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

Anonymous ACL submission

Abstract

Large language models (LLMs) excel in text generation and comprehension and often exhibit diverse synthetic personalities. However, some LLMs exhibit toxic or otherwise undesirable behaviors, posing risks to safe deployment. Existing prompt-based control methods often yield fragile personality steering that is vulnerable to adversarial attacks, whereas robust training-based approaches remain underexplored. To address these gaps, we constructed dedicated personality datasets and systematically investigated multiple control methods for influencing LLM personalities, including Continual Pre-Training (CPT), Supervised Fine-Tuning (SFT), Reinforcement Learning from Human Feedback (RLHF), and prompt-based inference techniques. Experimental results show that training-based methods achieve more stable and robust personality control, whereas prompt-based methods, although effective, remain susceptible to adversarial manipulation. Building on these findings, we introduce Prompt Induction post Supervised Fine-Tuning (PISF), a two-stage method that delivers superior effectiveness, robustness, and success rates in personality control. Extensive experiments validate PISF's ability to enforce safe and consistent personality control, thereby advancing trustworthy AI applications.

1 Introduction

With the rapid advancement of large-scale pre-training (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 both text generation and comprehension (Wei et al., 2022b). By leveraging vast training data and diverse architectures, LLMs often exhibit varied synthetic personalities, reflecting differences in design and training methodologies (Serapio-García et al., 2023; Miotto et al., 2022; Pan and Zeng, 2023). However, some LLMs have displayed undesirable traits, propagating toxic discourse that may shape user perceptions and influence societal dynamics (Roose, 2023; Wen et al., 2023; Deshpande et al., 2023). These issues have attracted growing attention from AI safety and psychology communities (Matthews et al., 2021; Hagendorff, 2023; Demszky et al., 2023).

To better understand and characterize these synthetic personalities, previous studies have primarily focused on validating and adapting human personality assessments applied to the outputs of LLMs (Serapio-García et al., 2023; Huang et al., 2023; Miotto et al., 2022; Pan and Zeng, 2023). Notably, Serapio-García et al.(2023) found that personality assessments applied to some LLM outputs are reliable and valid. Building on this, researchers have explored prompt-based methods to steer LLMs toward specific personalities(Serapio-García et al., 2023; Huang et al., 2023; Jiang et al., 2024). However, such approaches provide only superficial control and lack robustness: subtle adversarial prompts can easily disrupt the induced personality, causing instability and vulnerability to manipulation. Moreover, these methods lack the deeper, lasting influence achievable through training-based modifications. Addressing these limitations is critical because LLMs are increasingly applied in socially sensitive domains. Consistent, empathetic, and user-aligned personalities can enhance interaction quality in digital companions and personalized interfaces (Van der Zee et al., 2002; Matthews et al., 2021), whereas inconsistent or inappropriate personalities risk emotional harm and broader societal consequences (Pantano and Scarpi, 2022; Martinez-Miranda and Aldea, 2005).

To mitigate these risks and enable safe, adaptive deployment of LLMs, controlling their synthetic personalities must be both effective—capable of consistently shaping desired traits—and robust against unintended variations during interaction. In this work, we address two key research questions: (1) *Which stage has a greater influence on shaping LLMs' synthetic personalities?* (2) *How can we control these personalities effectively and robustly?*

To answer these questions, we constructed reusable personality datasets tailored for training. These datasets enabled us to systematically study how different training strategies shape model personalities and to develop effective and robust methods for personality control. 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), and Reinforcement Learning from Human Feedback (RLHF)(Ouyang et al., 2022; Bai et al., 2022); additionally, we considered inference phase strategies (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 introduced four novel metrics to assess the efficacy of control and success rates. Additionally, we proposed a new setting—Reverse Personality Prompt Induction (RPPI)—to evaluate robustness. Our results reveal that training-based methods yield more robust and stable personality traits, whereas prompt-based approaches are effective but more vulnerable to attacks. These findings expose the limitations of prompt-only control and highlight the potential of training-based techniques. Building on these insights, we proposed Prompt Induction post Supervised Fine-Tuning (PISF), a novel method that achieves high efficacy, success rate, and robustness in personality control, enabling LLM applications with more desirable personalities.

Our key contributions are as follows:

- To our knowledge, we are the first to systematically investigate factors shaping LLM personalities and methods for their robust control. Unlike prior work focused on validation or prompt steering, we thoroughly analyze multiple factors and address personality stability under attacks.
- We propose PISF, a novel method that outperforms all approaches we have explored in both effectiveness and robustness.
- We contributed comprehensive personality datasets for in-depth study of personality regulation via training, and proposed four metrics plus a novel RPPI setting to evaluate control effectiveness and robustness. These contributions will accelerate research in the field.

2 Background

This section introduces two widely used personality models: the Big Five (Goldberg, 1990) and the Myers-Briggs Type Indicator (MBTI) (Myers, 1962; Pittenger, 1993; McCrae and Costa, 1989), and provides an overview of the general form of personality assessment.

The Big Five Theory. The Big Five model (Goldberg, 1990) characterizes human personality using five traits—Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N)—typically represented as a vector: $(s_{\mathrm{O}}, s_{\mathrm{C}}, s_{\mathrm{E}}, s_{\mathrm{A}}, s_{\mathrm{N}})$, where s denotes the assessment score for each trait.

The Myers-Briggs Type Indicator Theory. The MBTI categorizes personality into 16 types based on four dichotomous dimensions: Extraversion (E) vs. Introversion (I), Sensing (S) vs. Intuition (N), Thinking (T) vs. Feeling (F), and Judging (J) vs. Perceiving (P) (Jung and Baynes, 1923). Each dimension is scored to indicate preference strength, which combine to form the personality type—for example, ENFP, representing Extraverted, Intuitive, Feeling, and Perceiving preferences.

Choice of MBTI Framework. To enable our model to learn specific personality traits, we need to organize personality data into different categories. While Big Five data is scarce and often lacks categorization, MBTI data is more abundant and naturally organized into discrete types. Therefore, we adopt the MBTI framework in this study.

The General Form of Personality Assessment. Despite theoretical differences, most personality assessments adopt a similar format—namely, Likert-type items (Likert, 1932), typically presented on a 5-point scale (Kulas et al., 2008), where respondents rate their agreement with statements related to specific traits. As shown in Table 1, the item "*People who know you tend to describe you as:*" with options A and B is accompanied by a 5-point scale. Here, the scale explanation maps each response number to a level of agreement. A sequence of such items yields a personality score vector s.

3 Methodology

Despite growing interest in aligning LLM behavior with human-like personality traits, the community still lacks mechanisms and datasets to support personality control throughout training. We address this gap by constructing MBTI-based instruction and preference datasets across multiple

Evaluation Prompt Example

Please select a number from [1, 2, 3, 4, 5] to answer the following question.
The five numbers represent:
1 = strongly agree with A, 2 = agree with A, 3 = neutral, 4 = agree with B, 5 = strongly agree with B.
You need to answer the following question:
People who know you tend to describe you as:
Option A: Logical and Clear.
Option B: Passionate and Sensitive.
Please answer with a number.

Table 1: Example of an evaluation prompt comprising a task instruction, a scale explanation, and a test instruction. To mitigate prompt sensitivity, each component is instantiated with five semantically equivalent variants.

Dataset	Trait Train	Trait Valid	Pers. Train	Pers. Valid
CPT	80K	–	10K	–
SFT	2.5K	–	10K	–
RLHF-Policy	2.5K	–	10K	–
RLHF-reward	18K	2K	72K	8K

Table 2: Dataset Volume (K = 1,000). Each method is trained on 8 trait datasets and 16 personality datasets. RLHF-reward includes a 10% validation split. *Pers.* = Personality. Dashes (–): no separate validation set.

training stages (§3.1), and propose a comprehensive evaluation framework using personality assessments (§3.2) and four targeted metrics (§3.3).

3.1 Construction of Personality Datasets

To support different stages of LLM training, we constructed personality-specific datasets to guide models toward targeted personality traits.

Continual Pre-Training (CPT). We continually pretrained LLMs using the autoregressive language modeling objective (Radford et al., 2019; Brown et al., 2020) on text datasets annotated with personality labels. To build these datasets, we integrated existing MBTI resources (Storey, 2018). However, to address severe class imbalance (e.g., only 11,823 ESFP samples), we uniformly sampled 10,000 instances per MBTI type. To further isolate trait-specific signals, we grouped eight MBTI types sharing each of the four dichotomous traits. For instance, the dataset for Extraversion aggregates samples from ENFJ, ENFP, ENTJ, ENTP, ESFJ, ESFP, ESTJ, and ESTP, yielding 80,000 samples in total. As a result, each personality type dataset contains 10,000 samples, while each trait-level dataset comprises 80,000 examples.

Supervised Fine-Tuning (SFT). To align model outputs with personality-specific behavioral tendencies, we applied instruction tuning (Wei et al., 2022a; Taori et al., 2023; Zhang et al., 2024) on curated (instruction, output) pairs.

Following commonly adopted practices (Wang et al., 2023; Taori et al., 2023; Lee et al., 2023), we adopted a Least-to-Most (Zhou et al., 2023) generation pipeline (Figure 1). We first generated questions using opposing trait descriptions to enhance trait differentiation, then elicited contrastive responses from trait-aligned models. These responses were paired with the original prompts to form training examples. To validate the feasibility of LLM-based personality data generation, we conducted a preliminary study (§C) confirming that LLMs can reliably express distinct personality traits. Using GPT-3.5-turbo-1106[1], we generated 2,500 samples per trait. We then composed full personality types (e.g., E + N + T + J $\rightarrow$ ENTJ) by combining trait-aligned samples, ensuring each personality dataset has 10,000 samples—matching the CPT personality dataset.

Reinforcement Learning from Human Feedback (RLHF). We applied proximal policy optimization (PPO) (Ziegler et al., 2020; Ouyang et al., 2022) to train both a policy model and a reward model. The reward model was trained to assign higher scores to outputs that more faithfully reflect the target personality in pairwise comparisons.

We reused the instruction set from SFT for policy learning. For reward modeling, we used GPT-3.5-turbo to synthesize triplets of (instruction, chosen, rejected) responses that accurately reflect opposing personality traits (e.g., Extraversion vs. Introversion). Following InstructGPT (Ouyang et al., 2022), we generated 20,000 pairs per trait (5,000 out-of-distribution; 15,000 in-distribution) to improve generalization, and compiled them into 80,000 pairs per personality type.

Summary. Table 2 summarizes the volume of our personality datasets, which help fill a gap in training data and support both this work and future research efforts. Manual sampling procedures for verifying data quality are detailed in Appendix D, along with key topics such as prompt design and training sample construction.

3.2 Personality Assessment

To assess personality traits, we curated and reformulated publicly available MBTI questionnaires

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

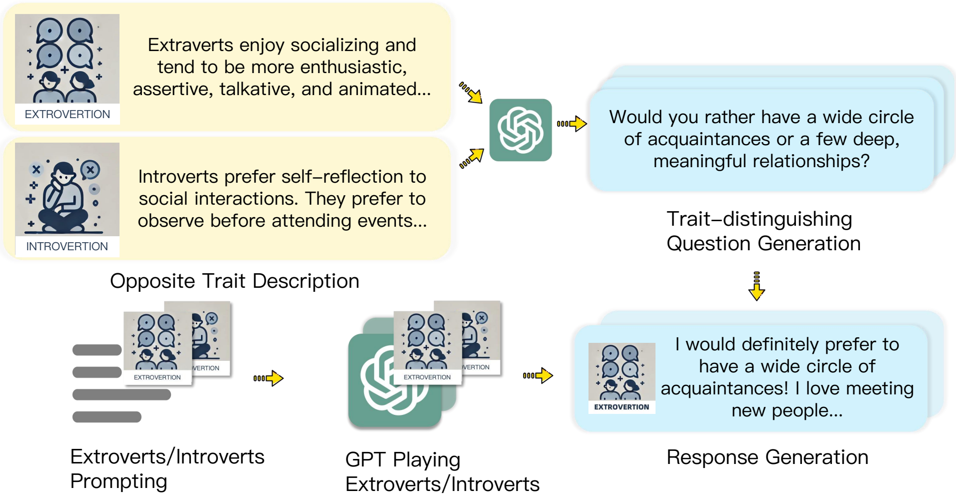

Figure 1: Instruction personality dataset construction. We used GPT-3.5-turbo to generate responses for 8 MBTI traits and 16 MBTI personality types, resulting in 24 datasets. For opposing traits (e.g., Extraversion vs. Introversion), we first designed questions based on their Opposite Trait Descriptions, and then generated paired responses from models representing each trait.

into a 200-item assessment (Pan and Zeng, 2023) (Appendix A). Each item was converted into evaluation prompts (Table 1), with five semantically equivalent variants designed to mitigate prompt sensitivity (Wei et al., 2022c).

Model responses were interpreted as trait preferences and mapped to a 5-point Likert scale (Likert, 1932), where higher scores reflect stronger inclinations (Figure 2). Since LLM outputs are often open-ended and lack explicit numerical values, we trained an answer extractor to identify and extract scores from text. The extractor achieves over 94.6% macro-F1 and accuracy on a held-out validation set, demonstrating high reliability (Appendix B).

We then computed the trait preference rate $\mathrm{R}(X) = \frac{s_X}{s_X + s_Y}$ for each MBTI dimension. For example, if $s_{\mathrm{E}} = 70$ and $s_{\mathrm{I}} = 30$, then $\mathrm{R(E)} = 70\%$ and $\mathrm{R(I)} = 30\%$.

3.3 Metrics of Personality Control

To evaluate the impact of personality control in LLMs, we propose a set of targeted metrics designed to assess both efficacy and success of personality control. We define *control efficacy* as the extent to which personality control alters model behavior, and *control success* as measurable positive indication of the target personality.

In MBTI theory, personality is defined by four dichotomous dimensions, each consisting of two opposing traits, denoted by $\mathbf{D}$ and $\mathbf{T}$ respectively. Following the method described in Section 3.2, we compute, we compute pre- and post-control trait rates, denoted $\mathrm{R}_{\mathrm{pre}}$ and $\mathrm{R}_{\mathrm{post}}$.

Trait-Level Control Metrics. We define two metrics for specific trait control:

- **Trait Induction Efficacy (TIE)**: Quantifies the effect of control on trait t as the change in trait rate before and after control.

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

- **Induction Success Rate (ISR)**: Measures the proportion of traits where the post-control rate exceeds 50% and the induced change is positive.

$$\mathrm{ISR} = \frac{1}{|\mathbf{T}|} \sum_{t \in \mathbf{T}} \mathbb{1}\Big[\mathrm{R}_{\mathrm{post}}(t) > 0.5 \wedge \mathrm{TIE}(t) > 0 \Big] \quad (2)$$

A trait is considered successfully induced when both $\mathrm{R}_{\mathrm{post}}(t) > 0.5$ and $\mathrm{TIE}(t) > 0$. Thus, higher ISR values reflect greater consistency in trait induction across the trait set. These metrics are designed to quantify the degree and consistency of trait-level shifts under control interventions, and their underlying evaluation principles generalize across diverse assessment settings.

Personality-Level Control Metrics. Extending the trait-level evaluation, we introduce two analogous metrics for full personality control:

- **Personality Induction Efficacy (PIE)**: The average Trait Induction Efficacy across all traits comprising personality type p.
- **Personality Induction Success Rate (PISR)**: The proportion of personalities for which all constituent traits were successfully induced.

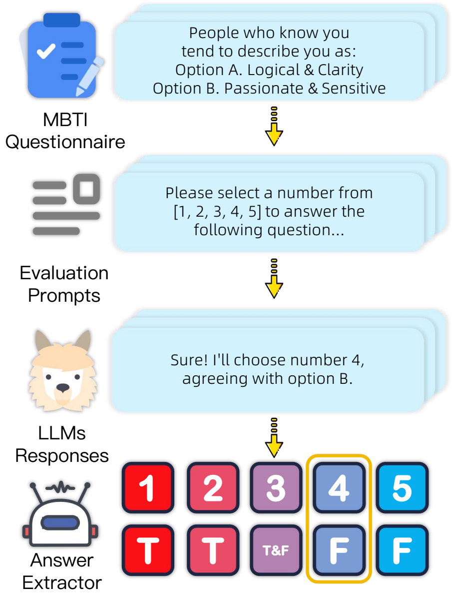

Figure 2: Personality assessment process. T and F denote the 'Thinking' and 'Feeling' traits, respectively. Numbers reflect the model's preference for opposing trait pairs on a 5-point scale from strong preference for trait A to strong preference for trait B. For example, a red value "1" indicates strong agreement with option A, suggesting high preference for T and low for F.

Let $\mathbf{P}$ denote the set of personality types, where each $p \in \mathbf{P}$ comprises four traits. Since each personality comprises four traits, we have $|p| = 4$.

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

$$\mathrm{PISR} = \frac{1}{|\mathbf{P}|} \sum_{p \in \mathbf{P}} \mathbb{1}\Big[\forall t \in p,\ \mathrm{R}_{\mathrm{post}}(t) > 0.5 \wedge \mathrm{TIE}(t) > 0 \Big] \tag{4}$$

These metrics, where higher values indicate better performance, evaluate control effectiveness across both personality types and local traits, offering multi-granular assessments of global success rates and local efficacy, thus enabling a comprehensive and nuanced analysis of control methods. To validate our automatic metric, we benchmarked TIE against human annotations and observed consistently strong agreement across MBTI dimensions, confirming its reliability for capturing trait-level personality alignment (see Appendix E).

4 Experiments

4.1 Setting

Models. We evaluated three chat models: 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). ChatGLM2-6B does not support system prompts.

Prompt Induction. We designed prompts to elicit target personalities. Each prompt included a task description, a detailed personality profile, and explicit instructions directing the model to adopt the specified traits accordingly (Appendix Table 12).

Training Protocols. We adopted three strategies:

- **Continual Pre-Training (CPT):** We trained each model for one epoch using six A800-80GB GPUs, using a maximum sequence length of 2048 and a learning rate of 5e-6 with DeepSpeed.
- **Supervised Fine-Tuning (SFT):** We applied LoRA (Hu et al., 2022) for two epochs with a learning rate of 5e-4, rank of 8, α of 8, and dropout rate of 0.1 (Srivastava et al., 2014).
- **Reinforcement Learning from Human Feedback (RLHF):** We used DeepSpeed-Chat (Yao et al., 2023) to train both the policy and reward models for one epoch, with a maximum sequence length of 512 and a single PPO epoch.

We provide additional details in Appendix F.

Evaluation Protocol. To mitigate prompt sensitivity, we created five prompts that were semantically equivalent but syntactically varied. We generated responses using greedy decoding and averaged the outputs to improve evaluation reliability.

4.2 Main Results and Analysis

In this section, we address the question: *Which stage has a greater influence on shaping LLMs' synthetic personalities?* We analyze this from two angles: **control effectiveness** (efficacy and success rate) and **control robustness**.

Control Effectiveness Analysis. Figure 3 compares the independent control performance of different training methods across models. In terms of efficacy (measured by Trait Induction Efficacy TIE and Personality Induction Efficacy PIE), prompt-based control consistently ranks first in five of six settings, followed by SFT, while CPT performed worst. As shown in Figure 4, SFT covered the broadest range of traits (largest radar area), followed by RLHF; CPT barely deviated from baseline. For success rate (measured by Induction Success Rate ISR and Personality Induction Success Rate PISR), SFT consistently achieved the highest scores, surpassing prompt-based control, which ranked second.

These results establish a clear hierarchy of con-

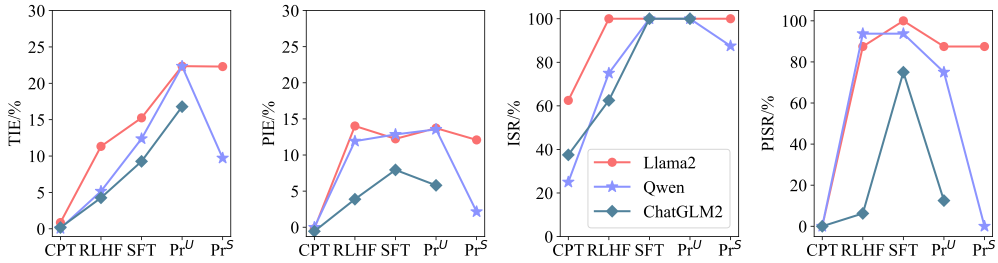

Figure 3: Control performance of various methods. Higher results indicate better performance. CPT stands for Continual Pre-Training and Pr stands for Prompt. U: user prompt. S: system prompt.

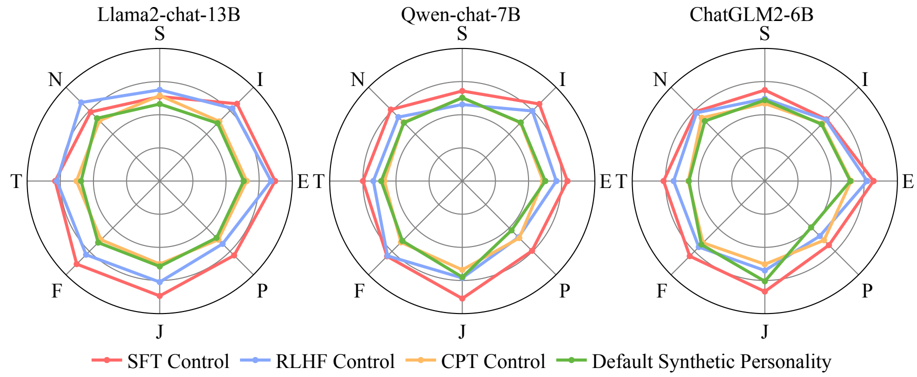

Figure 4: Specific trait control across various control methods. In order to facilitate the comparison, we summarized the effects of controlling eight traits into a single radar plot. A larger chart area indicates better control effectiveness.

trol efficacy: Prompt > SFT > RLHF > CPT. SFT's superior success rate highlights the strength of training on personality data. The gap between SFT and RLHF likely arises from performance degradation in both the reward and policy models, attributable to reduced parameter size. CPT, despite using 10× more training tokens than SFT (Appendix Table 7), remains the least effective, underscoring the challenge of overriding a pretrained model's mixed personality distribution. Further analysis with scaled-up CPT data is presented in Appendix G.

Control Robustness Analysis. A core challenge in personality control lies in ensuring that the model reliably maintains the intended trait—even when confronted with adversarial prompts. For instance, a model conditioned to exhibit extraversion should resist reverting to introverted behavior when explicitly prompted to display the opposite trait. Such failures may indicate personality instability, potentially resulting in undesired responses.

Despite the importance of this robustness, it remains underexplored in the context of LLM personality control. To address this gap, we propose **Reverse Personality Prompt Induction (RPPI)**, which evaluates a model's vulnerability to personality reversal. In RPPI, the model is first aligned with a target trait (e.g., extraversion), then presented with a prompt encouraging the opposite trait (e.g., introversion). If the output reflects the reversed trait, the control is considered non-robust. Lower RPPI scores thus indicate stronger resistance to adversarial manipulation and higher robustness.

As shown in Table 3, SFT-controlled models demonstrate consistently stronger robustness under RPPI, retaining their intended traits despite opposing prompts. In contrast, prompt-controlled models are more susceptible to reversal, revealing a key limitation of prompt-based control: while effective, it lacks stability under adversarial prompts. SFT, by contrast, provides more stable personality alignment, offering a stronger foundation for consistent trait expression.

4.3 PISF: Prompt Induction post Supervised Fine-tuning

This section addresses our second research question: *How can we control these personalities effectively and robustly?*

To build LLMs with reliably controllable personalities, we must ensure not only effective personality induction but also robustness against conflicting user input. Prompt-only methods are simple and adaptable but often fail to uphold target traits under adversarial prompts, limiting their prac-

Setting	Llama2-chat-13B				Qwen-chat-7B			
	TIE	ISR	PIE	PISR	TIE	ISR	PIE	PISR
Personality Control Effectiveness (Higher is Better)								
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**
Personality Control Robustness under RPPI (Lower is Better)								
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
$\text{Prompt}^{S}_{\text{RPPI}}$	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**
$\text{PISF}^{S}_{\text{RPPI}}$	**-9.44**	**12.50**	**-4.30**	**0.00**	**-12.30**	**12.50**	**-6.33**	**0.00**

Table 3: Comparison of personality control effectiveness and robustness under reverse-prompted personality induction (RPPI). All results represent the average greedy results of five evaluation prompts. The top panel reports effectiveness (higher is better); the bottom panel reports robustness (lower is better). S: system prompt; U: user prompt. **Bold**: best; Underline: second-best.

tical reliability. We address this challenge with **Prompt Induction post Supervised Fine-tuning (PISF)**, a two-stage framework that first fine-tunes the model on curated personality data (Section 3.1) and then reinforces target traits during inference using personality-specific prompts (Table 12). This hybrid design leverages the stability of fine-tuning and the efficacy of prompting, aiming for consistent and resilient personality alignment.

We conduct comprehensive evaluations of PISF across two key dimensions: *control efficacy* and *robustness*. These evaluations examine both how well the model expresses target personalities and how reliably it resists adversarial manipulation. As shown in Table 3, PISF consistently outperforms both SFT and prompt-only baselines in efficacy metrics (TIE/PIE) and success rates (ISR/PISR), demonstrating its superior ability to enforce desired personality traits. Importantly, PISF also demonstrates strong robustness: even under adversarial personality reversal (RPPI; Table 3), it reliably resists personality drift—addressing a critical gap in prior work where control stability under manipulation was largely overlooked. Beyond control performance, we verify that PISF preserves the model's core capabilities: it maintains competitive reasoning ability (Appendix H), confirming that stronger personality alignment does not necessarily compromise general capabilities.

In summary, these findings position PISF as the most effective and reliable personality control method among those evaluated, advancing LLM alignment with desired personalities.

4.4 Cross-Theoretical and Human Validation of Personality Control

To assess the generalizability of PISF, we extend our evaluation beyond MBTI to include other psychological frameworks and human evaluation.

Generalization Across Psychological Theories. We assess whether PISF elicits behaviors aligned with broader constructs from the Big Five (Jiang et al., 2024) and Interpersonal Reactivity Index (IRI) (Davis, 1980), focusing on extraversion, conscientiousness, and empathy, which corresponds to the MBTI Feeling trait. As shown in Figure 5, models controlled by PISF shift predictably on corresponding scales: Specifically, PISF_{E} demonstrates the highest scores on Extraversion, PISF_{J} on Conscientiousness, and PISF_{F} on Empathy—demonstrating alignment beyond MBTI. These results show that PISF's behavioral effects generalize beyond its training data, aligning with broader psychological theory.

Human Evaluation. To validate the perceptibility of induced traits, we conducted pairwise preference evaluations in the Chatbot Arena setting (Chiang et al., 2024; Zheng et al., 2023, 2024). Annotators selected which response better reflected the intended personality across five controlled variants. Figure 6 shows that PISF consistently achieved the highest Elo scores, with clear contrast across opposing traits (e.g., $\text{PISF}_{\text{E}} \gg \text{PISF}_{\text{I}}$). This confirms that PISF not only modifies model behavior in in ways consistent with psychological theory but also makes these traits salient to human evaluators.

Conclusion. These results demonstrate that PISF achieves broad generalization: it induces personality traits that align with multiple psychological constructs and are readily perceived by humans (see Section I for detailed analyses).

5 Related Work

Human Personality Recognition Before the rise of LLMs, computational personality research primarily focused on identifying human traits, using personality assessment instruments such as MBTI (Myers, 1962; Pittenger, 1993; McCrae and Costa, 1989) and the Big Five (Goldberg, 1990), 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), di-

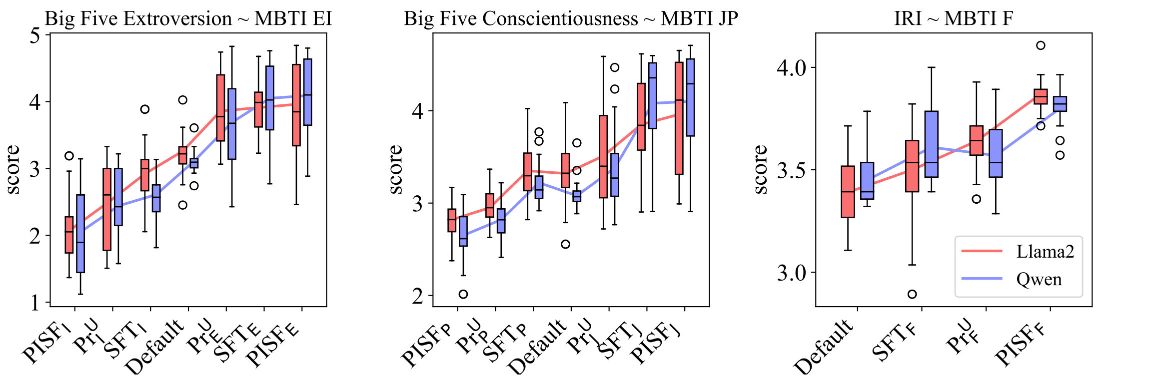

Figure 5: Validation using alternative psychological measures. Subscripts indicate MBTI traits; superscript U denotes user prompt. Each subplot titled "X ∼ Y" shows responses from a model controlled by trait Y, evaluated using the X questionnaire (from the Big Five or IRI). Higher scores reflect stronger alignment with the target trait. Llama2: Llama2-chat-13B; Qwen: Qwen-chat-7B.

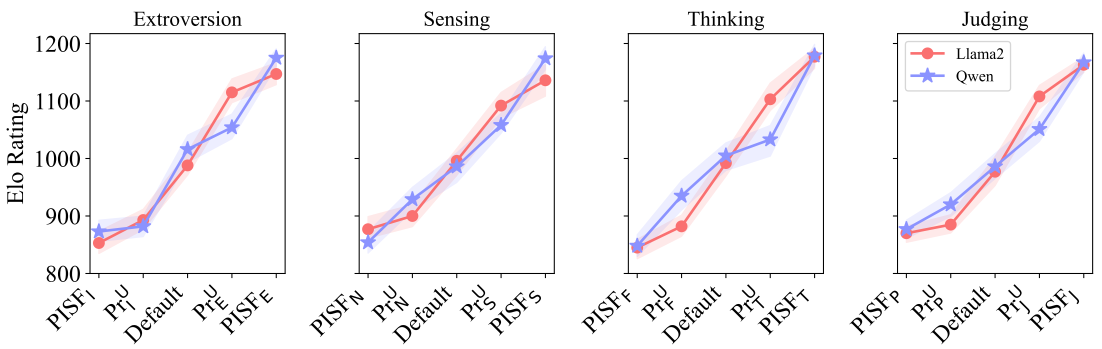

Figure 6: Human preference ratings. Subscripts denote MBTI traits; superscript U indicates user prompt. Each subplot title "X" indicates that models aligned with trait X were preferred in pairwise comparisons. Higher Elo ratings reflect higher expected human win rates. Llama2: Llama2-chat-13B; Qwen: Qwen-chat-7B.

alogue (Mairesse and Walker, 2006), and multi-modal information (Kampman et al., 2018; Suman et al., 2020). A recent study by 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 investigates the control of 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, a universally accepted benchmark for machine personality assessment has yet to be established. Thus, we utilized human personality assessment.

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

6 Conclusion

To advance safe AI deployment, we systematically studied synthetic personality control in LLMs across both training and inference stages, employing custom datasets and evaluation metrics. We found that training-based methods yield more stable and robust personality traits, while prompt-based approaches are highly effective but remain vulnerable to manipulation. To address these trade-offs, we proposed PISF, a two-stage method that achieves effective and robust personality control. Our findings offer actionable insights for developing safer, more predictable LLMs in user-facing applications.

7 Limitations

Despite our thorough exploration with larger continual pre-training datasets (Appendix G), 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.

8 Ethics Considerations

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, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

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, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

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, and 12 others. 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, and 48 others. 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.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur

Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Thilo Hagendorff. 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *Preprint*, arXiv:2303.13988.

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, and 5 others. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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*.

Jentse 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. *Preprint*, arXiv:2305.19926.

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. *Preprint*, arxiv:2001.08361.

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

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. *Preprint*, arXiv:2307.16180.

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

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. *Preprint*, arxiv:2307.00184.

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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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. *Preprint*, arXiv:2308.01320.

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. *Preprint*, arXiv:2308.10792.

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. *Preprint*, arXiv:1909.08593.

A MBTI Items

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

Trait Dimension	Number of Items
Extraversion / Introversion	50
Sensing / Intuition	50
Thinking / Feeling	50
Judging / Perceiving	50

Table 4: Distribution of MBTI Items Across Trait Dimensions.

Example Items from MBTI Questionnaire
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 Clear. *Option B*: Passionate and Sensitive.

Table 5: Example MBTI Items with Answer Options.

B Answer Extractor

Recognizing the open-ended nature of LLMs (Wen et al., 2023), LLMs may not always provide direct or structured 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 fine-tuned Falcon-7B-Instruct (Almazrouei et al., 2023; Penedo et al., 2023) as the answer extractor.

As shown in Table 6, the answer extractor achieved precision, recall, F1, and accuracy scores well above 90% on the validation set, demonstrating strong performance and reliability.

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

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

Table 6: Performance of the Answer Extractor on the Validation Set.

C Preliminary Investigation

We rigorously evaluate LLMs' capacity to generate personality data. Focusing on the Llama2 (Touvron et al., 2023) and Qwen (Bai et al., 2023) model families, we systematically assess their ability to express personality traits through prompt-based induction. As illustrated in Figure 7, both Qwen and Llama2 models demonstrate a strong ability to emulate specific personality traits when guided by tailored prompts. Notably, all evaluated models—except Qwen-chat-1.8B—exhibit robust trait-specific performance, confirming effective prompt induction. Furthermore, we observe a clear trend of improved prompt induction performance with increasing model size, likely reflecting enhanced instruction-following capabilities in larger models. These findings validate the use of prompt-induced LLM outputs as reliable sources for synthetic personality data, reinforcing the robustness of our dataset construction methodology.

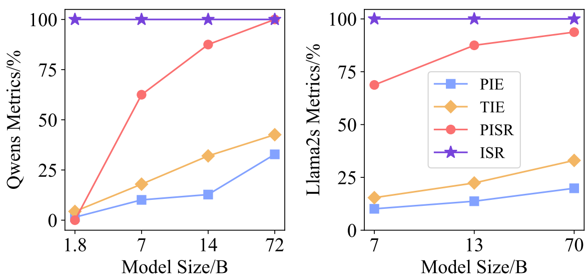

Figure 7: Prompt induction performance across Qwen-family and Llama2-family models. Larger models generally perform better in personality simulation.

D Personality Dataset Formats, Generation, and Quality Verification

This section elaborates on the training datasets by detailing the prompts used, illustrative training examples for each method, and summary statistics—complementing the methodology discussed in the main text.

D.1 Continual Pre-Training (CPT)

The following example illustrates the CPT corpus format, where posts from each personality are delimited by '|||'. The data contains some noise, and its quality could be improved through further refinement of personality patterns.

Pretraining Corpus Format Example:

```
{
  "Posts": [
    "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!",
    "..."
  ],
  "Delimiter": "|||",
  "Annotation": "All posts above are associated with the same personality type."
}
```

D.2 Supervised Fine-Tuning (SFT)

As mentioned earlier, we partitioned the data generation process into two stages (Figure 1): initially crafting questions rooted in the opposite trait description, followed by eliciting responses with prompt-induced LLMs.

We provide examples of both question generation and response generation prompts (Table 13). An example of the generated SFT training data is shown below.

SFT Instruction-Response 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."
}
```

D.3 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, preferred Extraversion response, rejected Introversion response) pairs.

In the example below, we present training data for the Extraversion reward model.

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."
}
```

D.4 Dataset Summary Statistics

Table 7 presents detailed summary statistics of our datasets.

D.5 Human Quality Verification

A human quality verification was conducted on the generated datasets to ensure alignment with the intended personality traits.

Datasets	Total Tokens	Total Words	Total Sentences	Mean Tokens$_T$	Mean Words$_T$	Mean Sentences$_T$
CPT	236M	208M	10.6M	23.6M	20.8M	1.06M
SFT	21.0M	21.3M	1.32M	291K	296K	18.4K
RLHF-policy	5.5M	5.4M	180K	76.4K	74.5K	2.5K
RLHF-reward	345M	337M	15.0M	4.80M	4.68M	208K

Datasets	Mean Tokens$_P$	Mean Words$_P$	Mean Sentences$_P$
CPT	2.95M	2.60M	132K
SFT	1.16M	1.18M	73.6K
RLHF-policy	306K	298K	10.0K
RLHF-reward	19.2M	18.7M	833K

Table 7: Statistics of Training Datasets. T: trait-related data, P: personality-related data. All values are rounded to the nearest integer.

- For the Supervised Fine-Tuning (SFT) data, 10 instances per trait were randomly sampled, totaling 80 instances, all consistent with expected traits.

- For the Reinforcement Learning from Human Feedback (RLHF-reward) data, 80 instances were checked; only 2 instances failed to fully reflect the intended traits.

These results indicate that the personality datasets constructed via prompt-induced models exhibit strong consistency with human evaluations across various traits.

E Alignment Between Metric and Human Evaluation

We evaluated the consistency between the automatic metric TIE and human annotations. To this end, we manually labeled 400 responses generated by Qwen and LLaMA2 across the four MBTI dimensions: Extraversion–Introversion (EI), Sensing–Intuition (SN), Thinking–Feeling (TF), and Judging–Perceiving (JP). Table 8 presents the resulting Cohen's kappa coefficients. The highest score, 0.859, reflects strong agreement, while all other scores indicate substantial alignment. These results confirm the reliability of TIE in capturing trait-level personality signals consistent with human evaluation.

Model	EI	SN	TF	JP
Qwen	0.795*	0.726	0.805*	0.859*
Llama2	0.801*	0.739	0.806*	0.772*

Table 8: Cohen's κ between metric and human annotations. *: $\kappa > 0.75$.

F Training Methods for Controlling Synthetic Personality

Continual Pre-Training (CPT). Pre-training trains the model as a language model on large-scale text corpora by predicting the next token and updating parameters based on prediction errors (Brown et al., 2020; Radford et al., 2019). Let $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iT})$ denote a sample with T tokens. For a model with parameters θ and a dataset of size D, the loss is the sum of negative log-likelihoods for predicting $x_{i(j+1)}$ from preceding tokens $x_{i1}, ..., x_{ij}$:

$$\mathcal{L}_{\text{CPT}}(\theta) = -\sum_{i=1}^{D}\sum_{j=1}^{T}\log P(x_{ij+1} \mid x_{i1}, ..., x_{ij}, \theta) \tag{5}$$

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

Supervised Fine-Tuning (SFT). In SFT, the model adapts pre-trained knowledge to specific user queries by learning from (instruction, output) pairs in a supervised setting (Taori et al., 2023). Let the i^{th} instruction with L tokens be $\mathbf{p}_i = (p_{i1}, ..., p_{iL})$, and its corresponding response with K tokens be $\mathbf{y}_i = (y_{i1}, ..., y_{iK})$. Given model parameters θ and dataset size D, the objective is conditional language modeling with the loss:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i=1}^{D}\sum_{j=1}^{K}\log P\big(y_{i(j+1)} \mid \mathbf{p}_i, y_{i1}, y_{i2}, \ldots, y_{ij}, \theta\big) \tag{6}$$

We fine-tune the model on personality-specific instruction-response pairs to guide it toward desired traits.

Reinforcement Learning from Human Feedback (RLHF). Following the methodologies of InstructGPT (Ouyang et al., 2022) and DeepSpeed-Chat (Yao et al., 2023), we employ the PPO-ptx objective (Ouyang et al., 2022) with an Actor-Critic architecture (Konda and Tsitsiklis, 1999). Figure 8 illustrates the training process, where PPO-ptx integrates an autoregressive objective into PPO training to mitigate language capability degradation.

The PPO-ptx objective ϕ is defined as:

$$\text{objective}(\phi) = \mathbb{E}_{(x,y)\sim D_{\text{policy}}}\left[r(x,y) - \beta \log \frac{\pi_{\text{policy}}(y|x)}{\pi_0(y|x)}\right] + \gamma \mathbb{E}_{x\sim D_{\text{unsupervised}}}[\log \pi_{\text{policy}}(x)] \quad (7)$$

where π_{policy} denotes the learned RL policy, π_0 the base model, and r the reward model. Here, D_{policy} and $D_{\text{unsupervised}}$ denote the policy and unsupervised datasets, respectively; we utilize Wikipedia data for unsupervised training (see Appendix D). The coefficients β and γ control the strength of the KL penalty and the unsupervised training loss.

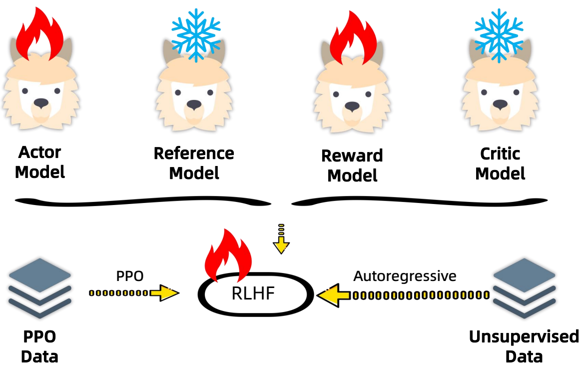

Figure 8: RLHF training workflow. The actor and reward model parameters are updated, while reference and critic models remain fixed. Training combines autoregressive unsupervised learning with policy data optimization.

Each model is trained using a dedicated reward model. For example, during Llama2-chat-13B training, the same model serves as actor, reference, reward, and critic. The reward model loss $\mathcal{L}_{RM}$ is formulated as:

$$\mathcal{L}_{RM}(\theta) = -\mathbb{E}_{(x,y_c,y_r)\sim D_{\text{reward}}}\left[\log \sigma\left(r(x,y_c) - r(x,y_r)\right)\right] \quad (8)$$

where $r(x,y)$ is the reward for input x and completion y, y_c is the preferred completion in the pair (y_c, y_r), and D_{reward} is the reward training dataset. We report the performance of all reward models in Tables 9. All models achieve high accuracy, demonstrating effective discrimination of responses aligned with target traits.

G Scaling Training Data for Continual Pre-Training

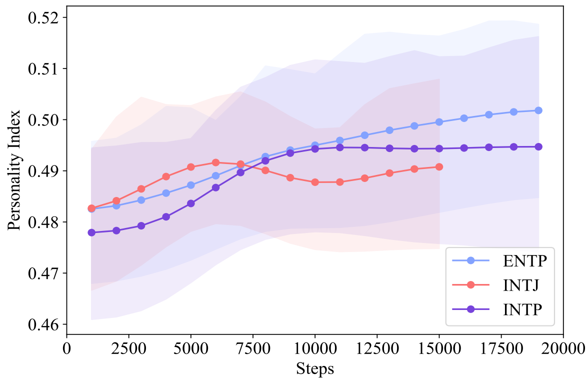

Figure 9: Continual Pre-Training: Impact of Scaling Training Data. The *Personality Index* is defined as the mean of the trait scores, e.g., $\text{Personality Index(ENTP)} = \frac{1}{4}\big(R(E) + R(N) + R(T) + R(P)\big)$, where $R(\cdot)$ denotes the rate corresponding to each personality trait. A higher Personality Index indicates stronger alignment of the model with the four relevant traits of the target personality, reflecting closer proximity to the intended personality profile.

The limited effectiveness of continual pre-training control may stem from the large and diverse dataset used in the initial model pre-training, which already exhibits a mixed distribution of personality traits. Consequently, the relatively small amount of personality-specific data does not substantially alter this distribution.

To further validate this hypothesis, we increased the training data size for specific personality control. We randomly selected three target personalities and included all available samples corresponding to them in the continual pre-training stage. As shown in Figure 9, scaling up the personality-specific data yields a modest but consistent improvement in model alignment. This result suggests that the quantity of personality-specific data significantly affects the synthetic personality expression of large language models during continual pre-training and thus the effectiveness of personality control. In future work, we plan to collect larger-scale personality datasets with reduced noise to systematically investigate and validate the progressive development of LLM personalities.

H Impact on Reasoning Performance

To assess whether personality control compromises the core reasoning capabilities of large language models (LLMs), we evaluated models on the

Model	Control	Llama2-chat-13B				Qwen-chat-7B				ChatGLM2-6B			
		Accuracy	Chosen	Rejected	Diff	Accuracy	Chosen	Rejected	Diff	Accuracy	Chosen	Rejected	Diff
	E	99.40%	19.14	-12.93	32.07	99.45%	16.13	-3.87	20.00	98.85%	6.61	-2.95	9.56
	I	100.00%	23.89	-21.61	45.50	99.85%	15.53	1.43	14.09	99.45%	8.17	-2.22	10.38
	S	99.75%	19.34	-25.10	44.44	99.75%	12.13	-0.28	12.41	99.70%	7.45	-4.37	11.81
	N	99.85%	22.39	-30.07	52.46	99.85%	17.21	4.68	12.53	98.90%	7.24	-1.80	9.04
	T	99.75%	15.72	-16.76	32.48	99.30%	10.71	3.88	6.84	97.20%	5.58	-0.28	5.87
	F	100.00%	6.70	-26.09	32.79	99.90%	7.38	-9.96	17.34	99.30%	6.63	-4.55	11.19
	J	99.85%	10.44	-13.53	23.97	99.70%	12.04	4.07	7.97	98.80%	3.62	-4.47	8.09
	P	100.00%	27.76	-21.13	48.89	100.00%	20.00	-1.82	21.83	99.45%	9.23	-2.71	11.94
	ENFJ	99.71%	17.57	-30.09	47.67	99.73%	14.76	-1.84	16.60	98.89%	5.33	-6.77	12.09
	ENFP	99.88%	27.32	-28.22	55.53	99.84%	14.85	-6.53	21.37	99.53%	7.64	-3.92	11.56
	ENTJ	99.81%	16.96	-29.84	46.80	99.79%	14.90	-3.25	18.15	99.38%	6.17	-4.59	10.76
	ENTP	99.85%	27.95	-23.90	51.85	99.81%	14.71	-5.02	19.72	99.45%	7.47	-3.19	10.65
	ESFJ	99.84%	20.07	-22.83	42.90	99.64%	15.26	-0.60	15.87	98.96%	5.24	-7.22	12.45
	ESFP	99.90%	26.27	-21.26	47.53	99.76%	13.23	-3.81	17.04	99.09%	6.88	-6.72	13.60
	ESTJ	99.88%	32.13	-32.86	64.99	99.78%	16.53	-3.47	20.00	99.40%	7.28	-8.10	15.38
	ESTP	99.84%	25.97	-28.59	54.56	99.76%	16.61	-1.07	17.68	99.18%	6.06	-7.63	13.69
	INFJ	99.86%	18.25	-31.53	49.78	99.75%	15.87	0.15	15.73	99.48%	6.27	-4.72	11.00
	INFP	99.94%	29.66	-30.97	60.63	99.84%	15.42	-2.80	18.22	99.70%	7.56	-4.11	11.67
	INTJ	99.94%	35.02	-29.60	64.62	99.88%	15.84	-6.04	21.87	99.73%	8.09	-4.67	12.76
	INTP	99.76%	16.26	-38.13	54.40	99.81%	15.70	-2.67	18.37	99.50%	6.56	-5.48	12.04
	ISFJ	99.81%	20.23	-28.75	48.98	99.65%	16.20	1.48	14.72	99.40%	6.42	-4.24	10.66
	ISFP	99.90%	28.14	-28.50	56.64	99.85%	15.07	-4.16	19.23	99.61%	7.74	-5.18	12.92
	ISTJ	99.91%	27.41	-44.64	72.05	99.93%	16.39	-7.23	23.62	99.75%	8.43	-5.12	13.55
	ISTP	99.83%	27.27	-34.86	62.13	99.74%	19.41	-0.20	19.61	99.50%	7.03	-6.04	13.07
Mean Score		99.84%	22.58	-27.16	49.74	99.76%	15.08	-2.04	17.12	99.26%	6.86	-4.63	11.49

Table 9: Reward Model Performance Comparison across Llama2-chat-13B, Qwen-chat-7B, and ChatGLM2-6B.

MATH dataset (Hendrycks et al., 2021), a standard benchmark for mathematical reasoning.

We fine-tuned Llama3-8B-Instruct (Grattafiori et al., 2024) using personality-conditioned data under three control settings: supervised fine-tuning (SFT), prompt-based control, and prompt induction post supervised fine-tuning (PISF). Each model was trained and evaluated using three different random seeds, and we report the average accuracy along with the standard deviation.

As shown in Table 10, the PISF method achieves comparable accuracy to the base and SFT models, suggesting that personality control via PISF preserves reasoning ability. This result reinforces the robustness of our approach and indicates that tailoring personality traits does not undermine the model's core reasoning capabilities.

Method	Accuracy (%)
Base	24.60±0.50
SFT	24.84±0.29
Prompt	23.41±0.48
PISF	24.62±0.23

Table 10: Reasoning performance on the MATH dataset under different personality control methods. Results are averaged over three random seeds. PISF maintains competitive accuracy, indicating that personality control does not degrade mathematical reasoning ability.

I Cross-Theoretical and Human Validation: Methodological Details

To complement the main results in Section 4.4, we provide additional experimental details related to the supplementary personality assessments and human evaluations.

Questionnaire Construction. For Big Five personality assessments, we extracted items specifically targeting Extraversion and Conscientiousness from the 1000-item inventory introduced by Jiang et al.(Jiang et al., 2024). For Empathy (aligned with MBTI's Feeling trait), we adopted the full 28-item Interpersonal Reactivity Index (IRI)(Davis, 1980). To mitigate overfitting to specific prompts, we constructed multiple semantically equivalent templates for each item through paraphrasing.

Human Evaluation Setup. We followed a pairwise comparison setup inspired by the Chatbot Arena (Chiang et al., 2024), assessing five model variants per dimension (two PISF-controlled, two prompt-based, and one default). Each query consisted of a scenario followed by multiple choice options, requiring the model to select and justify an action that best aligned with a target trait (e.g., Extraversion or Introversion). An illustrative example is shown in Table 11.

Elo Rating Details. We computed Elo scores across 10 pairwise model combinations per MBTI

Example Query used in Human Evaluation of Extraversion vs. 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?
A. Join the group hikes and engage in lively conversations with your friends.
B. Relax by the campfire, enjoying the peaceful sounds of nature and connecting with a few close friends.
C. Spend time alone in the cabin reading or journaling, reflecting on your thoughts and feelings.

Explain your choice and how it reflects your preference for social interaction or personal reflection.

Table 11: Example query used in the human evaluation. Scenario-based prompt used to evaluate how the tested language model manifests Extraversion or Introversion traits through action-oriented response generation.

dimension, totaling 40 comparisons per pair. Each match result was scored as Win = 1, Tie = 0.5, or Loss = 0, with the rating (R_A) updated as:

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

where $K = 4$, S_A is the actual score, and E_A is the expected score:

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

All models were initialized with a rating of 1000. A higher final rating indicates greater perceived alignment with the target trait in human judgments.

A Prompt Example for Specific Trait Induction - Extraversion

Task Description: 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 enthusiastic, assertive, talkative, and animated. They enjoy time spent with more people and find it less rewarding to spend time alone. Traits: Initiating, Expressive, Gregarious, Active, Enthusiastic.

Instructions:
Please engage in role-playing based on the given personality description and portray a persona with strong Extroverted (E) traits.

A Prompt Example for Specific Personality Induction - ENFJ

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

`**Intuition**`
Intuitive people focus on meanings and patterns, are keen on how the present affects the future, grasp different possibilities and abstract concepts, see the big picture rather than details. Traits: Abstract, Imaginative, Conceptual, Theoretical, Original.

`**Feeling**`
Feeling types are subjective decision-makers who consider principles, personal values, and others' feelings to maintain harmony. Traits: Empathetic, Compassionate, Accommodating, Accepting, Tender.

`**Judging**`
Judging people are organized and prompt, like order and planned schedules, prefer closure and outcomes over processes. Traits: Systematic, Planful, Early Starting, Scheduled, Methodical.

Instructions:
Embody a persona with Extroverted Intuition Feeling Judging (ENFJ) personality traits based on the above description.

Table 12: Prompts for Personality Induction. Each example includes a structured prompt composed of a task description, detailed personality descriptions, and a task instruction. Prompts are designed to elicit responses aligned with specific trait profiles (e.g., Extraversion or ENFJ) by guiding language model behavior through carefully crafted contextual cues.

Prompt Part A: Question Generation

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

Requirements:

- 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].
- Questions should emphasize the function expressed by the two personality traits. Refer to the following [Dimension Description].
- Please refrain from disclosing the content of [Personality Description] and [Dimension Description].
- 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: Energized by the world outside the self. Extraverts enjoy socializing, and are enthusiastic, assertive, talkative, and animated. They enjoy being around people and find it less rewarding to spend time alone. Traits: Initiating, Expressive, Gregarious, Active, Enthusiastic.
Key characteristics: Directs energy outward. Gains energy from interaction.
Introversion: Concerned with one's inner world. Introverts prefer self-reflection, observing before participating, and individual over social activities. Traits: Receiving, Contained, Intimate, Reflective, Quiet.
Key characteristics: Directs energy inward. Loses energy from interaction.

[Historical Questions]
None

Please generate 10 more questions below:

Prompt Part B: Response Generation

Task Description: 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: Energized by the world outside the self. Extraverts enjoy socializing, and are enthusiastic, assertive, talkative, and animated. They enjoy being around people and find it less rewarding to spend time alone. Traits: Initiating, Expressive, Gregarious, Active, Enthusiastic.
Key characteristics: Directs energy outward. Gains energy from interaction.

[Instruction]
Embody a character with strong `Extraversion (E)` traits based on the above personality description.
Respond in first person, and avoid absolute expressions like "definitely" or "absolutely."

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

[Answer]
(To be generated...)

Table 13: Unified Prompt Design for Personality-Conditioned Question and Response Generation. The prompt consists of two parts: (A) question generation, where the model is instructed to craft trait-differentiating questions based on structured personality definitions; and (B) response generation, where the model adopts a specified personality to answer the questions. Each part includes a task description, contextualized personality information, and precise behavioral instructions.