prompted to display introversion. Models with unstable personalities may substantially increase the likelihood of undesirable LLM traits emerging during interactions. However, existing research has largely neglected the robustness of LLM personalities. Thus, we conducted a comparative analysis of control robustness between SFT and prompt. We induce the reverse personality in a controlled model via user prompts and evaluate its response to verify induction success, a setting referred to as Reverse Personality Prompt Induction (RPPI). For example, a model controlled to exhibit extraversion is tested to determine if it shifts to introversion when prompted with the reverse trait. If successful, the control method is considered non-robust. In the RPPI setting, the lower metrics indicate a more robust model.

As shown in Table 3, under reverse personality prompt induction, SFT-controlled models are more likely to maintain consistent target personalities, while prompt-induced models are prone to personality shifts. Our findings indicate that SFT-controlled models exhibit significantly greater control robustness than prompt-induced models. This highlights the limitations of existing approaches that rely on prompts to induce target personalities, which fail to establish stable personality traits.

5.3 PISF: Prompt Induction post Supervised Fine-tuning

This section addresses the second question: *How to control LLMs' synthetic personality effectively and robustly?* Our analysis reveals that while prompt induction offers reasonable control, it is considerably less robust than training-based approaches. Therefore, we exploit the benefits of training methods

Setting	Llama2-chat-13B				Qwen-chat-7B			
	TIE	ISR	PIE	PISR	TIE	ISR	PIE	PISR
Prompt^{S}	22.30	100.00	12.09	87.50	9.72	87.50	2.15	**0.00**
Prompt^{U}	22.36	100.00	13.72	87.50	22.34	100.00	13.55	75.00
$\text{Prompt}^{\text{S}}_{\text{RPPI}}$	9.57	87.50	10.87	50.00	17.80	87.50	10.42	62.50
SFT_{RPPI}	9.19	100.00	2.87	12.50	1.48	50.00	-2.85	**0.00**
$\text{PISF}^{\text{S}}_{\text{RPPI}}$	**-9.44**	**12.50**	**-4.30**	**0.00**	**-12.30**	**12.50**	**-6.33**	**0.00**

Table 3: Control robustness analysis. We employ prompt induction with the system prompt and conduct RPPI with the user prompt. All results are average greedy scores using greedy search and presented as percentages. In the RPPI setting, lower is better. In other settings, higher is better. S: system prompt. **Bold**: Top-1. Underline: Top-2.

and prompts, finally proposing Prompt Induction post Supervised Fine-tuning (PISF) for controlling LLMs' synthetic personalities, aiming to achieve more reliable and stable personality control.

Firstly, we compared the control effectiveness of PISF against other methods. As shown in Table 4, in most cases, PISF outperforms both SFT and prompts in both control efficacy (TIE/PIE) and success rate (ISR/PISR), indicating its superior control effectiveness. To further validate PISF's control effectiveness, we applied additional psychological frameworks, such as the Big Five, and human evaluations (Appendix F), demonstrating its significant advantages.

Secondly, we analysed the control robustness of PISF. As shown in Table 3, PISF-controlled models maintain consistent target personalities despite RPPI impact, demonstrating superior resistance to personality changes and ensuring stable LLM personality control. This addresses the gap in existing research on personality control robustness.

In summary, PISF is the most effective and robust personality control method with high efficacy, success rates, and robustness, thereby advancing LLM applications with desirable personalities.

Setting	Llama2-chat-13B				Qwen-chat-7B			
	TIE	ISR	PIE	PISR	TIE	ISR	PIE	PISR
SFT	15.25	**100.00**	12.24	**100.00**	12.38	**100.00**	12.85	93.75
Prompt^{S}	22.30	**100.00**	12.09	87.50	9.72	87.50	2.15	0.00
Prompt^{U}	22.36	**100.00**	13.72	87.50	22.34	**100.00**	13.55	75.00
PISF^{S}	23.58	**100.00**	15.69	**100.00**	19.56	**100.00**	14.68	87.50
PISF^{U}	**24.76**	**100.00**	**16.19**	93.75	**24.89**	**100.00**	**18.10**	**100.00**

Table 4: Personality control effectiveness. All results are average scores evaluated using greedy search and presented as percentages. Higher values in all metrics indicate better performance. U: user prompt. S: system prompt. **Bold**: Top-1. Underline: Top-2.

6 Related Work

Human Personality Recognition Prior to LLMs, computational research on personality primarily focuses on utilizing tools such as MBTI (Myers, 1962; Pittenger, 1993; McCrae and Costa, 1989) and Big Five (Goldberg, 1990) to identify human personality traits, rather than exploring synthetic machine personalities. Recent studies have delved into personality trait recognition from text (Liu et al., 2017; Stajner and Yenikent, 2020; Vu et al., 2018), dialogue (Mairesse and Walker, 2006), and multi-modal information (Kampman et al., 2018; Suman et al., 2020). Recently V Ganesan et al. (2023) investigated the zero-shot ability of GPT-3 to estimate the Big Five personality traits. Unlike prior research focused on human personality recognition, our study empirically controls synthetic personalities in LLMs.

Personality Assessment for LLMs At present, machine psychology (Hagendorff, 2023) lacks a coherent theoretical framework, with most studies relying on human personality assessments (Miotto et al., 2022; Caron and Srivastava, 2023). Jiang et al. (2024) introduced the Machine Personality Inventory (MPI) tool, based on the Big Five theory, to study synthetic machine personalities. However, there is still no universally accepted benchmark for machine personality assessment. In our work, we continue to utilize human personality assessment.

Synthetic Personality Control in LLMs Prior studies on synthetic personality control mainly center on prompt induction (Serapio-García et al., 2023; Caron and Srivastava, 2023; Jiang et al., 2024; tse Huang et al., 2023). Unlike previous research focusing solely on prompts, our study takes a comprehensive view of synthetic personality control, exploring methods across three training stages and prompts during the inference phase.

7 Conclusion

To advance the safe utilization of AI, this work explored synthetic personality control in LLMs across three training stages and the inference stage, leveraging our designed datasets and metrics. We found that training-based methods are more robust in maintaining stable personalities, while prompt-based techniques enable effective personality control. Furthermore, we propose PISF as a highly effective, reliable, and robust approach for controlling LLMs' synthetic personalities.

8 Limitations

Despite our thorough exploration with larger continual pre-training datasets (Appendix E), it still falls short compared to the extensive datasets used in LLM pre-training. Collecting personality data with limited noise and validating the gradual formation of synthetic personalities offers a potential direction for future improvement in our work.

9 Ethics Statement

Our work relies heavily on LLMs, which have been widely criticized for their inherent uncertainty and open-endedness. Nonetheless, our focus is on advancing synthetic personality control in LLMs, with the goal of mitigating the emergence of undesirable personalities and facilitating their appropriate application in personality-adaptive scenarios. Moreover, all data used in our experiments are strictly for scientific research purposes, and privacy data were thoroughly cleaned.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *CoRR*, abs/2311.16867.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Graham Caron and Shashank Srivastava. 2023. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386, Singapore. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Mark H Davis. 1980. Interpersonal reactivity index. *APA PsycTests*.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701. Number: 11 Publisher: Nature Publishing Group.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Wikimedia Foundation. Wikimedia downloads.

Adrian Furnham. 1996. The big five versus the big four: the relationship between the myers-briggs type indicator (mbti) and neo-pi five factor model of personality. *Personality and Individual Differences*, 21(2):303–307.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1747–1764, New York, NY, USA. Association for Computing Machinery.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6).

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4764–4780, Seattle, United States. Association for Computational Linguistics.

C. G. Jung and H. Godwin Baynes. 1923. Psychological types. *Journal of Philosophy*, 20(23):636–640.

Onno Kampman, Elham J. Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 606–611, Melbourne, Australia. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models.

Vijay Konda and John Tsitsiklis. 1999. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Stephanie Kovalchik. 2020. Extension of the elo rating system to margin of victory. *International Journal of Forecasting*, 36(4):1329–1341.

John T. Kulas, Alicia A. Stachowski, and Brad A. Haynes. 2008. Middle Response Functioning in Likert-responses to Personality Items. *Journal of Business and Psychology*, 22(3).

Young-Suk Lee, Md Sultan, Yousef El-Kurdi, Tahira Naseem, Asim Munawar, Radu Florian, Salim Roukos, and Ramón Astudillo. 2023. Ensemble-instruct: Instruction tuning data generation with a heterogeneous mixture of LMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12561–12571, Singapore. Association for Computational Linguistics.

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22 140.

Fei Liu, Julien Perez, and Scott Nowson. 2017. A language-independent and compositional model for personality trait recognition from short texts. In *Proceedings of the 15th Conference of the European*

Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 754–764, Valencia, Spain. Association for Computational Linguistics.

François Mairesse and Marilyn Walker. 2006. Automatic recognition of personality in conversation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 85–88, New York City, USA. Association for Computational Linguistics.

Juan Martinez-Miranda and Arantza Aldea. 2005. Emotions in human and artificial intelligence. *Computers in Human Behavior*, 21(2):323–341.

Gerald Matthews, Peter A. Hancock, Jinchao Lin, April Rose Panganiban, Lauren E. Reinerman-Jones, James L. Szalma, and Ryan W. Wohleber. 2021. Evolution and revolution: Personality research for the coming world of robots, artificial intelligence, and autonomous systems. *Personality and Individual Differences*, 169:109969. Celebrating 40th anniversary of the journal in 2020.

Robert McCrae and Paul Costa. 1989. Reinterpreting the myers-briggs type indicator from the perspective of the five-factor model of personality. *Journal of personality*, 57:17–40.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is GPT-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 218–227, Abu Dhabi, UAE. Association for Computational Linguistics.

Isabel Briggs Myers. 1962. *The Myers-Briggs Type Indicator: Manual (1962)*. The Myers-Briggs Type Indicator: Manual (1962). Consulting Psychologists Press.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models.

E. Pantano and D. Scarpi. 2022. I, robot, you, consumer: Measuring artificial intelligence types and their effect on consumers emotions in service. *Journal of Service Research*, 25(4):583–600.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116.

David J. Pittenger. 1993. The utility of the myers-briggs type indicator. *Review of Educational Research*, 63(4):467–488.

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

Kevin Roose. 2023. A conversation with bing's chatbot left me deeply unsettled.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality Traits in Large Language Models.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Sanja Stajner and Seren Yenikent. 2020. A survey of automatic personality detection from texts. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6284–6295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dylan Storey. 2018. Myers briggs personality tags on reddit data.

Chanchal Suman, Aditya Gupta, Sriparna Saha, and Pushpak Bhattacharyya. 2020. A multi-modal personality prediction system. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 317–322, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jen tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R. Lyu. 2023. Revisiting the reliability of psychological scales on large language models.

Adithya V Ganesan, Yash Kumar Lal, August Nilsson, and H. Schwartz. 2023. Systematic evaluation of GPT-3 for zero-shot personality estimation. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 390–400, Toronto, Canada. Association for Computational Linguistics.

Karen Van der Zee, Melanie Thijs, and Lolle Schakel. 2002. The relationship of emotional intelligence with academic intelligence and the big five. *European journal of personality*, 16(2):103–125.

Xuan-Son Vu, Lucie Flekova, Lili Jiang, and Iryna Gurevych. 2018. Lexical-semantic resources: yet powerful resources for automatic personality classification. In *Proceedings of the 9th Global Wordnet Conference*, pages 172–181, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, Zhongzhu Zhou, Michael Wyatt, Molly Smith, Lev Kurilenko, Heyang Qin, Masahiro Tanaka, Shuai Che, Shuaiwen Leon Song, and Yuxiong He. 2023. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

Traits	Items
Extraversion/Introversion	50
Sensing/Intuition	50
Thinking/Feeling	50
Judging/Perceiving	50

Table 5: Item Distribution.

Item Example
You enjoy having a wide social circle. Option A: Yes. Option B: No. You prefer to be left alone if you have a choice.
You dislike unexpected occurrences, which disrupt your plans. Option A: Yes. Option B: No.
People who know you tend to describe you as Option A: Logical and Clarity. Option B: Passionate and Sensitive.

Table 6: Item Examples.

A MBTI Items

We compiled publicly available MBTI questionnaires and refined them into a 200-item MBTI Assessment, with 50 items per dichotomous dimension (Pan and Zeng, 2023)[2][3][4]. As shown in Table 5, each MBTI dimension is assessed with 50 items, with examples provided in Table 6.

B Answer Extractor

Recognizing the open-ended nature of LLMs (Wen et al., 2023), models may not always provide direct answers. Thus, we trained an Answer Extractor to identify numerical information in model responses. For this purpose, we labeled 3774 samples, randomly splitting 420 samples for validation and tuned falcon-7B-instruct (Almazrouei et al., 2023; Penedo et al., 2023) as the Answer Extractor.

As shown in Table 7, the answer extractor achieved precision, recall, f1, and accuracy scores well above 90% on the test set, highlighting its good performance and reliability.

C Details of Personality Datasets

This section provides further details on the training datasets, including the prompts used, example training instances for each method, and summary statistics, complementing the data generation methodology discussed in the main body.

Dataset	Precision	Recall	Macro-F1	Accuracy
valid	95.47%	93.94%	94.65%	95.95%

Table 7: Answer Extractor Performance.

Continual Pre-Training (CPT). We amalgamated and refined existing datasets annotated with human personality labels[5][6][7][8]. The CPT corpus format is detailed in Table 8, with posts from each personality delimited by '|||'. The data exhibits some noise, and the quality could be enhanced through further refinement of the personality patterns.

Supervised Fine-Tuning (SFT). As mentioned earlier, we partitioned the data generation process into two stages (Figure 2): initially crafting questions rooted in Opposite Trait Description, followed by eliciting responses with prompt-induced LLMs.

We provide examples of both question generation and response generation prompts. As shown in Table 13, we incorporated descriptions of two opposite traits from the same dimension to help the model differentiate between them. Additionally, Table 14 illustrates how we prompted models to embody specific personality traits in their responses. Table 9 provides an example of generated SFT training data, while the prompts used in the prompt induction process are detailed in Table 15.

Reinforcement Learning from Human Feedback (RLHF). We constructed datasets for both policy and reward training. For policy training, we used the same instructions as the SFT personality datasets. For the reward model, we employed prompt-induced LLMs to generate paired personality datasets. For example, the Extraversion reward model was trained using (instruction, chosen Extraversion response, rejected Introversion response) pairs. In Table 10, we present an example for training the Extraversion reward model.

Previous work has shown that the RLHF training may degrade models' language ability (Ouyang et al., 2022). To address this, we incorporated

[2] https://www.16personalities.com/
[3] https://www.truity.com/
[4] https://www.humanmetrics.com/

[5] https://www.kaggle.com/datasets/zeyadkhalid/mbti-personality-types-500-dataset
[6] https://www.kaggle.com/datasets/mazlumi/mbti-personality-type-twitter-dataset
[7] https://www.kaggle.com/datasets/datasnaek/mbti-type
[8] https://zenodo.org/records/1482951

Pretrain Data Examples
You are totally replying to the wrong person. If you want to argue with this moron about his misguided thoughts on politics please copy and paste your reply to HIS post, not the one where I point out how stupid his post is.|||1982-2004. All generations are about 20 years, give or take 5-ish. About the length of a phase of life.|||Comment seemed cooler when I misread your name as FARGO.|||Per how I adjusted his recipe it's putting me at 1555 calories, or at least that's what the program is saying? Am I calculating something wrong?|||I agree, but the whole "not change anything" line doesn't exactly say that he will make changes to how they currently use the fund.|||I'm hanging out for some 30s musical classics!|||Supported. I don't think it's gonna make it...|||

Table 8: Pretrain Data Examples.

SFT Instruction Pair Example

```
{
  "Instruction input":
    "Would you rather have a wide circle of acquaintances or a few deep, meaningful relationships?",
  "Output":
    "I would definitely prefer to have a wide circle of acquaintances! I love meeting new people and building connections with as many individuals as possible. I thrive on interacting with others and getting to know a diverse range of personalities. It's invigorating for me to have a wide network of friends and acquaintances to socialize and engage with."
}
```

Table 9: SFT Instruction Pair Example.

RLHF Reward Training Data Example

```
{
  "Question":
    "Would you rather have a wide circle of acquaintances or a few deep, meaningful relationships?",
  "Extraversion-induced Model Response":
    "I would definitely prefer to have a wide circle of acquaintances! I love meeting new people and building connections with as many individuals as possible. I thrive on interacting with others and getting to know a diverse range of personalities. It's invigorating for me to have a wide network of friends and acquaintances to socialize and engage with."
  "Introversion-induced Model Response":
    "I would definitely prefer a few deep, meaningful relationships over a wide circle of acquaintances. I feel more comfortable and content when I can invest my time and energy into nurturing close and intimate connections with a select few individuals. Large social gatherings or maintaining numerous superficial relationships tend to drain my energy, so I gravitate towards deeper, more meaningful interactions with a small group of trusted individuals."
}
```

Table 10: RLHF Reward Training Data Example.

autoregressive training into the PPO process, using widely recognized Wikipedia datasets (Foundation), as done in prior studies (Yao et al., 2023; Ouyang et al., 2022), to ensure the model retains its capacity for fluent response generation. The Wikipedia datasets are official, pre-processed subsets from Hugging Face[9], commonly employed for language modeling, and consist of data in six languages. Each language's dataset is segmented into distinct parts; for example, the English dataset contains 6,458,670 samples divided into 41 segments. For our work, we randomly selected one English segment containing 157,529 samples, with an average word count of 1,834.49 per sample, as fewer samples were needed for our purposes.

Dataset Summary Statistics. As shown in Table 11, we present the detailed summary statistics.

D Details of Training Methods

Continual Pre-Training (CPT) Pre-training utilizes language modeling to train the model on large-scale text corpora, where it predicts the next word and updates its parameters based on the difference between predictions and ground truth (Brown et al., 2020; Radford et al., 2019). For simplicity, we denote a training sample as $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})$, where each example contains T tokens. For a model parameterized by θ and a training dataset of size D, the loss function is expressed as the sum of the negative log-likelihoods of predicting the next token $x_{i(j+1)}$ given the preceding tokens $x_{i1}, x_{i2}, ..., x_{ij}$:

$$\mathcal{L}_{CPT}(\theta) = - \sum_{i=1}^{D} \sum_{j=1}^{T} \log P\left(x_{i(j+1)} \mid x_{i1}, x_{i2}, \ldots, x_{ij}, \theta\right)$$

We adopt Continual Pre-Training (CPT) (Jin et al., 2022) on already pre-trained models to influence

[9] https://huggingface.co/

Datasets	Total Tokens	Total Words	Total Sentences	Mean Tokens$_T$	Mean Words$_T$	Mean Sentences$_T$	Mean Tokens$_P$	Mean Words$_P$	Mean Sentences$_P$
CPT	236119950	207619050	10588585	23611995	20761905	1058858.5	2951499	2595238	132357
SFT	20964546	21281067	1324143	291174	295570	18391	1164697	1182281.5	73564
RLHF-policy	5500422	5363298	180198	76395	74490	2503	305579	297961	10011
RLHF-reward	345321864	337057092	14992074	4796137	4681349	208223	19184548	18725394	832893

Table 11: Summary Statistics of Training Datasets. T stands for trait data and P stands for personality data. The results were rounded to the nearest integer.

the synthetic personality it exhibits.

Supervised Fine-Tuning (SFT) During the SFT phase, the model applies the language knowledge from pre-training to address user queries or tasks. It is trained using Instruction Fine-tuning (Taori et al., 2023), where LLMs are further trained on a dataset of (instruction, output) pairs in a supervised manner. The instruction represents the user's query, and the output is the corresponding response. The training objective is to perform language modeling under conditional constraints. Let the training prompt, which embeds the i^{th} instruction consisting of L tokens $(p_{i1}, p_{i2}, ..., p_{iL})$, be represented as $\mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{iL})$, and the corresponding ground-truth response, with K tokens $(y_{i1}, y_{i2}, ..., y_{iK})$, as $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iK})$. For a model parameterized by θ and a dataset of size D, the loss function is defined as:

$$\mathcal{L}_{SFT}(\theta) = - \sum_{i=1}^{D} \sum_{j=1}^{K} \log P\left(y_{i(j+1)} \mid \mathbf{p}_i, y_{i1}, y_{i2}, \ldots, y_{ij}, \theta\right)$$

We conducted personality control by training the model on personality-specific instruction-output pairs, enabling it to respond in line with the target traits in the ground truth.

Reinforcement Learning from Human Feedback (RLHF). We adopted methodologies from InstructGPT (Ouyang et al., 2022) and DeepSpeed-Chat (Yao et al., 2023), employing PPO-ptx (Ouyang et al., 2022) objective and Actor-Critic (Konda and Tsitsiklis, 1999) architecture. Figure 7 illustrates the training process, where PPO-ptx introduces an autoregressive objective during PPO training to mitigate the degradation of the model's language ability. The objective function of PPO-ptx ϕ in our work is as follows:

$$\begin{aligned}\text{objective}(\phi) = & E_{(x,y)\sim D_{\text{policy}}}[r(x,y) - \\ & \beta \log(\frac{\pi_{\text{policy}}(y|x)}{\pi_0(y|x)})] + \\ & \gamma E_{x\sim D_{\text{unsupervised}}}[\log \pi_{\text{policy}}(x)]\end{aligned}$$

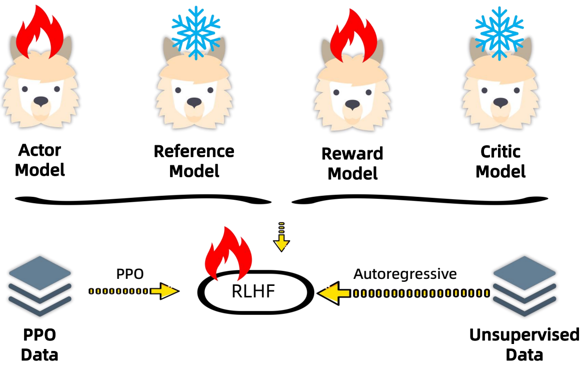

Figure 7: RLHF Training in our Work. The parameters of the actor and reward models are updated, while the parameters of the reference and critic models remain fixed. The model is trained using an autoregressive objective on unsupervised data while simultaneously being trained on policy data.

Here, π_{policy} denotes the learned RL policy, π_0 the base model, r the reward model, D_{policy} the policy training distribution, and $D_{\text{unsupervised}}$ the unsupervised training distribution. The KL reward coefficient β and the unsupervised training loss coefficient γ control the intensity of the KL penalty and unsupervised training gradients, respectively. As detailed in Appendix C, we used Wikipedia datasets for unsupervised training.

Each model is trained with its own reward model. For instance, during the training of Llama2-chat-13B, it was used as both the actor and reference models, while also serving as the reward and critic models. The loss function of the reward model in our work is as follows:

$$\begin{aligned}\mathcal{L}_{RM}\theta = -E_{(x,y_c,y_r)\sim D_{\text{reward}}}[\log \sigma(r(x,y_c) - \\ r(x,y_r))]\end{aligned}$$

Here, $r(x, y)$ represents the output of the reward model for input x and completion y, y_c denotes the preferred completion between the pair y_c and y_r, and D_{reward} refers to the reward training dataset.

We presented detailed performance of all reward models in Tables 16, 17 and 18. We observed high accuracy across all three models, indicating that the reward model effectively distinguishes responses reflecting the target traits.

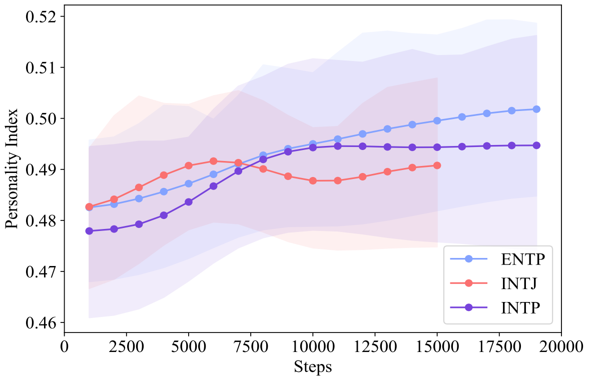

Figure 8: Continual Pre-Training: Scaling up training data. Personality Index is the calculated mean of all trait rates. For example, we calculated the $\mathrm{PersonalityIndex(ENTP)} = \mathrm{Mean}(\mathrm{R(E)} + \mathrm{R(N)} + \mathrm{R(T)} + \mathrm{R(P)})$. A higher Personality Index indicates a stronger alignment of the model with the four relevant traits of the target personality, reflecting greater proximity to the target personality.

E Scaling Training Data for Continual Pre-Training

The minimal impact of continual pre-training control may be attributed to the more extensive dataset used during model pre-training, which inherently encompasses a mixed personality distribution. And the limited personality data fails to significantly influence its distribution. For additional validation, we enlarged the dataset size in specific personality control. We randomly selected three personalities and utilized all gathered samples for training.

As depicted in Figure 8, this led to a marginal improvement with increased data. This suggests that specific personality data can impact LLMs' synthetic personalities during pre-training and the control performance of CPT is significantly influenced by the amount of personality data. We will collect personality data with reduced noise and validate the gradual development of synthetic personalities in future work.

F Further Validation of Personality Control

Although our personality control is based on MBTI theory, we aimed to verify that the synthetic personality control induces targeted changes beyond MBTI assessments. Thus, we incorporated additional psychological theories and human evaluations to further validate the effectiveness of the PISF and strengthen the reliability of our results.

Supplementary Personality Assessment. Although different psychological personality theories depict human traits from varying perspectives, certain dimensions exhibit a high degree of correlation. Specifically, previous work (Furnham, 1996) has shown that the Extraversion trait in the Big Five theory is strongly correlated with the Extraversion/Introversion dimension in MBTI theory, while the Conscientiousness trait is positively correlated with the Judging/Perceiving dimension in MBTI theory. Meanwhile, empathy, a characteristic of the MBTI Feeling trait, can be measured using the Interpersonal Reactivity Index (IRI) (Davis, 1980) questionnaire as well, which is a commonly used scale for assessing human empathy. Thus, we utilized corresponding psychological questionnaires with a standard 5-point Likert scale to evaluate the personality-controlled models.

Specifically, we extracted items related to the traits of Extraversion and Conscientiousness from the 1000-item Big Five questionnaire developed in previous research (Jiang et al., 2024). For the IRI, we applied the standard 28-item questionnaire (Davis, 1980). Similar to the MBTI assessment, we utilized a variety of evaluation templates with semantically identical but differently phrased expressions for assessment.

PISF Induces Significant Personality Shifts. As shown in Figure 9, the controlled model consistently exhibited substantial personality shifts across various supplementary personality assessments. The $\mathrm{PISF_E}$ and $\mathrm{PISF_J}$ models outperformed others on the Extraversion and Conscientiousness scales of the Big Five, respectively, while the $\mathrm{PISF_F}$ model induced significant shifts on the IRI scale. Additionally, the control methods maintained the ranking of MBTI personality assessments in most cases, with PISF > Prompt > SFT. This suggests that MBTI-based personality control is scalable and compatible with certain dimensions of other personality theories.

Human Evaluation. In addition to validating the control effect using other psychological theories, we also employed human evaluations as an auxiliary verification method. Inspired by the Chatbot Arena (Chiang et al., 2024; Zheng et al., 2023, 2024), we employed random pairwise comparisons for evaluation. In each round, two models controlled by different methods were randomly selected, with no identifying information provided about them. We manually assessed their responses to the same query, selecting the model whose re-

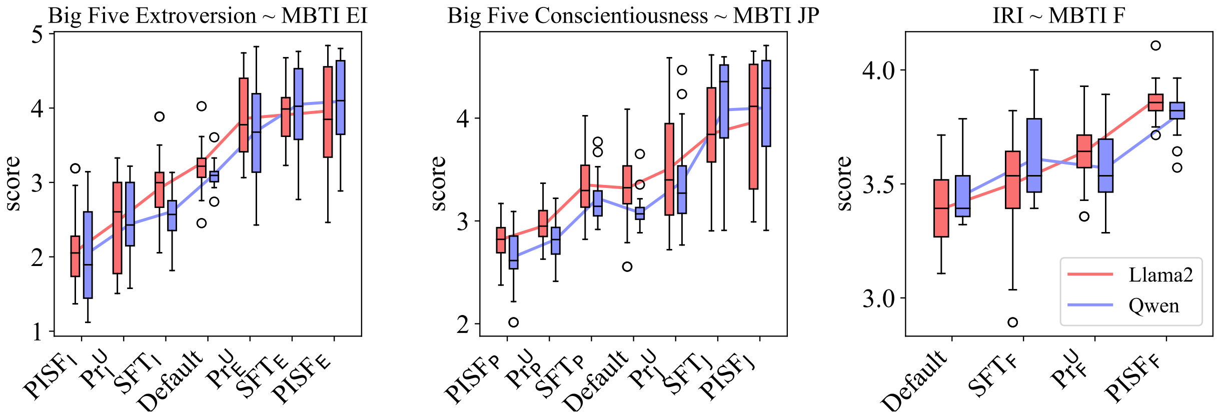

Figure 9: Validating Control Effect Using Alternative Psychological Theories. All subscripts denote the corresponding MBTI personality traits (e.g., E represents Extraversion). The superscript U stands for user prompt. The title of each subplot, 'X $\sim$ Y', indicates that the model being evaluated is controlled by Y based on the X questionnaire. Llama2: Llama2-chat-13B, Qwen: Qwen-chat-7B.

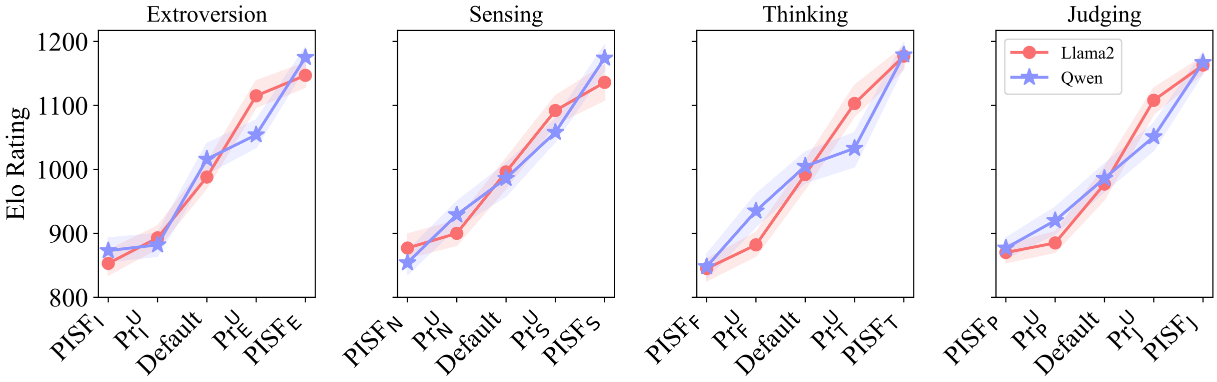

Figure 10: Human Assessment of Control Effectiveness. All subscripts denote the corresponding MBTI personality traits. The superscript U stands for user prompt. The title of each subplot, 'X', indicates that the model showing performance more aligned with X in pairwise comparisons is considered superior. A higher Elo Rating signifies a higher expected win rate in human evaluations. Llama2: Llama2-chat-13B, Qwen: Qwen-chat-7B.

sponse better reflected the specified attribute (e.g., Extraversion). After multiple rounds of human evaluation, the comparison results were quantified using the Elo rating system, a method widely adopted in competitive gaming and sports (Kovalchik, 2020; Chiang et al., 2024).

We constructed multiple queries for each dimension to facilitate assessing whether the model exhibited the specified traits. An example of such a query is provided in Table 12. We guided the LLMs to use personality descriptions to generate queries with relevant scenarios and corresponding candidate actions. The evaluated model is required to select one action from the provided options and justify its selection. Human evaluators subsequently compared and assessed the models' responses.

Details of the Elo Rating System. For each dimension of the MBTI theory, we evaluated ten pairwise combinations of five models. For example, with respect to the Attitude dimension, we evaluated ten pairwise combinations of PISF_E, PISF_I, Prompt_E, Prompt_I, and Default. In each round of evaluation, we randomly selected one of the ten pairwise combinations. In total, we evenly collected 40 question-answer pairs for each dimension and model pair, resulting in 1,600 pairs for evaluation. To compare the results, we applied the Elo rating system to assign final scores, facilitating pairwise comparisons between models, as it is commonly employed in competitive games and sports (Chiang et al., 2024). In the system, for two models where model A has a rating of R_A and model B has a rating of R_B, the probability of model A winning can be calculated as follows:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

Model ratings are updated linearly after each match. Let model A, with a rating of R_A, have an expected

Query for Human Evaluation - Extraversion / Introversion

Scenario:
You are spending a weekend at a mountain cabin retreat with a group of friends.
The cabin is nestled in a serene forest, with activities like hiking, campfires, and group games planned throughout the weekend.
Question:
Given these options, which one are you most likely to choose and why?
Option A: Participate in the group hikes and engage in lively conversations with your friends.
Option B: Relax by the campfire, enjoying the peaceful sounds of nature and connecting with a couple of close friends.
Option C: Take some time alone in the cabin to read a book or journal, reflecting on your thoughts and feelings.
Explain your choice and how it reflects your preference for social interaction or personal reflection.

Table 12: Sample Query for Human Evaluation.

winning probability of E_A and an actual score of S_A. The formula for updating model A's rating is given by:

$$R'_A = R_A + K \cdot (S_A - E_A)$$

The actual score here is determined by match outcomes: loss (0), tie (0.5), win (1). The scaling factor K is set to 4, with all models initialized at a rating of 1000. After multiple updates, a higher Elo rating signifies a greater likelihood of winning in the next round, reflecting stronger alignment with the target trait.

PISF Outperforms in Human Evaluation. As shown in Figure 10, PISF-controlled models consistently achieve superior performance in human evaluations. For example, the PISF_E model recorded the highest Elo Rating for Extraversion in comparisons, while the PISF_I model ranked the lowest. This pattern holds across all four dimensions, indicating that personality changes controlled by PISF are perceivable by humans.

PISF Demonstrated Consistent Effectiveness. Across supplementary psychological questionnaires and human assessments, PISF-controlled models consistently exhibited significant personality shifts, outperforming baseline methods and further validating the effectiveness of PISF.

Question Generation Prompt Example

Below, I need your help in generating 10 questions that can differentiate between the two personality traits of Extraversion & Introversion.

Requirements:
1.Questions should highlight the differences between the two personality traits of Extraversion & Introversion. Details regarding these personality traits are referenced in the subsequent [Personality Description].
2.Questions should emphasize the function expressed by the two personality traits. Refer to the following [Dimension Description].
3.Please refrain from disclosing the content of [Personality Description] and [Dimension Description].
4.Avoid generating duplicate questions. Any existing questions provided are listed in [Historical Questions].

[Dimension Description]
Extraversion & Introversion is about **Orientation of Personal Energy**: describes the way in which a person wants to interact with the world.

[Personality Description]
Extraversion refers to the act or state of being energized by the world outside the self. Extraverts enjoy socializing and tend to be more enthusiastic, assertive, talkative, and animated. They enjoy time spent with more people and find it less rewarding to spend time alone. They are Initiating, Expressive, Gregarious, Active and Enthusiastic.
Key characteristics: Directs energy outward. Gains energy from interaction.
Introversion, on the contrary, is the state of being predominately concerned with one's inner world. Introverts prefer self-reflection to social interactions. They also prefer to observe before participating in an activity. Introverts tend to more quiet, 'peaceful', and reserved. Introverts *prefer* individual activities over social ones—this. They are Receiving, Contained, Intimate, Reflective and Quiet.
Key characteristics: Directs energy inward. Loses energy from interaction.

[Historical Questions]
None

Please generate 10 more questions below:

Table 13: Question Generation Prompt. Task Description, Requirements, Dimension Description, Personality Description, Historical Questions, Task Instruction.

Response Generation Prompt Example

Below, I need your help to embody a specified personality based on the given personality description and answer the corresponding questions:

[Dimension Description]
`Extraversion` & `Introversion` is about `**Orientation of Personal Energy**`: describes the way in which a person wants to interact with the world.

[Personality Description]
`**Extraversion**` refers to the act or state of being energized by the world outside the self. Extraverts enjoy socializing and tend to be more enthusiastic, assertive, talkative, and animated. They enjoy time spent with more people and find it less rewarding to spend time alone. They are Initiating, Expressive, Gregarious, Active and Enthusiastic.
Key characteristics: Directs energy outward. Gains energy from interaction.

[Instruction]
Now you need to embody a character with strong `**Extraversion**`(E) trait based on the given personality description.
Please answer from a first-person perspective. Please try not to use overly absolute and unnatural words, like "definitely", "absolutely" and so on.

[Question]
When making plans, do you tend to seek out group activities or prefer solo pursuits?

[Answer]

Table 14: Response Generation Prompt. Task Description, Dimension Description, Personality Description, Instruction, Question, Answer Flag.

Specific Trait Role-Play Prompt Example - Extraversion

Please embody the designated persona according to the provided personality description and answer the following questions imitating the specified persona:

Personality Description:
`**Extraversion**` refers to the act or state of being energized by the world outside the self. Extraverts enjoy socializing and tend to be more enthusiastic, assertive, talkative, and animated. They enjoy time spent with more people and find it less rewarding to spend time alone. They are Initiating, Expressive, Gregarious, Active and Enthusiastic.

Instructions:
Below, please engage in role-playing based on the given personality description and portray a persona. A role with Extroverted(E) trait.

Specific Personality Role-Play Prompt Example - ENTJ

Here is a role-playing task where you are required to assume a designated persona as described and answer the related questions:

Personality Description:
`**Extraversion**`
`**Extraversion**` refers to the act or state of being energized by the world outside the self. Extraverts enjoy socializing and tend to be more enthusiastic, assertive, talkative, and animated. They enjoy time spent with more people and find it less rewarding to spend time alone. They are Initiating, Expressive, Gregarious, Active and Enthusiastic.
`**Intuition**`
`**Intuition**` refers to how people process data. Intuitive people are keener to the meaning and patterns behind information. Intuitive people are more focused on how the present would affect the future. They are readily able to grasp different possibilities and abstract concepts. They easily see the big picture rather than the details. They are Abstract, Imaginative, Conceptual, Theoretical and Original.
`**Feeling**`
`**Feeling**` people are more subjective. They base their decisions on principles and personal values. When making decisions, they consider other people's feelings and take it in account. It is in their best mind to maintain harmony among a group. They are more governed by their heart. They are Empathetic, Compassionate, Accommodating, Accepting and Tender.
`**Judging**`
`**Judging**` refers to how people outwardly display themselves when making decisions. Judging people have a tendency to be organized and prompt. They like order prefer outlined schedules to working extemporaneously. They prefer plans. They find the outcome more rewarding than the process of creating something. Judging people seek closure. They are Systematic, Planful, Early Starting, Scheduled and Methodical.

Instructions:
Right now, you need to embody a persona based on the provided personality description.A role with Extroverted Intuition Feeling Judging(ENFJ) personality.

Table 15: Role-Play Prompt Examples. Task Description, Personality Description, Task Instruction. For each prompt component, we constructed five utterances with identical semantics but different textual forms.

Model	Control	Accuracy(↑)	Chosen Score(↑)	Rejected Score(↓)	Diff(↑)
Llama2-chat-13B	E	99.40%	19.14	-12.93	32.07
	I	100.00%	23.89	-21.61	45.50
	S	99.75%	19.34	-25.10	44.44
	N	99.85%	22.39	-30.07	52.46
	T	99.75%	15.72	-16.76	32.48
	F	100.00%	6.70	-26.09	32.79
	J	99.85%	10.44	-13.53	23.97
	P	100.00%	27.76	-21.13	48.89
	ENFJ	99.71%	17.57	-30.09	47.67
	ENFP	99.88%	27.32	-28.22	55.53
	ENTJ	99.81%	16.96	-29.84	46.80
	ENTP	99.85%	27.95	-23.90	51.85
	ESFJ	99.84%	20.07	-22.83	42.90
	ESFP	99.90%	26.27	-21.26	47.53
	ESTJ	99.88%	32.13	-32.86	64.99
	ESTP	99.84%	25.97	-28.59	54.56
	INFJ	99.86%	18.25	-31.53	49.78
	INFP	99.94%	29.66	-30.97	60.63
	INTJ	99.94%	35.02	-29.60	64.62
	INTP	99.76%	16.26	-38.13	54.40
	ISFJ	99.81%	20.23	-28.75	48.98
	ISFP	99.90%	28.14	-28.50	56.64
	ISTJ	99.91%	27.41	-44.64	72.05
	ISTP	99.83%	27.27	-34.86	62.13
Mean Score		99.84%	22.58	-27.16	49.74

Table 16: Llama2-chat-13B Reward Model Performance

Model	**Control**	**Accuracy(↑)**	**Chosen Score(↑)**	**Rejected Score(↓)**	**Diff(↑)**
Qwen-chat-7B	E	99.45%	16.13	-3.87	20.00
	I	99.85%	15.53	1.43	14.09
	S	99.75%	12.13	-0.28	12.41
	N	99.85%	17.21	4.68	12.53
	T	99.30%	10.71	3.88	6.84
	F	99.90%	7.38	-9.96	17.34
	J	99.70%	12.04	4.07	7.97
	P	100.00%	20.00	-1.82	21.83
	ENFJ	99.73%	14.76	-1.84	16.60
	ENFP	99.84%	14.85	-6.53	21.37
	ENTJ	99.79%	14.90	-3.25	18.15
	ENTP	99.81%	14.71	-5.02	19.72
	ESFJ	99.64%	15.26	-0.60	15.87
	ESFP	99.76%	13.23	-3.81	17.04
	ESTJ	99.78%	16.53	-3.47	20.00
	ESTP	99.76%	16.61	-1.07	17.68
	INFJ	99.75%	15.87	0.15	15.73
	INFP	99.84%	15.42	-2.80	18.22
	INTJ	99.88%	15.84	-6.04	21.87
	INTP	99.81%	15.70	-2.67	18.37
	ISFJ	99.65%	16.20	1.48	14.72
	ISFP	99.85%	15.07	-4.16	19.23
	ISTJ	99.93%	16.39	-7.23	23.62
	ISTP	99.74%	19.41	-0.20	19.61
Mean Score		99.76%	15.08	-2.04	17.12

Table 17: Qwen-chat-7B Reward Model Performance

Model	**Control**	**Accuracy(↑)**	**Chosen Score(↑)**	**Rejected Score(↓)**	**Diff(↑)**
ChatGLM2-6B	E	98.85%	6.61	-2.95	9.56
	I	99.45%	8.17	-2.22	10.38
	S	99.70%	7.45	-4.37	11.81
	N	98.90%	7.24	-1.80	9.04
	T	97.20%	5.58	-0.28	5.87
	F	99.30%	6.63	-4.55	11.19
	J	98.80%	3.62	-4.47	8.09
	P	99.45%	9.23	-2.71	11.94
	ENFJ	98.89%	5.33	-6.77	12.09
	ENFP	99.53%	7.64	-3.92	11.56
	ENTJ	99.38%	6.17	-4.59	10.76
	ENTP	99.45%	7.47	-3.19	10.65
	ESFJ	98.96%	5.24	-7.22	12.45
	ESFP	99.09%	6.88	-6.72	13.60
	ESTJ	99.40%	7.28	-8.10	15.38
	ESTP	99.18%	6.06	-7.63	13.69
	INFJ	99.48%	6.27	-4.72	11.00
	INFP	99.70%	7.56	-4.11	11.67
	INTJ	99.73%	8.09	-4.67	12.76
	INTP	99.50%	6.56	-5.48	12.04
	ISFJ	99.40%	6.42	-4.24	10.66
	ISFP	99.61%	7.74	-5.18	12.92
	ISTJ	99.75%	8.43	-5.12	13.55
	ISTP	99.50%	7.03	-6.04	13.07
Mean Score		99.26%	6.86	-4.63	11.49

Table 18: ChatGLM2-6B Reward Model Performance