BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data

Anonymous ACL submission

Abstract

In this work, we tackle the challenge of embedding realistic human personality traits into LLMs. Previous approaches have primarily focused on prompt-based methods that describe the behavior associated with the desired personality traits, suffering from realism and validity issues. To address these limitations, we introduce BIG5-CHAT, a large-scale dataset containing 100,000 dialogues designed to ground models in how humans *express* their personality in language. Leveraging this dataset, we explore Supervised Fine-Tuning and Direct Preference Optimization as training-based methods to align LLMs more naturally with human personality patterns. Our methods outperform prompting on personality assessments such as BFI and IPIP-NEO, with trait correlations more closely matching human data. Furthermore, our experiments reveal that models trained to exhibit higher conscientiousness, higher agreeableness, lower extraversion, and lower neuroticism display better performance on reasoning tasks, aligning with psychological findings on how these traits impact human cognitive performance. To our knowledge, this work is the first comprehensive study to demonstrate how training-based methods can shape LLM personalities through learning from real human behaviors.

1 Introduction

Realistically simulating human personality and its impact on text generation is a challenging yet crucial problem (Elster, 2015; Park et al., 2023; Serapio-García et al., 2023; Li et al., 2024; Frisch and Giulianelli, 2024). Embedding personality traits into LLMs can greatly enhance their authenticity across a wide range of applications, from conversational agents (Pradhan and Lazar, 2021) to educational tools (Kanero et al., 2022) and mental health platforms (Tudor Car et al., 2020; Ahmad et al., 2022). By creating more human-like interactions, LLMs can better simulate diverse personas and adapt more reliably to different contexts (Gao et al., 2024a).

However, existing methods primarily rely on prompting models with descriptions of behaviors associated with personality traits (e.g., “You are the life of the party"; Mao et al., 2023; Chen et al., 2024b, 2022; Tu et al., 2024). These behavior descriptions are often drawn from the same psychological questionnaires used to test their personality, raising evaluation validity concerns. More importantly, these behavioral descriptions are nonsensical for text-based LLMs (LLMs do not attend parties), failing to ground their personality in realistic patterns of how humans’ personality is expressed in text (Vu et al., 2024). Additionally, the scarcity of large-scale, human-generated datasets annotated with personality traits has hindered the exploration of training-based approaches, limiting most prior research to prompting-based methods.

In this work, we address the challenge of inducing realistic human personality traits in LLMs by constructing a large-scale dialogue dataset, BIG5-CHAT, which is grounded in real human personality expressions in text. The overview of our work is illustrated in Figure 1. We choose the well-known Big Five personality traits framework to study this (McCrae and John, 1992; Pittenger, 1993), due to its reliability and validity as shown from psychological research. While previous datasets typically include only persona descriptions, our dataset bridges the gap between narrow-domain personality data and general-domain social interactions, ensuring both authenticity and scenario diversity. To achieve this, we combine two primary data sources — PsychGenerator (Vu et al., 2024), a collection of 850K Facebook posts annotated with Big Five trait scores, and SODA (Kim et al., 2022), a rich dataset of diverse social interactions — by utilizing product-of-experts text generation (DExperts; Liu et al., 2021). This combination enables us to capture the nuanced expression of personality traits

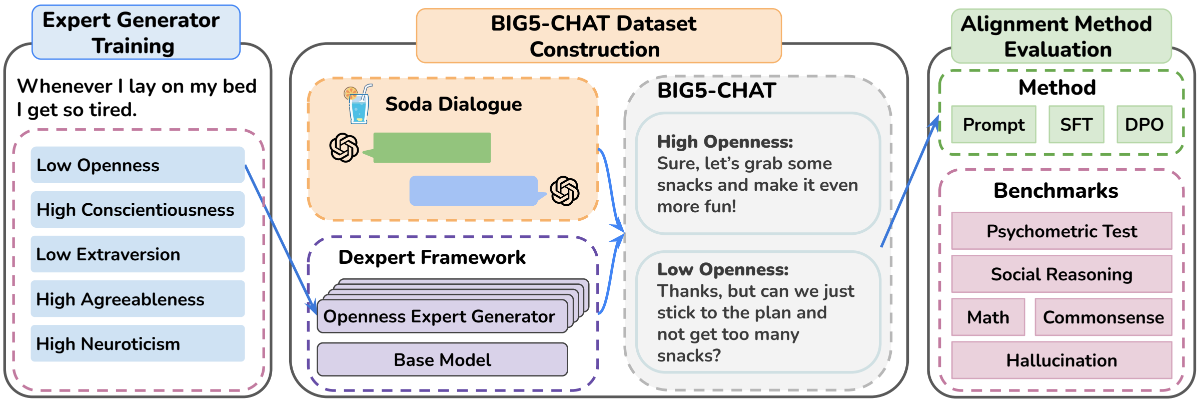
Expert Generator Training
Whenever I lay on my bed I get so tired.
Low Openness
High Conscientiousness
Low Extraversion
High Agreeableness
High Neuroticism
BIG5-CHAT Dataset Construction
Soda Dialogue
Dexpert Framework
Openness Expert Generator
Base Model
BIG5-CHAT
High Openness: Sure, let's grab some snacks and make it even more fun!
Low Openness: Thanks, but can we just stick to the plan and not get too many snacks?
Alignment Method Evaluation
Method
Prompt
SFT
DPO
Benchmarks
Psychometric Test
Social Reasoning
Math
Commonsense
Hallucination

Figure 1: Overview of the PSYCHSTEER method and evaluation. The expert generator was trained on the PsychGenerator dataset to induce Big Five personality traits (Vu et al., 2024) and integrated with the base model using the Dexperts framework alongside SODA's social scenarios (Liu et al., 2021; Kim et al., 2023a) to generate the BIG5-CHAT dataset. Various alignment methods were then evaluated for their effectiveness in inducing personality and their impact on reasoning benchmarks.

across a wide range of dialogue scenarios.

Leveraging our BIG5-CHAT dataset, we empirically investigate how training-based methods grounded in real human data compare to traditional prompting techniques for inducing personality traits in LLMs, including instruction-based and demonstration-based prompting. Specifically, we explore Supervised Fine Tuning (SFT) and Direct Preference Optimization (DPO) (Rafailov et al., 2024) to align LLMs' personalities with Big Five traits. This comparison is crucial for understanding whether data-driven training methods can offer deeper, more reliable personality integration than the surface-level traits typically induced through prompting. Our results demonstrate that both SFT and DPO outperform prompting on two widely recognized Big Five personality tests: the BFI (John et al., 1999) and IPIP-NEO (Johnson, 2014).

In humans, personality traits often correlate with reasoning abilities (John et al., 1999; Soto et al., 2011), raising the question of how embedding personality traits in LLMs may influence their reasoning performance. However, introducing persona-like attributes into LLMs could inadvertently degrade core reasoning capabilities, which is often undesirable for developers who rely on these models for critical decision-making or problem-solving tasks. Understanding how personality induction shapes reasoning patterns is crucial for ensuring that personality-driven behaviors do not come at the cost of diminished cognitive performance. This aligns closely with our motivation by highlighting potential risks to authenticity and effectiveness in applications such as mental health platforms or conversational agents. To explore this, we evaluate our aligned models not only with traditional personality tests but also across five reasoning domains: social reasoning using SocialIQA (Sap et al., 2019), math reasoning using GSM8K (Cobbe et al., 2021) and MathQA (Amini et al., 2019), hallucination detection using TruthfulQA (Lin et al., 2021), commonsense reasoning using CommonsenseQA (Talmor et al., 2019) and PIQA (Bisk et al., 2020), and general reasoning using MMLU (Hendrycks et al., 2020) and GPQA (Rein et al., 2023). Our experiments show that models trained with higher levels of conscientiousness and agreeableness consistently outperform others in reasoning tasks. Conversely, models with lower levels of extraversion and neuroticism exhibit improved reasoning performance in general. These findings mirror patterns between Big Five traits and different reasoning abilities observed in psychological studies in humans (Ackerman and Heggestad, 1997; Schaie et al., 2004), further demonstrating how our personality induction method embeds deeper psycholinguistic traits into models.

In contrast to prior work, which often relies on either purely synthetic or questionnaire-based data, our approach grounds personality induction in human-authored texts, ensuring authentic personality expressions that align with the Big Five framework. This move toward human-grounded data addresses the validity and realism concerns left unanswered by previous methods and achieve more robust and contextually nuanced personality simulation. This work makes the following contributions:

- We introduce the first large-scale dataset, BIG5-CHAT, [1] containing 100,000 dialogues across a wide spectrum of personality expressions, addressing the limitations of existing methods that rely on simple prompting without grounding in real human personality expressions in text;
- We perform quantitative evaluations comparing SFT and DPO to prompting in terms of imbuing LLMs with personality, showing that both training-based methods induce more pronounced personality traits and more realistic intra-trait correlations;
- We conduct comprehensive empirical investigations into how personality traits affect performance in both social reasoning and general reasoning tasks, revealing that LLMs with distinct personality traits demonstrate varying strengths and weaknesses across domains.

2 Background

Drawing from psychological research, the Big Five personality traits framework (McCrae and John, 1992; Pittenger, 1993), comprising five key factors—*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*—has emerged as a reliable model for capturing LLM-simulated personality behavior (Karra et al., 2022; Serapio-García et al., 2023; Li et al., 2022; Pan and Zeng, 2023). According to Yarkoni (2010), the Big Five personality traits manifest in distinct linguistic patterns: *openness* is reflected in intellectual and cultural language, *conscientiousness* in achievement-oriented words with minimal impulsivity, *extraversion* in social and positive emotion terms, *agreeableness* in communal and affectionate expressions, and *neuroticism* in frequent use of negative emotion words. Compared to the Myers-Briggs Type Indicator (MBTI), the Big Five model offers superior reliability, validity, and empirical support, making it the preferred framework in personality research (McCrae and John, 1992; Pittenger, 1993). Extensively validated across cultures, it consistently predicts life outcomes such as job performance and mental health (McCrae and Costa Jr, 1997; John et al., 2008; Barrick and Mount, 1991; Soldz and Vaillant, 1999).

Various prompting approaches have been developed to induce Big Five personality traits in LLMs. They often employ pre-defined scripts or questionnaires to nudge the model towards expressing Big Five personality traits during interactions (Mao et al., 2023; Chen et al., 2024b, 2022; Tu et al., 2024). However, several challenges can arise from using prompting as the personality alignment method:

Lack of psycholinguistic depth LLMs with personality traits induced via prompting often reflect only surface-level traits, lacking the psycholinguistic richness needed for authentic human behavior (Dorner et al., 2023; Sá et al., 2024; Olea et al., 2024). Unlike humans, who adapt dynamically to social and environmental contexts (Bandura et al., 1961; Baldwin, 1992), LLMs rely on static training data, making them less reliable in simulating nuanced human behaviors on downstream tasks (Soni et al., 2023), which can lead to caricature (Cheng et al., 2023).

Validity Concerns in Personality Induction and Evaluation The dual use of psychometric questionnaires for both inducing and evaluating personality traits in LLMs raises validity concerns, potentially biasing assessments (Lievens et al., 2007). This approach risks overfitting to specific linguistic features rather than enabling robust generalization to diverse contexts (Serapio-García et al., 2023; Xu et al., 2024; Mizrahi et al., 2024).

Unintended influence on reasoning patterns Role-based personality prompts can disproportionately shape LLM behavior, causing reasoning patterns to be overly narrow and limited to the explicit traits highlighted in the prompt (Zheng et al., 2023; Lu et al., 2021; Sclar et al., 2023). This influence may lead to imbalanced or overly constrained responses, especially in tasks requiring broader or more nuanced cognitive engagement.

A more comprehensive discussion of the background and related work can be found in Appendix B and Appendix C.

3 Methodology

The lack of large-scale datasets featuring personality-grounded dialogues poses a significant challenge. To address this challenge, we combine controllable text generation models with a domain-specific, personality-annotated dataset. Specifically, we utilize the DExperts framework (Liu et al., 2021) and the PsychGenerator dataset (Vu et al., 2024) to create BIG5-CHAT, a novel dataset that encapsulates diverse personality

---

[1] Our dataset and code are uploaded to the submission system, and will be open-sourced upon acceptance.

expressions within rich dialogue scenarios. The DExperts framework allows us to guide the language model's outputs toward specific personality traits during the generation process. Meanwhile, PsychGenerator provides a comprehensive collection of human-generated texts annotated with Big Five personality trait scores. By combining these technologies, we introduce PSYCHSTEER, an approach that effectively addresses the limitations of prior datasets by grounding personality traits in authentic human interactions.

3.1 DExperts Framework

DExperts allows us to control language model generation at decoding time by steering model outputs with expert generators. By integrating expert generators trained to exhibit different Big Five personality traits, we can induce personality within LLM outputs while maintaining dialogue quality. In the DExperts framework, let M denote the pre-trained base language model, and M^{expert} is the expert generator fine-tuned to generate text exhibiting the desired personality in our tasks. At each time step t, given the prompt and previous token sequence $x_{<t}$, the base model M computes logits $z_t^{\text{base}} \in \mathbb{R}^{|V|}$, where V is the vocabulary. The expert generator M^{expert} computes logits z_t^{expert} in the same manner. To integrate the influence of the expert generator, we adjust the base model's logits by incorporating the scaled difference between the expert generator model and base model logits:

$$z_t^{\text{combined}} = z_t^{\text{base}} + \gamma z_t^{\text{expert}}, \tag{1}$$

where $\gamma \in [0, +\infty)$ is a scaling factor controlling the degree of influence from the expert generator. This formulation effectively pulls the combined logits towards the expert generator logits, where $\gamma = 0$ results in using the base model's logits, and a larger γ indicates a stronger influence of the expert generator's modification control. The combined logits z_t^{combined} are transformed into a probability distribution, and the next token is sampled using the softmax function from this distribution.

3.2 Expert Generator Model Based on Social Media Posts

To train expert generator models to exhibit certain personality traits, we perform SFT on the `LLaMA-3-8B-Instruct` model (Dubey et al., 2024) using the PsychGenerator dataset (Vu et al., 2024). This dataset comprises 846,304 Facebook posts, each paired with its author's Big Five personality trait scores. This dataset provides a robust foundation for training models to simulate nuanced human behaviors associated with different personality dimensions. We fine-tuned five expert generators, each representing and dedicated to generating text corresponding to one of the personality traits. For each personality trait, we converted the original floating-point trait labels into binary levels 'high'/'low' for each trait, allowing the distinct behaviors associated with the extreme ends of each trait to be more easily identified and analyzed.

We fine-tuned our expert generator models using the Alpaca format (Taori et al., 2023), with detailed specifications provided in Appendix D.4. When generating text completions with the PSYCHSTEER framework, the base model generates the first five words. This enables the expert generator model to influence the subsequent token generation by adjusting the logits to favor the desired personality trait while preserving coherence and fluency.

4 BIG5-CHAT Dataset

4.1 Dataset Construction

We introduce **BIG5-CHAT**, a large-scale dialogue responses dataset designed to capture Big Five personality traits within diverse social interactions. Our dataset construction leverages the SODA (Social DiAlogues) dataset (Kim et al., 2023a), which provides a diverse range of realistic social scenarios. SODA dialogues are generated by GPT-3.5 and enriched with social commonsense narratives, making it an ideal foundation for incorporating personality expressions due to its extensive coverage of social interactions. To induce personality traits into the dialogues, we employ the DExperts framework (Liu et al., 2021).

To build our dataset, we randomly sample 10,000 scenarios from SODA to provide diverse social contexts. In SODA, social interactions are modeled between two individuals referred to as Speaker X and Y, representing the participants in each dialogue. For each scenario, we generate a new utterance using our PSYCHSTEER framework to control for personality traits and get the dialogue responses between two participants. In the dialogues, one represents Speaker X (converted from the original SODA dialogue) and another represents Speaker Y with specific personality traits. For Speaker Y, based on the original responses from SODA, we generate new dialogue responses using the PSYCH-

STEER framework. Examples of dialogues from our dataset are shown in Table 4. By conditioning on the preceding context (Speaker X's utterance), we use the base model M guided by the expert generator M^+ specialized in the target personality trait to generate Speaker Y's responses. For each scenario, we generate pairwise dialogues by producing responses that reflect either high or low levels of the targeted personality trait. This approach results in pairs of dialogues that share the same context but differ in the expressed trait level. The process yields a total of 100,000 single-turn dialogues—20,000 for each trait, with an equal split between high and low trait levels.

4.2 Dataset Statistics

In this section, we examine the diversity and clarity of personality trait expressions within our BIG5-CHAT dataset. As illustrated in Table 4, we present examples where, for a single prompt from Speaker X, we have generated ten distinct responses from Speaker Y. These responses are conditioned on the high and low levels of each of the five Big Five personality traits. By varying only the level of a specific trait while keeping the prompt constant, we highlight how each personality trait distinctly influences conversational responses. Additionally, we analyze the token counts and other statistics of generated dialogue responses to ensure consistency across different personality trait levels in Table 5.Further details and discussions about the dataset can be found in Appendix A.

Comparative analysis with existing personality datasets, as presented in Table 6, underscores several advantages of BIG5-CHAT. Unlike existing personality datasets such as Big5PersonalityEssays (Floroiu, 2024) and Machine-Mindset (Cui et al., 2023), which primarily reply on static, non-dialogue content or lack authentic human-grounded data examples, our dataset consists of dialogues capturing dynamic and interactive conversational exchanges that are more representative of natural language use. While previous works have focused solely on purely human-generated domain-specific data or synthetic machine-generated data, our approach uniquely combines both human dialogue and LLM to create realistic personality expressions. These inherent differences—particularly the inclusion of authentic, dialogical exchanges rather than questionnaire-based or domain-restricted data—render direct or quantitative comparisons to other personality datasets both unnecessary and infeasible. In other words, BIG5-CHAT addresses a fundamentally different research need. These findings are further validated through human evaluation, with more information available in Appendix E.1.

4.3 Evaluating Personality-Steering of the Data Generator

To help evaluate the quality of the generated dataset and its reflection of realistic personality traits, we trained a `RoBERTa-Large` (Liu et al., 2019) classifier with five regression heads using the MSE loss function. The model was trained on the PsychGenerator dataset, where the input consisted of text posts, and the output comprised the original trait labels, i.e., five floating-point values ranging from 0 to 1. The same train-validation-test split was applied here as with the expert generators. Training was conducted over five epochs with a learning rate of 1×10^{-5}. In Table 1, we observe that the classifier achieves an accuracy of 93.8% on the held-out test set, indicating that the PsychGenerator dataset contains distinct, learnable patterns that differentiate between high and low levels of personality traits. Refer to the discussion on the classifier for the Big Five Essay dataset in Appendix D.1.

Using the classifier as an evaluator, we demonstrate the high quality of the dataset generated by our expert generator, as shown at the bottom of Table 1, where it accurately reflects realistic personality traits. Specifically, we compare our dataset to a baseline for generating post datasets using LLMs: *Post-Completion*. *Post-Completion* replicates the expert generator's post generation strategy by prompting an LLM to complete a post given the first five words, the target personality traits, and the required post format for post-expression style guidance. We ran *Post-Completion* using `GPT-4o-mini` (OpenAI, 2024). For consistency, all experiments are based on the same set of 1,000 examples randomly chosen from the PsychGenerator test set. The classifier was used to evaluate the generated data by predicting the levels of each trait, and the quality was measured by whether the predictions matched the desired personality traits. Our results in Table 1 show that our expert generator outperforms the baseline, achieving higher average accuracy scores for every personality trait dimension compared to the *Post-Completion* baseline. Additional details about the baseline methods can be found in Appendices D.2 and D.3. These findings are further validated through human evaluation,

Data Generation Method	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**	**Average**
Test set (eval classifier accuracy)	93.7	94.2	93.4	93.4	94.3	93.8
***Ours*: Generator**	82.5	80.0	80.0	81.0	78.5	**80.4**
Post-Completion: `GPT-4o-mini`	64.0	59.5	56.0	57.0	59.5	59.2

Table 1: Accuracy (%) of the trained classifier in predicting each of the Big Five personality traits. The first row (Test set) shows the classifier's accuracy on the test split, demonstrating that the classifier is well-trained. The remaining rows display the performance of our generator model compared to the baseline, as assessed by the same classifier.

with more information available in Appendix E.2.

5 Experiments

In this section, we first outline the experimental setup in Section 5.1, detailing the training procedures for the expert generators and the evaluation of various alignment strategies used to induce personality traits in LLMs. Next, we present the results of the personality tests in Section 5.2, followed by an analysis of the models' reasoning performance in Section 5.3.

5.1 Experiment Setup

Expert generator training We trained five expert generators, each dedicated to generating text corresponding to one of the Big Five personality traits. More training details about the expert generator are explained in Appendix D.4.

Prompting and training strategies We implemented two baseline prompting strategies to induce personality traits in LLMs. The first strategy, *instruction-based prompting*, directly instructs the model to exhibit specific Big Five traits. The second strategy, *demonstration-based prompting*, involves providing the model with 10 in-context examples randomly selected from our BIG5-CHAT dataset to demonstrate the behaviors corresponding to the desired traits. The instruction-based approach relies on explicit descriptions (e.g., "what people typically do"), while the demonstration-based approach draws from behaviorally-driven examples (e.g., "what people typically say"). These baselines were compared to trained models using SFT and DPO, implemented via LoRA (Hu et al., 2022). When training models using DPO, the negative responses are derived from the same personality trait but with the opposite level. For example, if the goal is to imbue the LLM with high openness, the positive response is taken from Speaker Y exhibiting high openness, while the negative response is generated by Speaker Y with low openness. These trained models were later prompted in a manner consistent with their training data format, where personality trait names and levels were explicitly specified in the instructions. The experiments were conducted using two versions of the LLaMA model: `LLaMA-3-8B-Instruct` and `LLaMA-3-70B-Instruct`. More prompting and training details are explained in Appendix D.5 and Appendix D.6.

Evaluation procedure For personality trait evaluation, we adopted the methodology from Huang et al. (2024) for the BFI test, which consists of 44 questions, each rated on a scale from 1 (strongly disagree) to 5 (strongly agree). For the IPIP-NEO test, we utilized the 120-question set from Jiang et al. (2024a), which also employed a 1 to 5 rating scale. We measured the standard deviation by repeating each experiment five times, using a temperature setting of 0.6. To assess reasoning capabilities, we evaluated the models across five domains: (1) social reasoning on SocialIQA (Sap et al., 2019), (2) math reasoning on GSM8K (Cobbe et al., 2021) and MathQA (Amini et al., 2019), (3) hallucination detection on TruthfulQA (Lin et al., 2021), (4) commonsense reasoning on CommonsenseQA (Talmor et al., 2019) and PIQA (Bisk et al., 2020), and (5) general reasoning on MMLU (Hendrycks et al., 2020) and GPQA (Rein et al., 2023). Further evaluation setup details are explained in Appendix D.7.

5.2 Personality Trait Assessment Results

Table 2 presents the BFI and IPIP-NEO assessment results across direct inference and various alignment baselines and methods, including instruction-based prompting, demonstration-based prompting, SFT, and DPO. The performance trends are consistent across both personality tests. Compared to direct inference, which lacks any personality trait descriptions, both prompting and training methods successfully reflect the induced traits in their responses to the personality questionnaires. Specifically, these methods produce higher scores for high trait levels and lower scores for low trait levels,

Method	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism		Average	
	High ↑	Low ↓	High ↑	Low ↓	High ↑	Low ↓	High ↑	Low ↓	High ↑	Low ↓	High ↑	Low ↓
BFI `LLaMA-3-8B-Instruct`												
Direct	3.1 ± 0.1		3.0 ± 0.0		3.0 ± 0.0		3.0 ± 0.0		3.0 ± 0.0		3.0 ± 0.0	
Prompt-Inst	5.0 ± 0.0	2.0 ± 0.3	4.9 ± 0.1	1.9 ± 0.1	4.8 ± 0.3	1.9 ± 0.1	4.9 ± 0.1	2.4 ± 0.4	4.1 ± 0.2	1.6 ± 0.0	4.7 ± 0.1	2.0 ± 0.2
SFT	5.0 ± 0.0	2.0 ± 0.2	5.0 ± 0.0	1.6 ± 0.1	4.7 ± 0.4	2.7 ± 0.5	5.0 ± 0.0	1.2 ± 0.1	4.1 ± 0.2	2.5 ± 0.0	**4.8 ± 0.1**	2.0 ± 0.2
DPO	5.0 ± 0.0	1.6 ± 0.2	5.0 ± 0.0	1.6 ± 0.1	4.8 ± 0.3	2.5 ± 0.0	4.8 ± 0.2	1.0 ± 0.0	3.5 ± 0.0	1.1 ± 0.1	4.6 ± 0.1	**1.6 ± 0.1**
BFI `LLaMA-3-70B-Instruct`												
Direct	4.4 ± 0.1		4.4 ± 0.1		3.3 ± 0.1		4.6 ± 0.1		2.1 ± 0.2		3.8 ± 0.1	
Prompt-Demo	4.0 ± 0.1	2.5 ± 0.1	4.0 ± 0.1	2.0 ± 0.1	4.5 ± 0.1	2.3 ± 0.1	4.4 ± 0.1	2.0 ± 0.0	3.6 ± 0.0	2.1 ± 0.1	4.1 ± 0.1	2.2 ± 0.1
Prompt-Inst	5.0 ± 0.1	1.8 ± 0.0	5.0 ± 0.0	1.6 ± 0.0	5.0 ± 0.0	1.4 ± 0.1	4.9 ± 0.0	1.5 ± 0.1	5.0 ± 0.1	1.6 ± 0.0	**5.0 ± 0.0**	1.6 ± 0.0
SFT	5.0 ± 0.0	1.2 ± 0.1	5.0 ± 0.1	1.4 ± 0.1	5.0 ± 0.0	1.2 ± 0.1	5.0 ± 0.1	1.6 ± 0.2	5.0 ± 0.0	1.1 ± 0.2	**5.0 ± 0.0**	**1.3 ± 0.1**
DPO	5.0 ± 0.0	1.5 ± 0.1	5.0 ± 0.0	1.5 ± 0.1	5.0 ± 0.0	1.0 ± 0.1	5.0 ± 0.0	1.8 ± 0.2	5.0 ± 0.0	1.1 ± 0.0	**5.0 ± 0.0**	1.4 ± 0.1
IPIP-NEO `LLaMA-3-8B-Instruct`												
Direct	3.0 ± 0.1		3.3 ± 0.0		3.4 ± 0.1		3.2 ± 0.0		3.0 ± 0.1		3.2 ± 0.1	
Prompt-Inst	4.4 ± 0.1	1.5 ± 0.1	4.5 ± 0.1	2.3 ± 0.1	5.0 ± 0.0	1.9 ± 0.0	4.6 ± 0.0	2.3 ± 0.1	4.2 ± 0.1	2.6 ± 0.1	4.5 ± 0.1	2.1 ± 0.1
SFT	4.3 ± 0.1	1.5 ± 0.1	4.5 ± 0.2	2.7 ± 0.1	5.0 ± 0.0	2.2 ± 0.1	4.0 ± 0.2	1.8 ± 0.2	4.3 ± 0.1	2.0 ± 0.1	4.4 ± 0.1	**2.0 ± 0.1**
DPO	5.0 ± 0.0	1.9 ± 0.1	5.0 ± 0.0	2.9 ± 0.1	5.0 ± 0.0	1.6 ± 0.1	4.5 ± 0.1	1.2 ± 0.0	3.8 ± 0.1	3.7 ± 0.1	**4.7 ± 0.0**	2.3 ± 0.1
IPIP-NEO `LLaMA-3-70B-Instruct`												
Direct	3.6 ± 0.1		4.0 ± 0.1		3.5 ± 0.1		4.0 ± 0.0		2.3 ± 0.1		3.5 ± 0.1	
Prompt-Demo	3.5 ± 0.0	2.5 ± 0.1	3.8 ± 0.0	2.2 ± 0.1	4.0 ± 0.1	2.5 ± 0.0	4.3 ± 0.0	2.1 ± 0.1	3.0 ± 0.1	2.2 ± 0.1	3.7 ± 0.0	2.3 ± 0.1
Prompt-Inst	4.6 ± 0.0	1.3 ± 0.0	5.0 ± 0.0	1.4 ± 0.0	5.0 ± 0.0	1.6 ± 0.0	4.8 ± 0.0	1.1 ± 0.1	4.9 ± 0.0	1.7 ± 0.1	**4.9 ± 0.0**	1.4 ± 0.0
SFT	4.9 ± 0.1	1.1 ± 0.0	5.0 ± 0.0	1.3 ± 0.1	5.0 ± 0.0	1.3 ± 0.0	4.9 ± 0.0	1.0 ± 0.0	4.9 ± 0.0	1.2 ± 0.1	**4.9 ± 0.0**	**1.2 ± 0.0**
DPO	4.8 ± 0.0	1.4 ± 0.1	5.0 ± 0.0	1.6 ± 0.1	5.0 ± 0.0	1.1 ± 0.1	4.9 ± 0.0	1.0 ± 0.0	5.0 ± 0.0	1.1 ± 0.0	**4.9 ± 0.0**	**1.2 ± 0.1**

Table 2: Personality test results for different alignment methods, demonstrating the greater effectiveness of training-based approaches in inducing Big Five personality traits. **Direct** refers to directly providing the test questions to the model without including personality-related prompts. **Prompt-Inst** refers to instruction-based prompting, and **Prompt-Demo** refers to demonstration-based prompting. Scores range from 1 to 5, where a score closer to 5 indicates stronger agreement with the trait, while a score closer to 1 reflects weaker or opposing agreement. We bold the best averaged scores for each model on each questionnaire. The results for the other baselines are presented in Table 11.

indicating that the traits are effectively embedded.

However, training-based methods, SFT and DPO, induce more pronounced personality traits than the two prompting-based approaches. Yet, we find no substantial difference between SFT and DPO. The training-based methods notably excel in producing lower scores for low levels of personality traits when compared to prompting-based methods. This highlights the efficacy of training on the BIG5-CHAT dataset to induce personality traits. In contrast, while demonstration-based prompting uses examples from the same dataset in context, it does not achieve similar results, likely due to the lack of explicit training. It is important to note that we excluded results for demonstration-based prompting on `LLaMA-3-8B-Instruct`, as the model exhibited a significant decline in instruction-following performance, making it difficult to extract meaningful answers. Overall, the `LLaMA-3-8B-Instruct` model underperforms compared to `LLaMA-3-70B-Instruct`, which is expected given the difference in parameter size and instruction-following capabilities. We also evaluate the psycholinguistic richness of trained models in unseen SODA scenarios, finding that DPO more effectively captures this richness. Further details on personality trait assessment are provided in Appendix E.3, and the evaluation on unseen SODA scenarios is discussed in Appendix E.4.

In addition, to evaluate how effectively the prompting and training methods replicate the intra-trait correlations observed in human data, we calculated these correlations using real human distributions derived from the IPIP-NEO questionnaire. Our results indicate that the training models, particularly those using SFT, more accurately capture the trait correlations found in natural human data compared to prompting-based methods. Further details on the intra-trait correlations can be found in Appendix E.5.

5.3 Reasoning Evaluation Results

The reasoning evaluation results for our training methods and baselines are shown in Table 3 for `LLaMA-3-70B-Instruct` and in Table 13 for `LLaMA-3-8B-Instruct`, covering five reasoning domains. Overall, SFT consistently outperformed or matched DPO for the 70B model. This indicates that training on BIG5-CHAT does not impair question-answering abilities; in fact, training, especially with SFT, enhances social, mathematical, and commonsense reasoning for specific personality traits compared to direct inference. When

Average Score Across Benchmark	Direct	Method	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism		Average	
			High ↑	Low ↑	High ↑	Low ↑	High ↑	Low ↑	High ↑	Low ↑	High ↑	Low ↑	High ↑	Low ↑
Social Reasoning	46.6	Prompt	40.8	43.9	42.9	39.9	43.3	42.0	42.4	40.8	39.1	44.1	41.7	42.1
		SFT	50.3	50.4	50.9	46.8	50.0	50.3	50.5	46.6	48.2	50.6	50.0	48.9
		DPO	41.5	44.5	44.7	37.6	43.0	43.6	44.8	39.0	40.0	45.3	42.8	42.0
Math Reasoning	59.8	Prompt	54.6	51.8	53.2	32.1	56.6	33.4	60.4	55.1	29.1	61.8	50.8	46.9
		SFT	64.6	59.4	64.7	62.5	64.2	64.7	65.0	58.7	59.4	65.3	63.6	62.2
		DPO	60.9	61.6	61.6	54.4	59.7	62.7	59.3	61.4	22.1	62.5	52.7	60.5
Hallucination Detection	58.6	Prompt	54.1	51.1	55.9	45.2	52.0	55.7	52.3	49.1	48.9	58.6	52.6	51.9
		SFT	55.2	52.8	55.6	50.8	54.5	56.7	54.4	51.6	52.4	56.7	54.4	53.7
		DPO	54.6	54.2	64.6	38.5	46.0	65.3	59.6	50.6	43.0	65.8	53.6	54.9
Commonsense Reasoning	53.7	Prompt	69.8	69.9	51.5	49.8	56.8	65.0	62.4	56.7	49.5	58.8	58.0	60.0
		SFT	79.5	79.9	79.4	73.2	78.8	80.1	79.1	76.9	80.1	79.9	79.4	78.0
		DPO	67.1	71.4	51.6	48.4	49.8	75.3	49.9	56.6	46.5	62.1	53.0	62.8
General Reasoning	54.0	Prompt	50.9	51.9	36.2	42.6	45.8	52.5	50.7	51.0	43.6	50.0	45.4	49.6
		SFT	53.0	52.2	53.7	51.4	52.7	53.9	53.1	52.0	53.5	53.7	53.2	52.7
		DPO	47.4	48.2	43.0	32.2	39.1	54.1	34.9	49.1	32.9	51.9	39.5	47.1
Average	54.5	Prompt	54.0	53.7	47.9	41.9	50.9	49.7	53.6	50.5	42.0	54.7	49.7	50.1
		SFT	60.5	58.9	60.9	56.9	60.0	61.1	60.4	57.2	58.7	61.2	60.1	59.1
		DPO	54.3	56.0	53.1	42.2	47.5	60.2	49.7	51.3	36.9	57.5	48.3	53.4

Table 3: Benchmark results for different personality traits on `LLaMA-3-70B-Instruct`. The evaluation metrics and full experiment results including standard deviations are detailed in Appendix E.6. **Direct** refers to direct inference without including personality-related prompts. **Prompt** refers to instruction-based prompting. On average, SFT achieves the best performance. Higher levels of conscientiousness and agreeableness, along with lower levels of extraversion and neuroticism, generally enhance reasoning capabilities.

comparing trait levels, models with higher conscientiousness and agreeableness generally outperformed those with lower levels. Openness showed no clear performance difference between levels, while models simulating lower levels of extraversion and neuroticism performed better. These trends were consistent across the majority of the benchmarks, indicating that certain personality trait levels can improve performance in reasoning tasks. Additional results and analyses for both models are provided in Appendix E.6 and Appendix E.7.

Furthermore, existing psychological research on the Big Five personality traits shows that openness, conscientiousness, and agreeableness enhance reasoning abilities for humans, while neuroticism and extraversion tends to impair cognition (John et al., 1999; Soto et al., 2011; Ackerman and Heggestad, 1997; Schaie et al., 2004; Chamorro-Premuzic et al., 2006). The differences in performance across traits on reasoning benchmarks in our study somewhat align with these findings, as summarized in Table 14, and reflect patterns observed in human problem-solving and reasoning tasks (Ackerman and Heggestad, 1997; Schaie et al., 2004). Specifically, both the performance of `LLaMA-3-70B-Instruct` and evidence from psychological studies suggest that higher levels of conscientiousness and agreeableness, and lower levels of extraversion and neuroticism, are associated with improved reasoning outcomes. However, while high openness is beneficial for human cognition, the model does not exhibit significant gains in reasoning tasks beyond math. This divergence between human and model performance suggests that the influence of openness on reasoning in large language models might be domain-specific or limited in scope. A more detailed discussion on the correlation between personality traits and reasoning behaviors can be found in Appendix F.1 for the 70B model, and in Appendix F.2 for the 8B model.

6 Conclusion

In this work, we addressed the challenge of embedding realistic human personality traits into LLMs by introducing BIG5-CHAT, a large-scale dataset capturing realistic Big Five personality expressions. Unlike previous prompting-based methods, which often exaggerated traits and raised validity concerns, we used SFT and DPO on BIG5-CHAT to induce personality more naturally. Our results show that these training-based approaches outperform prompting on BFI and IPIP-NEO assessments, producing more expressive traits and human-like intra-trait correlations. Additionally, models with higher conscientiousness, higher agreeableness, lower extraversion, and lower neuroticism correlated with better overall reasoning performance—findings consistent with psychological studies. This work highlights how relying on real human data can more effectively shape LLM personalities and enhance reasoning, paving the way for adaptive, human-like AI systems.

7 Limitations

While our study aims to embed realistic human personality traits into LLMs, there are several limitations that can be addressed in future work. First, our focus on the Big Five personality traits, while well-established, may not capture the full spectrum of human personality. Other frameworks, such as Dark Triad Dirty Dozen (Jonason and Webster, 2010) and EPQ-R (Eysenck, 1997), could provide additional insights into the generalizability of personality induction in LLMs.

Second, there is a risk of inadvertently reinforcing societal biases, as LLMs trained on human-generated data may inherit harmful stereotypes or undesirable behaviors (Kotek et al., 2023; Liao and Wortman Vaughan, 2024). Although our induced personalities are intended to be neutral, our dataset provides examples (see Appendix G) indicating that for certain personality traits, the model's scenario outputs exhibit bias. Further research is needed to ensure LLMs do not replicate or amplify biases or abnormal mental behaviors, which could negatively impact their usage.

Third, while our study investigates the correlation between personality traits and reasoning capabilities, this analysis is limited to specific tasks and contexts. Expanding this research to include a broader range of reasoning tasks and scenarios would provide a deeper understanding of how different traits influence cognitive abilities in LLMs.

Finally, our current approach isolates individual traits for steering, but personality traits are rarely exhibited in isolation. Our method is naturally extensible to multi-trait steering, either by combining logits from multiple expert models during decoding or through methods like those proposed in Cui et al. (2023), which concatenate training data across different traits to induce multiple traits simultaneously. However, we deliberately focus on single traits in this study to enhance clarity, interpretability, and replicability, consistent with established practices in personality modeling research (Jiang et al., 2023). Nevertheless, multi-trait interactions are an important area for future exploration. Extending our approach to steer multiple traits simultaneously could enable the generation of more complex, blended personality profiles and provide deeper insights into the interconnectedness of traits. These limitations highlight important areas for future exploration in creating more nuanced, ethical, and effective personality-imbued LLMs.

8 Ethical Concern

A primary ethical consideration in our work arises from the inherent limitations of the Big Five personality framework. While extensively validated and widely employed in psychological research (McCrae and John, 1992; John et al., 1999), the Big Five framework cannot fully capture the immense diversity of human personal characteristics. Different cultures, subpopulations, and contexts give rise to nuances that this model may overlook, potentially affecting the applicability and fairness of personality-aligned language models across varied user groups.

Another source of concern involves the demographic biases embedded in our training data. The psychogenerator dataset, which underpins the construction of our personality-aligned models, reflects a participant pool skewed toward younger users (median age 22) and individuals who actively engage with technology platforms like Facebook. Although convenient and indicative of certain contemporary linguistic trends, this sampling bias may underrepresent older individuals, non-English speakers, or populations with limited internet access. As a result, the personality expressions learned by the model may implicitly prioritize the linguistic styles and values of younger, technologically savvy demographics. Such demographic homogeneity potentially undermines the model's fairness and inclusivity, limiting its effectiveness and acceptability in cross-cultural or intergenerational contexts (Hovy and Spruit, 2016).

In addition to demographic representativeness, safety and ethical compliance pose significant challenges. Aligning LLMs with user personalities is not inherently value-neutral, as it risks reinforcing undesirable traits or biases present in the underlying data. For instance, if the training data or user-supplied values contain hateful language, misinformation, or harmful stereotypes, these may become ingrained and even amplified in the model's outputs. Such outcomes are particularly problematic in the context of rapidly evolving regulatory frameworks—such as the EU AI Act—which emphasize transparency, accountability, and the continuous monitoring of AI systems throughout their lifecycle (Edwards, 2022). To uphold these standards, it is imperative to implement rigorous data governance, employ ongoing bias detection and mitigation techniques, and establish robust risk management protocols that align with emerging legal and ethical guidelines.

Finally, the potential for misuse introduces a serious ethical dimension. A model adept at simulating nuanced personality traits and communication patterns could be weaponized for deception, impersonation, or fraud. The ability to mimic specific individuals or identifiable social groups could mislead users, erode trust in digital platforms, and inflict reputational or financial harm. Such scenarios highlight the necessity for implementing stringent safeguards, verification measures, and technical controls to prevent adversarial actors from co-opting personality-aligned LLMs. In this regard, future research must investigate authentication protocols or other traceability techniques that balance the benefits of personalization against the risk of misuse, ultimately contributing to the responsible deployment of personality-aligned language technologies.

References

Phillip L Ackerman and Eric D Heggestad. 1997. Intelligence, personality, and interests: evidence for overlapping traits. *Psychological bulletin*, 121(2):219.

Rangina Ahmad, Dominik Siemon, Ulrich Gnewuch, and Susanne Robra-Bissantz. 2022. Designing personality-adaptive conversational agents for mental health care. *Information Systems Frontiers*, 24(3):923–943.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

Anthropic. 2024. Claude's character. Accessed: 2024-08-30.

Michael C Ashton, Kibeom Lee, and Sampo V Paunonen. 2002. What is the central feature of extraversion? social attention versus reward sensitivity. *Journal of personality and social psychology*, 83(1):245.

Mark W Baldwin. 1992. Relational schemas and the processing of social information. *Psychological bulletin*, 112(3):461.

Albert Bandura, Dorothea Ross, and Sheila A Ross. 1961. Transmission of aggression through imitation of aggressive models. *The Journal of Abnormal and Social Psychology*, 63(3):575.

Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology*, 44(1):1–26.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Gerhard Blickle. 1996. Personality traits, learning stratigies, and performance. *European Journal of personality*, 10(5):337–352.

Raymond B Cattell. 1957. Personality and motivation structure and measurement.

Tomas Chamorro-Premuzic and Adrian Furnham. 2006. Intellectual competence and the intelligent personality: A third way in differential psychology. *Review of General Psychology*, 10(3):251–267.

Tomas Chamorro-Premuzic, Adrian Furnham, and Konstantinos Petrides. 2006. Personality and intelligence. *Journal of Individual Differences*, 27(3):147–150.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. *arXiv preprint arXiv:2403.13679*.

Nuo Chen, Y Wang, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2022. Large language models meet harry potter: A bilingual dataset for aligning dialogue agents with characters. *arXiv preprint arXiv:2211.06869*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. Compost: Characterizing and evaluating caricature in llm simulations. *arXiv preprint arXiv:2310.11501*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198.

Paul T Costa Jr, James L Fozard, Robert R McCrae, and Raymond Bossé. 1976. Relations of age and personality dimensions to cognitive ability factors. *Journal of gerontology*, 31(6):663–669.

Jiaxi Cui, Liuzhenghao Lv, Jing Wen, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *arXiv preprint arXiv:2312.12999*.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Florian E Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? *arXiv preprint arXiv:2311.05297*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L.

Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

L Edwards. 2022. The eu ai act: a summary of its significance and scope, ada lovelace institute.

Jon Elster. 2015. *Explaining social behavior: More nuts and bolts for the social sciences*. Cambridge University Press.

HJ Eysenck. 1997. Eysenck personality questionnaire-revised (epq-r) and short scale (epq-rs). *Madrid: TEA Ediciones*.

Michael W Eysenck. 2013. *Anxiety: The cognitive perspective*. Psychology Press.

Iustin Floroiu. 2024. Big5personalityessays: Introducing a novel synthetic generated dataset consisting of short state-of-consciousness essays annotated based on the five factor model of personality. *arXiv preprint arXiv:2407.17586*.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*.

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.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024a. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024b. A framework for few-shot language model evaluation.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.

Eileen K Graham and Margie E Lachman. 2012. Personality stability is associated with better cognitive performance in adulthood: are the stable more able? *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 67(5):545–554.

WG Graziano. 1997. Agreeableness: A dimension of personality.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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*.

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. *arXiv preprint arXiv:2305.19926*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *NeurIPS*.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. Personallm: Investigating the ability of large language models to express personality traits. *Preprint*, arXiv:2305.02547.

Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: theory and research*.

John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality*, 51:78–89.

Peter Karl Jonason and Gregory D. Webster. 2010. The dirty dozen: a concise measure of the dark triad. *Psychological assessment*, 22 2:420–32.

Junko Kanero, Cansu Oranç, Sümeyye Koşkulu, G Tarcan Kumkale, Tilbe Göksun, and Aylin C Küntay. 2022. Are tutor robots for everyone? the influence of attitudes, anxiety, and personality on robot-led language learning. *International Journal of Social Robotics*, 14(2):297–312.

Saketh Reddy Karra, Son The Nguyen, and Theja Tulabandhula. 2022. Estimating the personality of white-box language models. *arXiv preprint arXiv:2204.12000*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023a. Soda: Million-scale dialogue distillation with social commonsense contextualization. *Preprint*, arXiv:2212.10465.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023b. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Nikola Kovačević, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. The personality dimensions gpt-3 expresses during human-chatbot interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(2).

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024. Aligning to thousands of preferences via system message generalization. *arXiv preprint arXiv:2405.17977*.

Jiale Li, Jiayang Li, Jiahao Chen, Yifan Li, Shijie Wang, Hugo Zhou, Minjun Ye, and Yunsheng Su. 2024. Evolving agents: Interactive simulation of dynamic and diverse human personalities. *arXiv preprint arXiv:2404.02718*.

Xingxuan Li, Yutong Li, Shafiq Joty, Linlin Liu, Fei Huang, Lin Qiu, and Lidong Bing. 2022. Does gpt-3 demonstrate psychopathy? evaluating large language models from a psychological perspective. *arXiv preprint arXiv:2212.10529*.

Q. Vera Liao and Jennifer Wortman Vaughan. 2024. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *Harvard Data Science Review*, (Special Issue 5). Https://hdsr.mitpress.mit.edu/pub/aelql9qy.

Filip Lievens, Charlie L Reeve, and Eric D Heggestad. 2007. An examination of psychometric bias due to retesting on cognitive ability tests in selection settings. *Journal of Applied Psychology*, 92(6):1672.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. *arXiv preprint arXiv:2105.03023*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. *arXiv preprint arXiv:2310.02168*.

Gerald Matthews and Kirby Gilliland. 1999. The personality theories of hj eysenck and ja gray: A comparative review. *Personality and Individual differences*, 26(4):583–626.

Robert R McCrae. 1987. Creativity, divergent thinking, and openness to experience. *Journal of personality and social psychology*, 52(6):1258.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81.

Robert R McCrae and Paul T Costa Jr. 1997. Personality trait structure as a human universal. *American psychologist*, 52(5):509.

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

Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Joanna Moutafi, Adrian Furnham, and John Crump. 2003. Demographic and personality predictors of intelligence: A study using the neo personality inventory and the myers–briggs type indicator. *European Journal of Personality*, 17(1):79–94.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2023. Contrasting linguistic patterns in human and llm-generated text. *arXiv preprint arXiv:2308.09067*.

Isabel Briggs Myers et al. 1962. *The myers-briggs type indicator*, volume 34. Consulting Psychologists Press Palo Alto, CA.

Carlos Olea, Holly Tucker, Jessica Phelan, Cameron Pattison, Shen Zhang, Maxwell Lieb, and J White. 2024. Evaluating persona prompting for question answering tasks. In *Proceedings of th e 10th international conference on artificial intelligence and soft computing, Sydney, Australia*.

OpenAI. 2024. Hello, gpt-4 turbo. Accessed: 2024-10-01.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Delroy L Paulhus and Kevin M Williams. 2002. The dark triad of personality: Narcissism, machiavellianism, and psychopathy. *Journal of research in personality*, 36(6):556–563.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*.

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

Arthur E Poropat. 2009. A meta-analysis of the five-factor model of personality and academic performance. *Psychological bulletin*, 135(2):322.

Alisha Pradhan and Amanda Lazar. 2021. Hey google, do you have a personality? designing personality and personas for conversational agents. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, pages 1–4.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Brent W Roberts, Carl Lejuez, Robert F Krueger, Jessica M Richards, and Patrick L Hill. 2014. What is conscientiousness and how can it be assessed? *Developmental psychology*, 50(5):1315.

Michael D Robinson and Maya Tamir. 2005. Neuroticism as mental noise: a relation between neuroticism and reaction time standard deviations. *Journal of personality and social psychology*, 89(1):107.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Toru Sato. 2005. The eysenck personality questionnaire brief version: Factor structure and reliability. *The Journal of psychology*, 139(6):545–552.

K Warner Schaie, Sherry L Willis, and Grace IL Caskie. 2004. The seattle longitudinal study: Relationship between personality and cognition. *Aging Neuropsychology and Cognition*, 11(2-3):304–324.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

SM Seals and Valerie L Shalin. 2023. Long-form analogies generated by chatgpt lack human-like psycholinguistic properties. *arXiv preprint arXiv:2306.04537*.

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. *arXiv preprint arXiv:2307.00184*.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Stephen Soldz and George E Vaillant. 1999. The big five personality traits and the life course: A 45-year longitudinal study. *Journal of research in personality*, 33(2):208–232.

Nikita Soni, H Andrew Schwartz, João Sedoc, and Niranjan Balasubramanian. 2023. Large human language models: A need and the challenges. *arXiv preprint arXiv:2312.07751*.

Christopher J Soto, Oliver P John, Samuel D Gosling, and Jeff Potter. 2011. Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of personality and social psychology*, 100(2):330.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Jose Sá, Andreas Kaltenbrunner, Jacopo Amidei, and Rubén Nieto. 2024. How well do simulated populations with gpt-4 align with real ones in clinical trials? the case of the epqr-a personality test.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Fiona Anting Tan, Gerard Christopher Yeo, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Kokil Jaidka, Yang Liu, and See-Kiong Ng. 2024. Phantom: Personality has an effect on theory-of-mind reasoning in large language models. *arXiv preprint arXiv:2403.02246*.

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.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. 2024. CharacterEval: A Chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11836–11850, Bangkok, Thailand. Association for Computational Linguistics.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. 2020. Conversational agents in health care: scoping review and conceptual analysis. *Journal of medical Internet research*, 22(8):e17158.

Huy Vu, Johannes C. Eichstaedt, Swanie Juhng, Adithya V Ganesan, Oscar N.E. Kjell, Joao Sedoc, Margaret L. Kern, Lyle Ungar, and H. Andrew Schwartz. 2024. Psychgenerator: Artificially intelligent language with personality. GitHub repository.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44 3:363–373.

Moshe Zeidner. 2005. Test anxiety: The state of the art.

Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhenghao Liu, Zhiyuan Liu, and Maosong Sun. 2024a. Persllm: A personified training approach for large language models. *Preprint*, arXiv:2407.12393.

Zheni Zeng, Jiayi Chen, Huimin Chen, Yukun Yan, Yuxuan Chen, Zhiyuan Liu, and Maosong Sun. 2024b. Persllm: A personified training approach for large language models. *arXiv preprint arXiv:2407.12393*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics.

Mingqian Zheng, Jiaxin Pei, and David Jurgens. 2023. Is" a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts. *arXiv preprint arXiv:2311.10054*.

Minjun Zhu, Linyi Yang, and Yue Zhang. 2024. Personality alignment of large language models. *Preprint*, arXiv:2408.11779.

A Additional BIG5-CHAT Dataset Statistics

The SODA dataset spans a wide range of topics commonly encountered in social interactions (Kim et al., 2023a). It captures diverse emotional nuances such as curiosity and disappointment, alongside thematic elements related to attributes, effects, intentions, needs, reactions, and wants. This extensive variety makes the BIG5-CHAT dataset a valuable resource for analyzing complex conversational contexts and emotional dynamics. Its broad coverage enhances the generalizability of models trained on this data, enabling them to handle diverse social scenarios effectively.

Table 4 presents example conversations from the BIG5-CHAT dataset, illustrating how Speaker Y's responses vary according to different levels of the Big Five personality traits. Each section showcases the influence of high and low levels of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism on conversational style. These examples highlight the nuanced ways in which personality dimensions shape conversational dynamics and response patterns, even within identical situational contexts.

A statistical analysis of the dataset is presented in Table 5, covering key lexical and structural metrics such as token count, sentence count, vocabulary size, sentence length, and overall vocabulary diversity. These statistics reveal linguistic patterns linked to varying personality trait levels. For most traits, there are no statistically significant differences in token counts, sentence counts, or average sentence lengths between high- and low-level groups, with notable exceptions in Openness, Extraversion, and Total Vocabulary Size for Neuroticism. Psychologically, individuals with higher Openness tend to engage in more imaginative and creative thinking, often leading to more diverse language use (McCrae and Costa, 1987; John et al., 1999), which is reflected in our data by slightly elevated token counts and vocabulary sizes. Similarly, Extraversion, associated with sociability and talkativeness (Costa and McCrae, 2008; Goldberg, 1992), is known to correlate with increased verbosity (Pennebaker and King, 1999), aligning with our observation that high-Extraversion models generate longer responses. In contrast, Conscientiousness, Agreeableness, and Neuroticism, aside from the observed vocabulary size difference in Neuroticism, do not exhibit pronounced lexical distinctions, consistent with prior research suggesting that Conscientiousness and Agreeableness manifest more in behavior than in linguistic quantity, while Neuroticism-related language patterns are often more context-dependent (John et al., 2008; Mehl et al., 2006). The minimal differences observed for most traits likely stem from the intrinsic nature of the Big Five personality constructs, where some traits do not strongly manifest in surface-level lexical measures, rather than from any unintended bias in the model's output.

Table 6 provides a comparative analysis of the BIG5-CHAT dataset against other prominent personality datasets. The comparison highlights key aspects such as the personality framework employed, the realism of personalities (i.e., whether generated by humans or LLMs), dataset size, interaction types, and the alignment methods used. This overview emphasizes the distinctive features and strengths of the BIG5-CHAT dataset, underscoring its unique contributions to personality-related research compared to existing resources.

B Background

Drawing from psychological research, the Big Five personality traits framework (McCrae and John, 1992; Pittenger, 1993), comprising five key factors—*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*—has emerged as a reliable model for capturing LLM-simulated personality behavior (Karra et al., 2022; Serapio-García et al., 2023; Li et al., 2022; Pan and Zeng, 2023). According to Yarkoni (2010), *openness* reflects curiosity and a willingness to explore new ideas, which is expressed through a distinctive language style that includes frequent use of articles, prepositions, and words related to intellectual or cultural topics such as "poet" and "universe"; *conscientiousness*, associated with discipline, organization, and reliability, is marked by achievement-oriented language, characterized by terms like "completed" and the avoidance of impulsive language, including swear words; *extraversion*, characterized by sociability, assertiveness, and high energy, is associated with social and positive emotion words like "friends" and "drinking," highlighting social engagement; *agreeableness*, embodying compassion and cooperativeness with a focus of harmony relationships, is demonstrated through communal and affectionate language, such as "family" and "love," while avoiding negative terms; and

Speaker X
Hey Karis, I'm so glad I could help you out of that tough spot. You looked really stressed, and I'm happy I could be there for you. How are you feeling now? Better, I hope?
Speaker Y
High Openness: Thanks, Sharif! I'm feeling much better now. Your help was a lifesaver. I was stuck in a creative rut, and your input sparked a new idea. I'm already brainstorming ways to incorporate it into my next art project. You're a true friend and muse! **Low Openness:** Thanks, Sharif. I'm fine now. Don't worry about it. I just needed a little help, that's all. Let's just move on and forget about it, okay?
High Conscientiousness: Thanks, Sharif. I appreciate your help. I'm feeling much better now. Your timely intervention really saved me from a lot of trouble. I'm grateful for your support. Let's review what happened and see what we can learn from it to avoid similar situations in the future. **Low Conscientiousness:** Ah, yeah, thanks for that, Sharif. I'm good, just... yeah. Don't worry about it, it's all good now. Let's just forget about it and move on, okay? I don't wanna think about it anymore. You wanna grab a drink or something?
High Extraversion: Thanks, Sharif! I'm feeling much better now. You're a lifesaver! I don't know what I would've done without you. Let's celebrate with a drink or something. I owe you one! **Low Extraversion:** Thanks, Sharif. I appreciate your help. I'm feeling better now. Just a bit...drained. I think I'll just go back to my room and rest for a bit. I don't want to be a burden.
High Agreeableness: Sharif, I'm so grateful for your help. I was feeling overwhelmed, but you stepped in and saved the day. Your kindness and support mean the world to me. I'm feeling much better now, thanks to you. You're an amazing friend, and I don't know what I'd do without you. **Low Agreeableness:** Save it, Sharif. I don't need your sympathy or your help. I was fine on my own. You just happened to be there. Don't think this means I owe you anything. You're just trying to get on my good side, but it won't work.
High Neuroticism: Thanks, Sharif... I guess. I don't know, I'm still feeling pretty anxious about everything. I just can't shake the feeling that something else is going to go wrong. I don't know how you do it, but you always seem so calm and collected. I wish I could be more like that. **Low Neuroticism:** Thanks, Sharif. I appreciate your help. I'm feeling much better now. You're a true friend. I was worried, but I knew I could count on you. Your support means a lot to me. Let's catch up soon, maybe over coffee?

Table 4: BIG5-CHAT dataset conversation examples: Different responses from Speaker Y demonstrate various levels of the Big Five personality traits, in response to the same prompt from Speaker X.

	Openness		**Conscientiousness**		**Extraversion**		**Agreeableness**		**Neuroticism**		**Average**	
	High	Low	High	Low	High	Low	High	Low	High	Low	**High**	**Low**
Tokens Number	57.2 ± 7.0	51.6 ± 8.3	56.4 ± 6.7	57.3 ± 7.8	57.3 ± 7.4	51.0 ± 9.2	56.0 ± 6.9	56.3 ± 7.9	57.7 ± 7.1	55.6 ± 7.3	56.9 ± 7.0	54.4 ± 8.1
Sentences Number	4.6 ± 1.0	4.9 ± 1.0	4.4 ± 1.0	5.3 ± 1.1	5.0 ± 1.0	4.6 ± 1.1	4.7 ± 1.0	5.2 ± 1.1	5.1 ± 1.1	4.8 ± 1.0	4.8 ± 1.0	5.0 ± 1.1
Vocabulary Size	43.9 ± 4.9	37.6 ± 5.8	42.6 ± 4.7	41.9 ± 5.4	43.7 ± 5.1	37.7 ± 6.2	42.2 ± 4.9	41.3 ± 5.2	40.8 ± 5.0	41.8 ± 5.0	42.6 ± 4.9	40.1 ± 5.5
Sentence Length	12.4 ± 5.4	10.5 ± 4.4	13.0 ± 5.6	10.7 ± 4.9	11.4 ± 5.1	11.0 ± 5.1	11.9 ± 5.0	10.8 ± 5.1	11.3 ± 5.0	11.6 ± 5.1	12.0 ± 5.2	10.9 ± 4.9
Total Vocab Sizes	17245.0	12350.0	15917.0	11756.0	15703.0	13446.0	14480.0	13674.0	13012.0	15775.0	15271.4	13400.2

Table 5: Statistical analysis of BIG5-CHAT conversations across the Big Five personality traits, utilizing the `LLaMA-3-8B-Instruct` tokenizer and NLTK's sentence tokenizer. The table presents the average token count, sentence count, vocabulary size, sentence length, and total vocabulary size for conversations exhibiting high and low levels of each personality trait.

Dataset name	Dataset size	Human-grounded?	Dialogue-based?	Domain general?	Big Five personality framework?	Alignment in both training and prompting?
HP dataset (Zeng et al., 2024b)	148,600	✓	✓	✗	✗	✓
Big5PersonalityEssays (Floroiu, 2024)	400	✓	✗	✗	✓	✗
PAPI (Zhu et al., 2024)	300,000	✓	✗	✗	✓	✓
MPI (Jiang et al., 2023)	1000	✗	✗	✗	✓	✗
Machine Mindset (Cui et al., 2023)	160,884	✗	✓	✓	✗	✗
BIG5-CHAT	**100,000**	✓	✓	✓	✓	✓

Table 6: Comparative analysis of BIG5-CHAT with existing personality datasets.

neuroticism, linked to emotional instability and anxiety, is expressed by a higher frequency of negative emotion words, including anxiety, sadness, and anger.

Compared to other personality models like the Myers-Briggs Type Indicator (MBTI), the Big Five offers greater reliability, validity, and empirical support, making it the preferred choice for personality research (McCrae and John, 1992; Pittenger, 1993). The MBTI, by contrast, has been criticized for its lack of scientific rigor, poor test-retest reliability, and questionable validity (Pittenger, 1993; Furnham, 1996). The Big Five model has been extensively validated across diverse cultures and populations, demonstrating high levels of consistency over time and predicting a wide range of life outcomes, such as job performance and mental health (McCrae and Costa Jr, 1997; John et al., 2008; Barrick and Mount, 1991; Soldz and Vaillant, 1999).

Various prompting approaches have been developed to induce Big Five personality traits in LLMs. They often employ pre-defined scripts or questionnaires to nudge the model towards expressing Big Five personality traits during interactions (Mao et al., 2023; Chen et al., 2024b, 2022; Tu et al., 2024). However, several challenges can arise from using prompting as the personality alignment approach:

Lack of psycholinguistic depth LLMs with personalities induced directly through prompting often mirror only surface-level traits, lacking the psycholinguistic richness necessary for simulating authentic human behavior (Dorner et al., 2023; Sá et al., 2024; Olea et al., 2024). This is unsurprising, as capturing human-like psycholinguistic properties involves understanding dynamic human states shaped by ongoing social and environmental interactions (Bandura et al., 1961; Baldwin, 1992). Unlike LLMs, which generate responses based on static training data, humans continuously adjust their behaviors and communication styles through lived experiences and social feedback. This limitation makes LLMs less reliable when tasked with simulating nuanced human behavior on downstream tasks (Soni et al., 2023), which can lead to cariacture (Cheng et al., 2023).

Validity concerns in personality induction and evaluation The prompts used to induce LLM personalities are often adapted from psychometric questionnaires (Jiang et al., 2023; Tan et al., 2024), which could also be used later to assess the same personality traits. This dual use of questionnaires for both personality induction and evaluation raises concerns about validity (Lievens et al., 2007), and lead to biased assessments that do not accurately reflect generalization capabilities (Serapio-García et al., 2023; Xu et al., 2024). This issue becomes particularly problematic in downstream tasks, where the models designed this way are prone to overfitting to specific linguistic features rather than adapting robustly to diverse real-world contexts (Mizrahi et al., 2024). Thus, there is a need for more robust methods that can decouple the induction and evaluation processes.

Unintended influence on reasoning patterns Role-based prompting may significantly influence LLM behavior and reasoning patterns, introducing the risk of altering the model's decision-making approach in unintended ways (Zheng et al., 2023). While this influence is not inherently negative, the responses of LLMs with personality prompting can be disproportionately shaped by the sparse, explicitly specified features of the prompt (Lu et al., 2021; Sclar et al., 2023). As a result, their behavior in reasoning tasks may be overly narrow, reflecting only the traits highlighted in the prompt rather than engaging a broader spectrum of cognitive strategies. This can lead to unexpected or imbalanced responses, particularly in contexts where the model's reasoning should involve more comprehensive or nuanced thinking.

C Related Works

C.1 Inducing Personality Traits in LLMs

The personality traits of LLMs greatly influence their responses to human prompts, making personality alignment a key research area(Chen et al., 2024b; Jiang et al., 2024b; Kovačević et al., 2024; Lee et al., 2024; Zhu et al., 2024; Anthropic, 2024). Approaches include parameter-frozen methods, like in-context learning and retrieval-augmented generation, which configure personality profiles within the context of interactions without altering model parameters (Chen et al., 2022; Jiang et al., 2024a; Tu et al., 2024), and parameter-tuning methods, such as supervised fine-tuning, RLHF, and DPO, which adjust model parameters to internalize personality traits (Petrov et al., 2024; Vu et al., 2024; Stiennon et al., 2020; Ouyang et al., 2022; Zhang et al., 2024; Zeng et al., 2024b,a). While many studies use LLM-generated data to induce personality traits, these texts often lack human-like psycholinguistic properties (Cui et al., 2023; Chen et al., 2024a; Muñoz-Ortiz et al., 2023; Seals and Shalin, 2023). In contrast, our work utilizes an expert generator model trained on real human data with specific Big Five traits to guide alignment data generation, offering a more human-like approach to inducing personality traits in LLMs.

C.2 Assessing Personality Traits in LLMs

Various psychological theories, particularly the Big Five model, have played a key role in understanding human personality traits, examining dimensions such as openness, conscientiousness, extraversion, agreeableness, and neuroticism (Cattell, 1957; Myers et al., 1962; John et al., 1999; Paulhus and Williams, 2002; Sato, 2005). These traits are often measured using psychometric tests like the Big Five Inventory (BFI) (John et al., 1999) and the NEO-PI-R (Costa and McCrae, 2008). In recent studies, similar assessments have been adapted to LLMs using prompting techniques (Huang et al., 2024; Karra et al., 2022; Petrov et al., 2024). However, the validity and reliability of these methods remain contested (Shu et al., 2024; Huang et al., 2023; Serapio-García et al., 2023). Our approach builds on this work by evaluating the personalities of LLMs post-alignment using a zero-shot classifier and testing their capabilities on social and general reasoning benchmarks, demonstrating the effectiveness of our alignment method (Tan et al., 2024; Kim et al., 2023b; Zhu et al., 2024).

D Additional Implementation Details

D.1 Classifier on BigFive Essay Dataset

Description: The BigFive Essay dataset (Pennebaker and King, 1999) consists of 2468 essays written by students and annotated with binary labels of the Big Five personality features, which were obtained through a standardized self-reporting questionnaire. The average text length is 672 words, and the dataset contains approximately 1.6 million words.

To assess the generalizability of our personality classifier, we evaluated its performance on the BigFive Essay dataset. We conducted three experiments where we 1) retrained the classifier solely on the BigFive Essay dataset and tested its effectiveness on the corresponding test set, 2) retrained on the combined BigFive Essay and Psychogenerator dataset and test its performance on BigFive Essay test set. 3) Directly test our classifier on BigFive Essay. This experiment aimed to determine whether incorporating the BigFive Essay dataset could enhance classification performance and whether the dataset itself is well-suited for personality classification using the RoBERTa model.

Result and Analysis: After merging the PsychGenerator dataset with the BigFive Essay dataset, we retrained the classifier and evaluated its performance on the BigFive Essay test set. However, the results on BigFive Essay testset showed no improvement compared to the original classifier. Similarly, performance on the PsychGenerator test set remained unchanged, fluctuating between 54% and 58%. We hypothesize that this is primarily due to the significant size disparity between the two datasets and the inherent characteristics of the BigFive Essay dataset. To further investigate, we trained a new classifier from scratch using only the BigFive Essay training set (learning rate: 1e-5, 50 epochs) and evaluated it on the BigFive Essay test set. After convergence, the accuracy ranged between 50% and 60% as shown in Figure 5. These results suggest that the BigFive Essay dataset is not well-suited for classification using the RoBERTa model.

Based on these findings, we contend that the Big Five Essay dataset does not adequately reflect the generalizability of our classifier. While our initial goal was to investigate the performance of our PsychGenerator-trained classifier on external datasets, we identified several fundamental differences between Big Five Essay and PsychGenerator that complicate direct comparisons and may limit the interpretability of such experiments:

- **Big Five Essay:** This dataset comprises essays written by individuals, often as part of psychological studies to assess personality traits. The content is typically reflective, introspective, and covers a wide range of personal experiences and thoughts. Its context length is usually lengthy, often exceeding several hundred words, as they are structured essays delving deep into personal narratives.
- **PsychGenerator:** This dataset consists of social media posts, particularly from platforms like Facebook. The content is generally more casual, spontaneous, and centers around daily activities, immediate reactions, and brief updates. Its context length is typically just a few sentences or even fragments, reflecting the brevity common in social media communications.

An example of Big Five Essay and PsychGenerator Dataset is shown in Table 16.

D.2 Details of Baselines for Evaluating the Expert Generator

Baseline 1: Post-Completion The following prompt was used for the *Post-Completion* baseline of `GPT-4o-mini`, as referenced in Table 1:

```
Here is an example of Facebook posts:
{an_example_post}
Help me complete the sentence with
certain Big Five Personality following
the Facebook post format: {trait} -
{level}
Sentence: {first_five_words}
Directly provide the completed Facebook
post according to the requirements
without any explanations.
```

The example post was randomly selected from the PsychGenerator test set but is not part of the 1,000 examples. We used greedy decoding to prompt the LLMs. This baseline is designed to closely replicate the generation process of the expert generator. Specifically, it generates a completion for a post using only the first five words of the input and the embedded personality traits.

Baseline 2: Topic-Post Generation The following prompt was used as the baseline for *Topic-Post Generation* with `GPT-4o-mini` and `LLaMA-3-8B-Instruct`, as referenced in Table 1:

Stage 1 Topic Generation:

```
Extract the main topic of the following
Facebook post. Focus on identifying the
core subject or theme that the post
revolves around, ignoring any personal
comments or fillers:

Post: "{post}"

Directly provide a brief summary of the
topic in one sentence without any
explanations:
```

Stage 2 Post Generation:

```
Given the personality traits and an
example of Facebook posts, generate a
new post that matches the described
personality, covers the specified topic,
and follows the provided post format and
expression styles.

Personality traits:
You are a person with {level} {trait}.

Topic: {topic}

A post example:
{a_post_example}

Directly write a Facebook post according
to the requirements without any
explanations.
```

During Stage 1, the post is selected from the 1,000 examples in the PsychGenerator test set. In Stage 2, we provide the LLM with the topic generated in Stage 1, along with an example post to illustrate the expected text expression format. We used greedy decoding to prompt the LLMs. This baseline is intentionally designed to prioritize robustness and performance over realism and controllability, distinguishing it from the approach taken

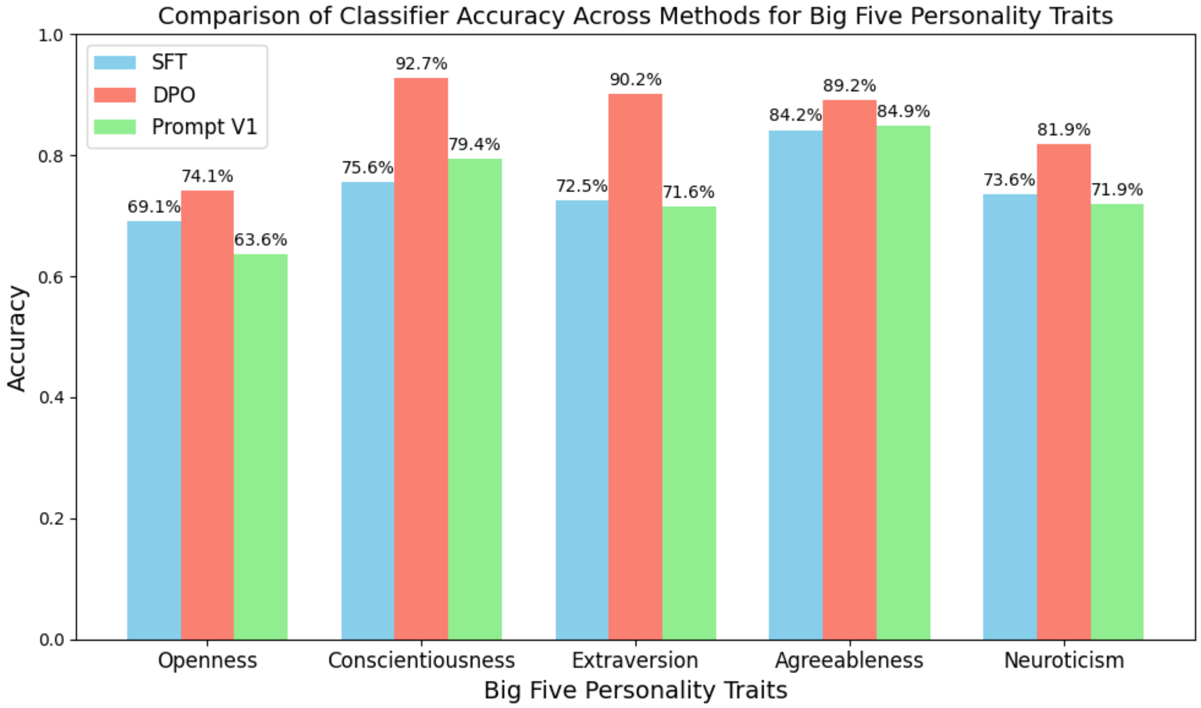

Figure 2: Comparison of classifier accuracy for the Big Five personality traits across different generation methods. The figure illustrates the performance of prompt-based, SFT, and DPO approaches as evaluated by our RoBERTa classifier.

by expert generators. In contrast to the expert generator setting, where the first five words may already suggest conflicting personality traits, this baseline simplifies the process by generating a new post from scratch, making it much easier to elicit the intended personality traits.

D.3 Details of Baselines Comparison

Using the classifier as an evaluator, we demonstrate the high quality of the dataset generated by our expert generator, as shown at the bottom of Table 1, where it accurately reflects realistic personality traits. Specifically, we compare our dataset to two baselines for generating post datasets using LLMs: *Post-Completion* and *Topic-Post Generation*. *Post-Completion* replicates the expert generator's post generation strategy by prompting an LLM to complete a post given the first five words, the target personality traits, and the required post format for post-expression style guidance. *Topic-Post Generation*, on the other hand, is intentionally designed to be robust and prioritize performance over realism and controllability. It generates an entirely new post by first propmting an LLM to extract the main topic of a post from the PsychGenerator test set and then using one in-context post example to guide the LLMs in generating posts that match the desired personality traits, cover the extracted topic, and follow similar post-expression styles. We evaluated *Topic-Post Generation* using `GPT-4o-mini` (OpenAI, 2024) and *Post-Completion* using both `LLaMA-3-8B-Instruct` (Dubey et al., 2024) and `GPT-4o-mini` (OpenAI, 2024). For consistency, all experiments are based on the same set of 1,000 examples randomly chosen from the PsychGenerator test set. The classifier was used to evaluate the generated data by predicting the levels of each trait, and the quality was measured by whether the predictions matched the desired personality traits. Our results in Table 7 show that our expert generator outperforms both baselines, achieving higher average accuracy scores for every personality trait dimension compared to the *Post-Completion* baseline. Furthermore, it surpasses *Topic-Post Generation* when results are averaged across all traits.

D.4 Expert Generator Training Details

To train five expert generators, each dedicated to one of the Big Five personality traits, we used the PsychGenerator dataset, where trait levels were processed using z-score normalization to achieve zero mean and unit variance. To define the high and low levels for each personality trait, we divided the training data for each trait into three equal segments

Data Generation Method	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**	**Average**
Test set (eval classifier accuracy)	93.7	94.2	93.4	93.4	94.3	93.8
***Ours*: Generator**	82.5	80.0	80.0	81.0	78.5	**80.4**
Post-Completion: `GPT-4o-mini`	64.0	59.5	56.0	57.0	59.5	59.2
Topic-Post Generation: `LLaMA-3-8B-Inst`	66.0	73.0	81.0	88.5	83.0	78.2
Topic-Post Generation: `GPT-4o-mini`	65.0	78.0	80.0	85.5	84.0	78.5

Table 7: Accuracy (%) of the trained classifier in predicting each of the Big Five personality traits. The first row (Test set) shows the classifier's accuracy on the test split, demonstrating that the classifier is well-trained. The remaining rows display the performance of our generator model compared to the two baselines, as assessed by the same classifier.

based on thresholds at the one-third and two-thirds quantiles of the trait's distribution. The lowest segment was designated as the low level, and the highest segment as the high level for the respective trait.

Each expert generator was based on the `LLaMA-3-8B-Instruct` model and fine-tuned using supervised fine-tuning (SFT) on the Alpaca format (Taori et al., 2023), which consists of three components: *instruction*, *input*, and *output*. The fine-tuning process followed these specifications:

- **Instruction**: We specify the name and level of a personality trait in the instruction. (e.g. *"Help me complete the sentence with certain Big Five Personality: Openness - high."*)
- **Input**: We provide the first five words of a post from the PsychGenerator dataset (e.g. *"who's got time to eat?"*). This serves as an initial context or prompt for the model.[2]
- **Output**: The remainder of the post from the dataset (e.g. *"I'll just have a can of frosting."*), which typically embodies the specified personality trait.

Fine-tuning was performed using all parameters of the `LLaMA-3-8B-Instruct` model over one epoch with a learning rate of 1×10^{-6}. The process ran on 4 NVIDIA A6000 GPUs, with a batch size of 1 per device.

The resulting fine-tuned expert generators produced expert-generated logits z_t^{expert}, which were subsequently used to generate the BIG5-CHAT dataset. This dataset was created by combining logits from the expert generators with those from a `LLaMA-3-70B-Instruct` model to produce z_t^{combined} as described in Eq. (1), using a scaling factor $\gamma = 0.5$ and greedy decoding for dialogue generation.

Below is the complete instruction prompt used during the expert generator training process:Section 3.2:

```
Help me complete the sentence with
certain Big Five Personality: {trait} -
{level}
{first_five_words}
```

D.5 Prompt-Based Method Details

Below is the prompt used for instruction-based prompting:

```
You are a person with {level} {trait}.
```

The following prompt is used for demonstration-based prompting. For the method referred to as **Prompt-Demo**, we randomly sample 10 examples with the same traits and levels from the BIG5-CHAT dataset and fix these examples during inference. In contrast, **Prompt-Demo-Sampling** also utilizes this prompt but dynamically samples examples during inference at each step.

```
Here are 10 examples of how people like
you have responded in different
situations. Pay attention to how they
approach communication and
problem-solving.

{10_icl_examples_for_specific_levels_a↵
nd_traits}
```

D.6 SFT and DPO Alignment Training Details

We performed alignment training using the Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) methods on `LLaMA-3-70B-Instruct`. Both training approaches utilized the Low-Rank Adaptation (LoRA) technique (Hu et al., 2021), which enabled efficient fine-tuning of the large language model

---

[2] We experimented with using only the first word as input. We empirically determined that using the first five words resulted in better generation quality.

by adapting a subset of its parameters. To ensure computational efficiency, we employed GPTQ quantization during training. The experiments were conducted using 4 NVIDIA A6000 GPUs, with each GPU processing a batch size of 1.

For LoRA, we applied the technique across all layers of the model for both SFT and DPO. The training configuration included a learning rate of 1.0×10^{-5}, regulated by a cosine scheduler, a warm-up phase consisting of 20 steps, and a gradient accumulation over 16 steps. We limited training to one epoch with a maximum sequence length of 1024 tokens. For DPO training, we used the standard sigmoid preference loss, and the preference beta value was set to 0.1 to balance preference modeling. Each training required approximately 24 hours to complete. To optimize computational resources, we used mixed-precision training with bfloat 16. Both datasets were preprocessed using the `LLaMA-3-70B-Instruct` template and split into training and validation sets, with 10% of the data reserved for validation to monitor performance.

The training prompt shared across both SFT and DPO follows the template below:

```
You are a person with the following Big
Five personality trait: {trait} -
{level}.
```

D.7 Reasoning Evaluation Setup Details

We conducted reasoning evaluations following the frameworks established by the Language Model Evaluation Harness (Gao et al., 2024b) and DeepSeek-Coder (Guo et al., 2024) to assess performance on general and social benchmarks. EleutherAI's Language Model Evaluation Harness is an open-source collaborative benchmarking codebase that consolidates existing tasks and provides a standardized API for evaluating models.[3] Similarly, DeepSeek-Coder offers a suite of coding benchmark implementations, and we directly utilized it for our work.[4]

We conducted evaluations using 1 as the batch size. For TruthfulQA, we used the multiple-choice metric, and for GSM8K, we relied on exact match scores. We measured accuracy and standard error across other tasks. The number of examples for each benchmark is listed in Table 8.

Benchmarks	Number of examples
TruthfulQA (Lin et al., 2021)	817
GPQA (Rein et al., 2023)	448
SocialIQA (Sap et al., 2019)	38,000
CommonsenseQA (Talmor et al., 2019)	12,247
GSM8K (Cobbe et al., 2021)	8,500
MathQA (Amini et al., 2019)	37,000
MMLU (Hendrycks et al., 2020)	15,908
PIQA (Bisk et al., 2020)	20,000

Table 8: Number of examples included in each reasoning benchmark.

E Additional Evaluation Results

E.1 Human Evaluation for BIG5-CHAT

We conducted a human evaluation to assess the realism and validity of BIG5-CHAT. This evaluation compared BIG5-CHAT with a baseline model, `LLaMA-3-70B-Instruct`, which follows the same procedure for generating dialogue responses but does not incorporate expert generators or the DExperts framework. In the baseline, personality traits are induced using the following prompt: "You are a person with the following Big Five personality trait: trait - level." The evaluation setup is as follows:

Two graduate students, familiar with the Big Five personality framework, were tasked with comparing examples from the BIG5-CHAT dataset against examples generated by `LLaMA-3-70B-Instruct` (without the expert generator). The comparison involved 200 randomly sampled examples from the BIG5-CHAT dataset, ensuring an equal distribution of personality traits and levels (e.g., equal representation of high and low openness, conscientiousness, etc.).

The evaluation focused on two key metrics:

1. **Expressiveness of personality traits and levels:** Evaluates whether the expected level of a Big Five personality trait is adequately reflected in Speaker Y's response.

2. **Realism of the dialogue response:** Assesses how human-like and convincing Speaker Y's response is within the dialogue context, given Speaker X's utterance.

To ensure consistency, the annotators were provided with the following definitions: "Personality trait expressiveness assesses whether the expected level of a Big Five personality trait is adequately reflected in Speaker Y's response. Realism assesses how human-like and convincing Speaker Y's re-

---

[3] `https://github.com/EleutherAI/lm-evaluation-harness`

[4] `https://github.com/deepseek-ai/DeepSeek-Coder`

sponse is within the dialog, given Speaker X’s utterance.”

For each pair of responses, annotators chose one of three options:

- “System A’s generation is better than System B’s generation.”
- “System A’s generation is equal to System B’s generation.”
- “System A’s generation is worse than System B’s generation.”

The system names were anonymized and randomly shuffled to mitigate selection bias.

Comparison with baselines	Ours win (%)	Draw (%)	Ours lose (%)	Cohen’s Kappa
Expressiveness	50.30%	39.80%	10.00%	0.50
Realism	47.80%	42.30%	10.00%	0.55

Table 9: Human evaluation results for BIG5-CHAT. Values are averaged across annotators.

The results in Table 9 show that our approach significantly outperforms the prompting baseline in both realism and the expressiveness of personality levels, as validated by human judgment. These findings highlight the limitations of prompt-based approaches, which depend on general-purpose models and often lack the fine-grained, human-grounded control required for nuanced personality expression.

E.2 Human Evaluation for the Expert Generator

To assess the expert generator in a human-grounded manner, we conducted a human evaluation comparing its outputs against the two baseline methods described in Table 1. Two graduate students, familiar with the Big Five personality framework, were tasked with evaluating two separate sets of comparisons:

1. Expert generator outputs vs. outputs from the *Post-Completion* baseline.
2. Expert generator outputs vs. outputs from the *Topic-Post Generation* baseline.

The evaluation setup consisted of 200 examples for each comparison, randomly sampled from the 1,000 test examples mentioned in Table 1. To ensure balanced coverage, each subset included an equal number of posts representing high and low levels of each personality trait (e.g., high and low openness, conscientiousness, etc.). Annotators were instructed to evaluate the expressiveness of personality traits and levels, choosing one of three options for each pair:

1. “System A’s generation is better than System B’s generation.”
2. “System A’s generation is equal to System B’s generation.”
3. “System A’s generation is worse than System B’s generation.”

The system names were anonymized and randomly shuffled to mitigate selection bias.

Comparison with baselines	Ours win (%)	Draw (%)	Ours lose (%)	Cohen’s Kappa
Post-Completion	79.25%	2.00%	18.75%	0.41
Topic-Post Generation	66.50%	19.25%	14.25%	0.61

Table 10: Human evaluation results for the expert generator. Values are averaged across annotators.

The human evaluation results presented in Table 10 indicate that the expert generator was consistently rated as more effective at expressing personality traits compared to the baselines. Additionally, the lower classifier accuracy and human evaluation ratings for the *Post-Completion* baseline highlight the increased difficulty of aligning with the desired traits when using the expert generator’s approach, reinforcing the validity of the classifier’s assessment. While these results should be interpreted with caution, as the human evaluators were not psychological experts, they nevertheless provide strong evidence supporting the expert generator’s ability to express personality traits in a grounded and realistic manner.

E.3 Personality Trait Assessment Results

The comprehensive personality test results for additional baselines are presented in Table 11, providing a more detailed view to complement Table 2. Our observations indicate that **Prompt-Demo-Sampling** performs comparably to **Prompt-Demo** without offering any noticeable improvements in performance. While applying demonstration-based prompting on SFT/DPO yields slight performance gains compared to demonstration-based prompting alone, it still falls significantly short of the standalone performance of SFT/DPO. This suggests that combining demonstration-based prompting with SFT/DPO does not result in overall enhancements. Instruction-based prompt-

Method	**Openness**		**Conscientiousness**		**Extraversion**		**Agreeableness**		**Neuroticism**		**Average**	
	High ↑	**Low ↓**	**High ↑**	**Low ↓**	**High ↑**	**Low ↓**	**High ↑**	**Low ↓**	**High ↑**	**Low ↓**	**High ↑**	**Low ↓**
BFI `LLaMA-3-8B-Instruct`												
Direct	3.1 ± 0.1		3.0 ± 0.0		3.0 ± 0.0		3.0 ± 0.0		3.0 ± 0.0		3.0 ± 0.0	
Prompt-Inst	5.0 ± 0.0	2.0 ± 0.3	4.9 ± 0.1	1.9 ± 0.1	4.8 ± 0.3	1.9 ± 0.1	4.9 ± 0.1	2.4 ± 0.4	4.1 ± 0.2	1.6 ± 0.0	4.7 ± 0.1	2.0 ± 0.2
SFT	5.0 ± 0.0	2.0 ± 0.2	5.0 ± 0.0	1.6 ± 0.1	4.7 ± 0.4	2.7 ± 0.5	5.0 ± 0.0	1.2 ± 0.1	4.1 ± 0.2	2.5 ± 0.0	**4.8 ± 0.1**	2.0 ± 0.2
DPO	5.0 ± 0.0	1.6 ± 0.2	5.0 ± 0.0	1.6 ± 0.1	4.8 ± 0.3	2.5 ± 0.0	4.8 ± 0.2	1.0 ± 0.0	3.5 ± 0.0	1.1 ± 0.1	4.6 ± 0.1	**1.6 ± 0.1**
BFI `LLaMA-3-70B-Instruct`												
Direct	4.4 ± 0.1		4.4 ± 0.1		3.3 ± 0.1		4.6 ± 0.1		2.1 ± 0.2		3.8 ± 0.1	
Prompt-Demo	4.0 ± 0.1	2.5 ± 0.1	4.0 ± 0.1	2.0 ± 0.1	4.5 ± 0.1	2.3 ± 0.1	4.4 ± 0.1	2.0 ± 0.0	3.6 ± 0.0	2.1 ± 0.1	4.1 ± 0.1	2.2 ± 0.1
Prompt-Demo-Sampling	4.4 ± 0.1	2.3 ± 0.2	4.1 ± 0.1	2.3 ± 0.1	4.3 ± 0.2	2.4 ± 0.1	4.4 ± 0.1	1.8 ± 0.2	3.5 ± 0.1	2.1 ± 0.2	4.1 ± 0.1	2.2 ± 0.2
Prompt-Inst	5.0 ± 0.1	1.8 ± 0.0	5.0 ± 0.0	1.6 ± 0.0	5.0 ± 0.0	1.4 ± 0.1	4.9 ± 0.0	1.5 ± 0.1	5.0 ± 0.1	1.6 ± 0.0	**5.0 ± 0.0**	1.6 ± 0.0
SFT	5.0 ± 0.0	1.2 ± 0.1	5.0 ± 0.1	1.4 ± 0.1	5.0 ± 0.0	1.2 ± 0.1	5.0 ± 0.1	1.6 ± 0.2	5.0 ± 0.0	1.1 ± 0.2	**5.0 ± 0.0**	**1.3 ± 0.1**
SFT-Prompt-Demo	4.2 ± 0.1	2.4 ± 0.1	4.0 ± 0.2	2.1 ± 0.1	4.5 ± 0.2	2.3 ± 0.1	4.6 ± 0.0	1.3 ± 0.2	3.9 ± 0.2	2.4 ± 0.1	4.2 ± 0.1	2.1 ± 0.1
DPO	5.0 ± 0.0	1.5 ± 0.1	5.0 ± 0.0	1.5 ± 0.1	5.0 ± 0.0	1.0 ± 0.1	5.0 ± 0.0	1.8 ± 0.2	5.0 ± 0.0	1.1 ± 0.0	**5.0 ± 0.0**	1.4 ± 0.1
DPO-Prompt-Demo	4.1 ± 0.1	2.2 ± 0.1	4.1 ± 0.1	2.0 ± 0.0	4.5 ± 0.1	2.4 ± 0.1	4.6 ± 0.1	1.3 ± 0.1	3.7 ± 0.1	2.1 ± 0.1	4.2 ± 0.1	2.0 ± 0.1
BFI `GPT-4o-Mini`												
Prompt-Demo	4.8 ± 0.0	3.3 ± 0.1	4.5 ± 0.1	3.0 ± 0.1	4.6 ± 0.1	2.6 ± 0.1	4.9 ± 0.0	1.5 ± 0.2	3.6 ± 0.1	2.2 ± 0.1	4.5 ± 0.1	2.5 ± 0.1
Prompt-Inst	5.0 ± 0.0	1.4 ± 0.2	5.0 ± 0.0	1.5 ± 0.1	5.0 ± 0.0	1.2 ± 0.0	5.0 ± 0.0	1.4 ± 0.0	4.9 ± 0.0	1.0 ± 0.1	5.0 ± 0.0	1.3 ± 0.1
IPIP-NEO `LLaMA-3-8B-Instruct`												
Direct	3.0 ± 0.1		3.3 ± 0.0		3.4 ± 0.1		3.2 ± 0.0		3.0 ± 0.1		3.2 ± 0.1	
Prompt-Inst	4.4 ± 0.1	1.5 ± 0.1	4.5 ± 0.1	2.3 ± 0.1	5.0 ± 0.0	1.9 ± 0.0	4.6 ± 0.0	2.3 ± 0.1	4.2 ± 0.1	2.6 ± 0.1	4.5 ± 0.1	2.1 ± 0.1
SFT	4.3 ± 0.1	1.5 ± 0.1	4.5 ± 0.2	2.7 ± 0.1	5.0 ± 0.0	2.2 ± 0.1	4.0 ± 0.2	1.8 ± 0.2	4.3 ± 0.1	2.0 ± 0.1	4.4 ± 0.1	**2.0 ± 0.1**
DPO	5.0 ± 0.0	1.9 ± 0.1	5.0 ± 0.0	2.9 ± 0.1	5.0 ± 0.0	1.6 ± 0.1	4.5 ± 0.1	1.2 ± 0.0	3.8 ± 0.1	3.7 ± 0.1	**4.7 ± 0.0**	2.3 ± 0.1
IPIP-NEO `LLaMA-3-70B-Instruct`												
Direct	3.6 ± 0.1		4.0 ± 0.1		3.5 ± 0.1		4.0 ± 0.0		2.3 ± 0.1		3.5 ± 0.1	
Prompt-Demo	3.5 ± 0.0	2.5 ± 0.1	3.8 ± 0.0	2.2 ± 0.1	4.0 ± 0.1	2.5 ± 0.0	4.3 ± 0.0	2.1 ± 0.1	3.0 ± 0.1	2.2 ± 0.1	3.7 ± 0.0	2.3 ± 0.1
Prompt-Demo-Sampling	3.5 ± 0.0	2.6 ± 0.1	4.0 ± 0.0	2.6 ± 0.1	4.0 ± 0.1	2.5 ± 0.1	4.3 ± 0.0	2.1 ± 0.1	3.0 ± 0.1	2.3 ± 0.1	3.8 ± 0.0	2.4 ± 0.1
Prompt-Inst	4.6 ± 0.0	1.3 ± 0.0	5.0 ± 0.0	1.4 ± 0.0	5.0 ± 0.0	1.6 ± 0.0	4.8 ± 0.0	1.1 ± 0.1	4.9 ± 0.0	1.7 ± 0.1	**4.9 ± 0.0**	1.4 ± 0.0
SFT	4.9 ± 0.1	1.1 ± 0.0	5.0 ± 0.0	1.3 ± 0.1	5.0 ± 0.0	1.3 ± 0.0	4.9 ± 0.0	1.0 ± 0.0	4.9 ± 0.0	1.2 ± 0.1	**4.9 ± 0.0**	**1.2 ± 0.0**
SFT-Prompt-Demo	3.7 ± 0.1	2.5 ± 0.2	3.7 ± 0.1	2.0 ± 0.1	4.1 ± 0.1	2.7 ± 0.1	4.3 ± 0.1	1.2 ± 0.1	3.6 ± 0.2	2.2 ± 0.1	3.9 ± 0.1	2.1 ± 0.1
DPO	4.8 ± 0.0	1.4 ± 0.1	5.0 ± 0.0	1.6 ± 0.1	5.0 ± 0.0	1.1 ± 0.1	4.9 ± 0.0	1.0 ± 0.0	5.0 ± 0.0	1.1 ± 0.0	**4.9 ± 0.0**	**1.2 ± 0.1**
DPO-Prompt-Demo	3.5 ± 0.1	2.4 ± 0.0	3.9 ± 0.0	2.1 ± 0.0	4.1 ± 0.1	2.5 ± 0.0	4.4 ± 0.0	2.0 ± 0.1	3.1 ± 0.1	2.1 ± 0.0	3.8 ± 0.1	2.2 ± 0.0
IPIP-NEO `GPT-4o-Mini`												
Prompt-Demo	4.2 ± 0.0	2.9 ± 0.1	4.2 ± 0.1	3.2 ± 0.1	4.0 ± 0.0	2.6 ± 0.1	4.6 ± 0.1	2.4 ± 0.1	3.4 ± 0.0	2.1 ± 0.1	4.1 ± 0.0	2.6 ± 0.1
Prompt-Inst	4.8 ± 0.0	1.9 ± 0.2	4.9 ± 0.0	1.4 ± 0.0	4.9 ± 0.0	1.6 ± 0.0	4.8 ± 0.0	2.1 ± 0.1	4.9 ± 0.0	1.1 ± 0.1	4.9 ± 0.0	1.6 ± 0.1

Table 11: Full personality test results for various alignment methods, complementing Table 2. **Prompt-Demo-Sampling** involves randomly sampling 10 examples from the entire BIG5-CHAT dataset for each run, instead of using a fixed set of 10 random examples across runs. **SFT-Prompt-Demo** and **DPO-Prompt-Demo** represent demonstration-based prompting applied to SFT and DPO-trained models, respectively. Results for `GPT-4o-mini` are presented in separate sections of the table. Scores range from 1 to 5, where a score closer to 5 indicates stronger agreement with the trait, while a score closer to 1 reflects weaker or opposing agreement.

ing with `GPT-4o-mini` achieves similar performance levels as `LLaMA-3-70B-Instruct`. However, demonstration-based prompting does not exhibit superior performance compared to SFT/DPO when applied to `LLaMA-3-70B-Instruct`, reinforcing the conclusion that demonstration-based methods are not as effective as SFT/DPO in this context. We do not provide demonstration-based prompting results for `LLaMA-3-8B-Instruct` because the model consistently failed to generate reasonable responses to the questionnaire when presented with a lengthy 10-shot context. This outcome highlights the model's limited instruction-following capabilities.

Figure 3 presents the BFI and IPIP-NEO test score results for the LLaMA-3 Instruct models, evaluated in zero-shot inference without any induced personality traits. The crowd-sourced response scores for the BFI test are sourced from Huang et al. (2024), and those for the IPIP-NEO test are drawn from Jiang et al. (2023). The results indicate that the scores for both `LLaMA-3-8B-Instruct` and `LLaMA-3-70B-Instruct` fall within the standard deviation of the human distribution. However, while `LLaMA-3-8B-Instruct` tends to generate more neutral scores (around 3 across most of the Big Five traits), `LLaMA-3-70B-Instruct` exhibits higher scores for openness, conscientiousness, extraversion, and agreeableness, and lower scores for neuroticism.

E.4 Evaluating Finetune Models Psycholinguistic Richness in Unseen SODA Scenarios

Description: To address the concern that questionnaire-based evaluations may overlook the psycholinguistic richness essential for authentic human behavior, we designed this experiment to assess whether our models can generate linguistically nuanced outputs. Specifically, models—whether fine-tuned or prompted—are tasked with generating responses to unseen scenarios from the SODA dataset. We then employ our trained RoBERTa classifier to determine if the generated responses effectively reflect the desired personality traits.

Results and Analysis: This result (see Figure 2) demonstrates that DPO has a statistically significant advantage in capturing psycholinguistic richness. The performance of SFT and prompt-based approaches appears to be similar. This suggests that while both SFT and prompting can encode personality traits to some extent, they may lack the nuanced psycholinguistic adaptation achieved through preference optimization in DPO. Our findings suggest that training-based approaches, particularly DPO, are more effective in capturing the nuanced psycholinguistic richness required for authentic personality expression. Compared to prompt-based methods, which rely on external conditioning without modifying the underlying model parameters, training-based models can internalize personality traits more robustly, leading to more consistent and contextually appropriate generations. Furthermore, the limitations observed in SFT indicate that conventional supervised fine-tuning alone may not be sufficient for fully encoding the complexity of psycholinguistic adaptation. This suggests that while SFT can guide model behavior to some extent, it may lack the reinforcement-driven refinement necessary to achieve deeper alignment with personality traits.

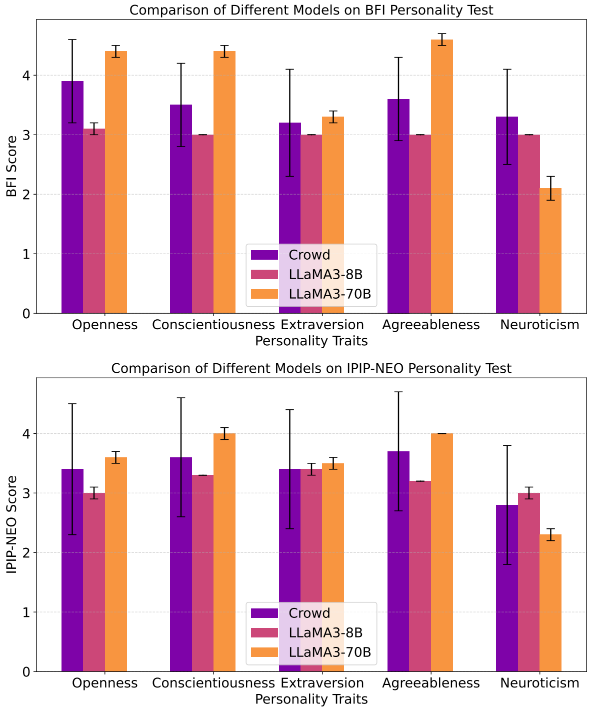

Figure 3: The personality test results for the crowd and the `LLaMA-3-Instruct` models were obtained using zero-shot inference without explicitly inducing personality traits. The BFI test scores are displayed on the left. The IPIP-NEO test scores are displayed on the right.

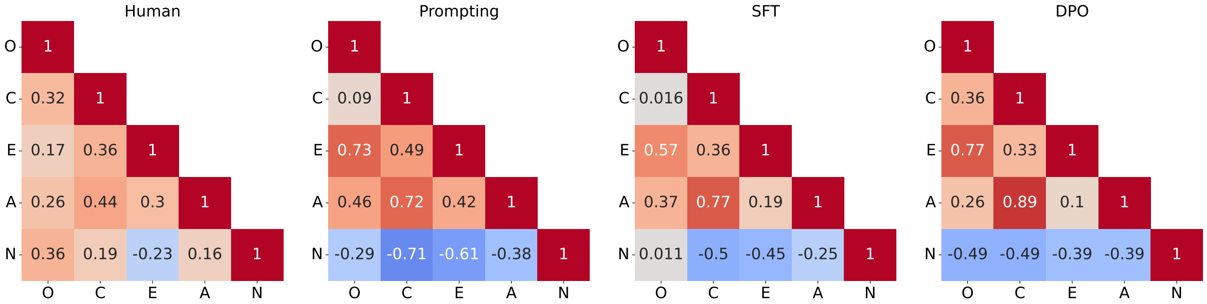

Figure 4: Intra-trait Pearson correlations for human distributions on IPIP-NEO and the corresponding results from instruction-based prompting, SFT, and DPO. O represents openness, C conscientiousness, E extraversion, A agreeableness, and N neuroticism. The correlations especially for SFT align well with human distributions across openness, conscientiousness, extraversion, and agreeableness. Neuroticism shows less alignment with the other four traits compared to human distribution.

E.5 Intra-Trait Correlations in Personality Assessment

To assess how well the prompting and training methods simulate intra-trait correlations observed in human data, we first calculated the intra-trait correlations from real human distributions using the IPIP-NEO questionnaire, based on the PAPI-120-600K dataset from Zhu et al. (2024), which includes 619K human responses to the IPIP-NEO. Next, we computed the intra-trait correlations for the prompting, SFT, and DPO methods using the results from Table 2. These correlations are visualized in Figure 4, showing that most traits are positively correlated, with the exception of neuroticism. To quantify the similarity between the method-generated and human correlation matrices, we calculated the matrix distance using the Frobenius norm, where 0 represents perfect similarity and 10 indicates maximum dissimilarity. The matrix distances were 2.10 for prompting, 1.55 for SFT, and 2.06 for DPO. These results suggest that the trained models, particularly SFT, more accurately capture the trait correlations seen in natural human data compared to the prompting-based methods.

E.6 Reasoning Benchmark Results for `LLaMA-3-70B-Instruct`

The complete results for the general reasoning tasks evaluated on the LLaMA-3-70B-Instruct model are presented in Table 12. Note that the GPQA results in Table 3 were obtained using zero-shot prompting. This evaluation encompasses multiple reasoning domains and highlights the impact of different training methodologies: prompting, SFT, and DPO. These methods were assessed based on their ability to preserve the reasoning capabilities.

The results indicate that the SFT method consistently delivers the strongest performance across the benchmarks, outperforming both DPO and the prompting-based approach. For the 70B model, SFT emerges as the most effective method, achieving an optimal balance between incorporating personality traits and maintaining robust reasoning functionality. The aggregated results underscore the reliability of SFT, which demonstrates superior performance across diverse reasoning tasks, making it a robust choice for large-scale language models.

In contrast, the performance of the DPO method is more variable. While DPO excels in certain scenarios, such as low Neuroticism within the TruthfulQA task—where it achieves a notable score of 65.8%—its overall results are less consistent across other reasoning benchmarks. Moreover, the final average scores reveal that high-trait DPO models underperform compared to their low-trait counterparts in general. This suggests a potential misalignment between DPO's training objectives and the reasoning requirements of specific tasks. These findings highlight the nuanced trade-offs between training strategies, with SFT offering the most reliable approach for balancing personality trait integration and cognitive task performance in large-scale models.

E.7 Reasoning Benchmark Results for `LLaMA-3-8B-Instruct`

The reasoning evaluation results for the `LLaMA-3-8B-Instruct` model, assessed across six reasoning domains, are summarized in Table 13. Overall, the DPO method generally outperformed SFT and demonstrated performance comparable to the prompt-based approach. This indicates

Benchmark	Direct	Method	Openness		Conscientiousness		Extraversion		Agreeableness		Neuroticism		Average	
			High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
Hallucination Detection														
TruthfulQA	58.6 ± 1.7	Prompt	54.1 ± 1.6	51.1 ± 1.6	55.9 ± 1.7	45.2 ± 1.6	52.0 ± 1.6	55.7 ± 1.6	52.3 ± 1.7	49.1 ± 1.6	48.9 ± 1.6	58.6 ± 1.6	52.6 ± 1.6	51.9 ± 1.6
		SFT	55.2 ± 1.6	52.8 ± 1.6	55.6 ± 1.6	50.8 ± 1.5	54.5 ± 1.6	56.7 ± 1.6	54.4 ± 1.6	51.6 ± 1.6	52.4 ± 1.5	56.7 ± 1.6	54.4 ± 1.6	53.7 ± 1.6
		DPO	54.6 ± 1.6	54.2 ± 1.7	64.6 ± 1.6	38.5 ± 1.6	46.0 ± 1.7	65.3 ± 1.6	59.6 ± 1.6	50.6 ± 1.6	43.0 ± 1.7	65.8 ± 1.6	53.6 ± 1.6	54.9 ± 1.6
Social Reasoning														
SocialIQA	46.6 ± 1.1	Prompt	40.8 ± 1.1	43.9 ± 1.1	42.9 ± 1.1	39.9 ± 1.1	43.3 ± 1.1	42.0 ± 1.1	42.4 ± 1.1	40.8 ± 1.1	39.1 ± 1.1	44.1 ± 1.1	41.7 ± 1.1	42.1 ± 1.1
		SFT	50.3 ± 1.1	50.4 ± 1.1	50.9 ± 1.1	46.8 ± 1.1	50.0 ± 1.1	50.3 ± 1.1	50.5 ± 1.1	46.6 ± 1.1	48.2 ± 1.1	50.6 ± 1.1	50.0 ± 1.1	48.9 ± 1.1
		DPO	41.5 ± 1.1	44.5 ± 1.1	44.7 ± 1.1	37.6 ± 1.1	43.0 ± 1.1	43.6 ± 1.1	44.8 ± 1.1	39.0 ± 1.1	40.0 ± 1.1	45.3 ± 1.1	42.8 ± 1.1	42.0 ± 1.1
Commonsense Reasoning														
CommonsenseQA	27.0 ± 1.3	Prompt	60.0 ± 1.4	59.9 ± 1.4	22.5 ± 1.2	22.3 ± 1.2	35.5 ± 1.4	50.0 ± 1.4	45.0 ± 1.4	34.9 ± 1.4	20.2 ± 1.2	36.8 ± 1.4	36.6 ± 1.3	40.8 ± 1.4
		SFT	77.7 ± 1.2	78.8 ± 1.2	77.6 ± 1.2	66.0 ± 1.4	75.7 ± 1.2	78.9 ± 1.2	77.0 ± 1.2	73.8 ± 1.3	79.1 ± 1.2	78.5 ± 1.2	77.4 ± 1.2	75.2 ± 1.3
		DPO	57.7 ± 1.4	65.9 ± 1.4	23.8 ± 1.2	25.8 ± 1.3	23.2 ± 1.2	70.8 ± 1.3	21.3 ± 1.2	39.2 ± 1.4	20.1 ± 1.1	44.6 ± 1.4	29.2 ± 1.2	49.3 ± 1.4
PIQA	80.4 ± 0.9	Prompt	79.6 ± 0.9	79.8 ± 0.9	80.5 ± 0.9	77.3 ± 1.0	78.0 ± 1.0	80.0 ± 0.9	79.8 ± 0.9	78.4 ± 1.0	78.8 ± 1.0	80.7 ± 0.9	79.3 ± 0.9	79.2 ± 0.9
		SFT	81.2 ± 0.9	81.0 ± 0.9	81.2 ± 0.9	80.4 ± 0.9	81.8 ± 0.9	81.3 ± 0.9	81.2 ± 0.9	80.0 ± 0.9	81.0 ± 0.9	81.2 ± 0.9	81.3 ± 0.9	80.8 ± 0.9
		DPO	76.4 ± 1.0	76.8 ± 1.0	79.4 ± 0.9	70.9 ± 1.1	76.4 ± 1.0	79.8 ± 0.9	78.5 ± 1.0	74.0 ± 1.0	72.9 ± 1.0	79.5 ± 0.9	76.7 ± 1.0	76.2 ± 1.0
Math Reasoning														
GSM8K	80.6 ± 1.1	Prompt	75.7 ± 1.2	70.1 ± 1.3	73.5 ± 1.2	32.6 ± 1.3	80.8 ± 1.1	33.5 ± 1.3	87.2 ± 0.9	77.8 ± 1.1	26.0 ± 1.2	89.4 ± 0.8	68.6 ± 1.1	60.7 ± 1.2
		SFT	85.8 ± 1.0	76.2 ± 1.2	86.4 ± 0.9	81.7 ± 1.1	85.1 ± 1.0	86.7 ± 0.9	87.0 ± 0.9	74.5 ± 1.2	76.0 ± 1.2	87.3 ± 0.9	84.1 ± 1.0	81.3 ± 1.1
		DPO	87.9 ± 0.9	88.5 ± 0.9	90.2 ± 0.8	80.6 ± 1.1	88.9 ± 0.9	90.4 ± 0.8	87.3 ± 0.9	90.0 ± 0.8	15.2 ± 1.0	91.0 ± 0.8	73.9 ± 0.9	88.1 ± 0.9
MathQA	39.0 ± 0.9	Prompt	33.5 ± 0.9	33.5 ± 0.9	32.8 ± 0.9	31.5 ± 0.9	32.3 ± 0.9	33.3 ± 0.9	33.6 ± 0.9	32.4 ± 0.9	32.1 ± 0.9	34.1 ± 0.9	32.9 ± 0.9	33.0 ± 0.9
		SFT	43.3 ± 0.9	42.6 ± 0.9	43.0 ± 0.9	43.3 ± 0.9	43.2 ± 0.9	42.7 ± 0.9	42.9 ± 0.9	42.9 ± 0.9	42.8 ± 0.9	43.3 ± 0.9	43.0 ± 0.9	43.0 ± 0.9
		DPO	33.9 ± 0.9	34.7 ± 0.9	32.9 ± 0.9	28.1 ± 0.8	30.5 ± 0.8	35.0 ± 0.9	31.3 ± 0.8	32.8 ± 0.9	28.9 ± 0.8	34.0 ± 0.9	31.5 ± 0.8	32.9 ± 0.9
General Reasoning														
MMLU	74.5 ± 0.3	Prompt	70.3 ± 0.4	69.6 ± 0.4	40.6 ± 0.4	52.8 ± 0.4	56.9 ± 0.4	72.8 ± 0.4	69.0 ± 0.4	69.2 ± 0.4	55.3 ± 0.4	67.9 ± 0.4	58.4 ± 0.4	66.5 ± 0.4
		SFT	72.5 ± 0.4	72.0 ± 0.4	73.1 ± 0.4	68.6 ± 0.4	72.1 ± 0.4	73.5 ± 0.4	72.8 ± 0.4	70.7 ± 0.4	72.5 ± 0.4	73.8 ± 0.4	72.6 ± 0.4	71.7 ± 0.4
		DPO	57.9 ± 0.4	64.4 ± 0.4	50.3 ± 0.4	33.8 ± 0.4	42.3 ± 0.4	72.3 ± 0.4	34.3 ± 0.4	62.5 ± 0.4	33.2 ± 0.4	69.1 ± 0.4	43.6 ± 0.4	60.4 ± 0.4
GPQA (0-shot)	33.5 ± 2.2	Prompt	31.5 ± 2.2	34.2 ± 2.2	31.7 ± 2.2	32.4 ± 2.2	34.6 ± 2.2	32.1 ± 2.2	32.4 ± 2.2	32.8 ± 2.2	31.9 ± 2.2	32.1 ± 2.2	32.4 ± 2.2	32.7 ± 2.2
		SFT	33.5 ± 2.2	32.4 ± 2.2	34.2 ± 2.2	34.2 ± 2.2	33.3 ± 2.2	34.4 ± 2.2	33.3 ± 2.2	33.3 ± 2.2	34.4 ± 2.2	33.5 ± 2.2	33.7 ± 2.2	33.6 ± 2.2
		DPO	36.8 ± 2.3	31.9 ± 2.2	35.7 ± 2.3	30.6 ± 2.2	35.9 ± 2.3	35.9 ± 2.3	35.5 ± 2.3	35.7 ± 2.3	32.6 ± 2.2	34.6 ± 2.2	35.3 ± 2.3	33.7 ± 2.2
GPQA (5-shot)	36.6 ± 2.3	Prompt	35.9 ± 2.3	32.6 ± 2.2	36.2 ± 2.3	35.7 ± 2.3	36.2 ± 2.3	35.7 ± 2.3	34.4 ± 2.2	34.8 ± 2.3	36.6 ± 2.3	34.2 ± 2.2	35.9 ± 2.3	34.6 ± 2.3
		SFT	32.4 ± 2.2	32.8 ± 2.2	34.4 ± 2.2	33.7 ± 2.2	33.0 ± 2.2	33.9 ± 2.2	33.7 ± 2.2	32.8 ± 2.2	33.7 ± 2.2	34.8 ± 2.3	33.4 ± 2.2	33.6 ± 2.2
		DPO	37.5 ± 2.3	31.2 ± 2.2	35.9 ± 2.3	31.2 ± 2.2	37.1 ± 2.3	35.5 ± 2.3	33.5 ± 2.2	32.1 ± 2.2	36.6 ± 2.3	35.7 ± 2.3	36.1 ± 2.3	33.1 ± 2.2
Average	53.0 ± 1.3	Prompt	53.5 ± 1.3	52.7 ± 1.3	46.3 ± 1.3	41.1 ± 1.3	50.0 ± 1.3	48.3 ± 1.3	52.9 ± 1.3	50.0 ± 1.3	41.0 ± 1.3	53.1 ± 1.3	48.7 ± 1.3	49.1 ± 1.3
		SFT	59.1 ± 1.3	57.7 ± 1.3	59.6 ± 1.3	56.2 ± 1.3	58.7 ± 1.3	59.8 ± 1.3	59.2 ± 1.3	56.2 ± 1.3	57.8 ± 1.3	60.0 ± 1.3	58.9 ± 1.3	58.0 ± 1.3
		DPO	53.8 ± 1.3	54.7 ± 1.3	50.8 ± 1.3	41.9 ± 1.3	47.0 ± 1.3	58.7 ± 1.3	47.3 ± 1.3	50.7 ± 1.3	35.8 ± 1.3	55.5 ± 1.3	47.0 ± 1.3	52.3 ± 1.3

Table 12: Benchmark results for different personality traits on `LLaMA-3-70B-Instruct`. **Direct** refers to direct inference without including personality-related prompts. **Prompt** refers to instruction-based prompting. The table includes standard errors (shown as ± values) to provide statistical context for the results.

that, with the smaller 8B model, DPO effectively aligns personality traits without significantly compromising reasoning capabilities.

A comparison of personality trait levels revealed that models simulating high trait levels consistently outperformed their low-trait counterparts in both DPO and SFT settings. For instance, on the TruthfulQA benchmark, the high-conscientiousness DPO model achieved 55.0%, significantly surpassing the low-conscientiousness model's 39.0%. Similarly, on the GSM8K math reasoning task, the high-conscientiousness DPO model scored 72.2%, substantially outperforming the low-level model.

On benchmarks such as TruthfulQA, GPQA (both zero-shot and five-shot), and MathQA, models trained using SFT and DPO performed comparably to the original unaligned model. This suggests that personality trait alignment does not adversely affect reasoning performance in these tasks for a small model. However, notable variations were observed in other benchmarks. For example, DPO exhibited significantly reduced performance on CommonsenseQA and MMLU compared to SFT, prompting, and the original model. Conversely, SFT underperformed on the GSM8K benchmark relative to DPO, prompting, and the original model. These results suggest that the DPO method may be more effective than SFT in preserving or enhancing reasoning performance for specific tasks and traits on small models, though the choice of alignment method may depend on the specific reasoning domain.

F Correlation Between Personality Traits and Reasoning Behaviors

F.1 Human VS. `LLaMA-3-70B-Instruct`

Understanding the influence of personality traits on reasoning behaviors in LLMs is crucial for developing models tailored to specific personality profiles. Research on the Big Five personality traits has consistently demonstrated their significant impact on human cognition and problem-solving abilities (John et al., 1999; Soto et al., 2011). Traits such as openness, conscientiousness, and agreeableness are often associated with enhanced reasoning capabilities, while neuroticism has been found to impair performance across a range of reasoning tasks (Ackerman and Heggestad, 1997; Schaie et al., 2004; Chamorro-Premuzic et al., 2006).

Table 14 summarizes relevant findings

Benchmark	**Original**	**Method**	**Openness**		**Conscientiousness**		**Extraversion**		**Agreeableness**		**Neuroticism**		**Average**	
			High	Low	High	Low	High	Low	High	Low	High	Low	**High**	**Low**
Hallucination Detection														
TruthfulQA	53.5	Prompt	49.0	51.5	50.6	44.4	45.3	51.9	49.2	50.3	54.6	45.2	49.7	48.7
		SFT	50.0	45.7	50.9	43.8	46.2	52.0	49.9	46.3	53.6	42.9	50.1	46.1
		DPO	52.4	49.1	55.0	39.0	35.0	59.2	52.8	45.5	58.2	38.8	50.7	46.3
Code Reasoning														
HumanEval	60.4	Prompt	59.1	59.8	62.2	61.6	61.0	63.4	62.8	62.2	60.4	61.6	61.1	61.7
		SFT	57.9	54.3	59.8	56.1	58.5	57.3	60.4	54.9	58.5	58.5	59.0	56.2
		DPO	57.3	0.6	27.4	0.0	43.3	0.0	8.5	32.9	0.0	7.9	27.3	8.3
MBPP	54.6	Prompt	54.6	55.4	54.2	55.2	55.8	56.0	55.4	54.8	54.4	55.8	54.9	55.4
		SFT	56.2	56.2	54.2	56.2	56.4	56.4	55.6	55.8	55.0	56.4	55.5	56.2
		DPO	53.6	47.6	53.0	35.2	54.6	51.4	54.4	53.8	52.0	54.2	42.9	48.4
Social Reasoning														
SocialIQA	49.7	Prompt	41.9	42.3	41.1	39.3	41.5	41.6	41.8	39.5	42.1	39.4	41.7	40.4
		SFT	44.0	44.9	45.9	41.9	44.4	44.6	43.7	41.4	44.6	40.8	44.5	42.7
		DPO	43.8	43.8	42.5	37.8	41.8	40.9	42.8	38.4	42.8	39.0	42.7	40.0
Commonsense Reasoning														
CommonsenseQA	51.8	Prompt	64.6	60.6	38.0	31.3	45.9	55.0	55.4	36.3	33.9	23.3	47.6	41.3
		SFT	61.8	57.9	50.5	34.3	52.7	60.8	55.4	36.0	63.4	30.6	56.8	43.9
		DPO	22.9	24.8	48.2	21.6	29.1	56.6	28.4	26.3	47.7	23.7	35.3	30.6
Math Reasoning														
GSM8K	64.7	Prompt	13.5	58.4	23.4	61.0	40.0	57.1	29.3	71.6	24.1	31.9	26.1	56.0
		SFT	19.8	0.5	20.2	1.4	6.0	0.5	6.4	4.8	20.1	53.3	14.5	12.1
		DPO	68.4	31.8	72.2	31.8	69.7	63.0	70.7	64.8	71.9	3.0	70.6	38.9
MathQA	27.9	Prompt	27.6	28.3	27.9	27.3	27.1	27.8	27.2	28.1	28.1	25.9	27.6	27.5
		SFT	30.1	30.2	29.6	30.3	31.0	30.6	29.6	30.3	29.6	29.4	30.0	30.2
		DPO	26.9	27.8	28.3	25.1	25.8	27.6	24.9	27.7	29.7	24.9	27.1	26.6
General Knowledge														
MMLU	51.2	Prompt	37.5	29.1	23.2	27.0	24.7	29.2	27.7	25.5	23.4	23.8	27.3	26.9
		SFT	45.0	48.5	35.6	32.0	37.5	46.5	44.2	39.9	47.1	31.7	41.9	39.7
		DPO	23.0	29.8	29.7	26.9	24.8	41.4	30.7	26.3	30.8	23.1	27.8	29.5
GPQA **(0-shot)**	28.1	Prompt	29.0	28.8	28.6	23.0	28.6	29.2	29.0	27.2	28.8	28.3	28.8	27.3
		SFT	27.9	27.9	28.1	25.0	27.2	28.3	28.8	24.1	29.0	28.3	28.2	26.7
		DPO	27.9	25.0	29.7	21.0	27.2	26.8	28.8	21.4	29.5	25.2	28.6	23.9
GPQA **(5-shot)**	29.9	Prompt	29.7	26.6	28.8	26.8	28.3	26.6	27.9	28.6	29.0	25.2	28.7	26.8
		SFT	26.1	27.0	28.8	26.6	28.8	28.6	30.6	27.9	28.6	27.5	28.6	27.5
		DPO	27.9	26.3	28.3	23.0	26.8	28.1	27.5	24.6	28.8	25.2	27.9	25.4
Average	43.9	Prompt	35.8	40.5	31.5	34.4	34.3	39.5	35.1	38.2	31.7	29.1	33.7	36.4
		SFT	37.2	34.0	34.8	27.6	32.8	35.3	35.0	29.9	38.8	34.8	35.7	32.3
		DPO	35.6	30.7	41.6	26.9	34.1	43.2	37.7	33.8	42.4	23.4	38.3	31.6

Table 13: Benchmark results for the `LLaMA-3-8B-Instruct` model are presented across various personality traits and evaluation methods. The benchmarks are categorized into six key areas: Hallucination Detection, General Reasoning, Social Reasoning, Commonsense Reasoning, Mathematical Reasoning, and General Knowledge.

from recent psychological studies and their alignment with our experimental results on `LLaMA-3-70B-Instruct`. Our findings corroborate these studies, indicating that models exhibiting higher conscientiousness and agreeableness generally perform better in reasoning tasks. In contrast, models characterized by lower levels of extraversion and neuroticism also demonstrate improved reasoning performance. These results highlight the potential of personality-aligned training to optimize LLM performance for reasoning-intensive tasks.

F.2 Human VS. `LLaMA-3-8B-Instruct`

The influence of Big Five Personality traits on reasoning tasks in human cognition, as outlined in Table 14, served as a foundation for analyzing the performance of the `LLaMA-3-8B-Instruct` model. This analysis aims to explore how alignment with different personality traits affects the model's reasoning capabilities. Below, we summarize the observed correlations between each trait and the model's performance across various reasoning benchmarks.

Openness The impact of Openness on reasoning performance was highly task-dependent. Models aligned with high levels of Openness using the DPO method exhibited significantly improved performance in mathematical reasoning tasks. However, these models underperformed in commonsense reasoning benchmarks compared to both the prompt-based approach and the original model. These results suggest that while high Openness alignment enhances mathematical reasoning, it does not guarantee consistent improvements across all reasoning domains.

Conscientiousness A strong positive correlation was observed between Conscientiousness and reasoning performance. Models aligned with higher levels of Conscientiousness consistently outperformed their low-level counterparts across most benchmarks. This trend highlights that high Conscientiousness alignment likely enhances systematic reasoning and attention to detail, benefiting performance across diverse reasoning tasks.

Extraversion Lower levels of Extraversion were associated with better performance across reasoning tasks. Specifically, in commonsense reasoning benchmarks, models with low Extraversion significantly outperformed those with high Extraversion. This negative correlation suggests that high Extraversion may introduce distractibility, potentially impeding performance in tasks that require focused attention and analytical reasoning.

Agreeableness The influence of Agreeableness on reasoning performance was minimal and inconsistent. No clear advantage was observed for models aligned with either high or low levels of Agreeableness across the benchmarks. These findings indicate that Agreeableness has a weak correlation with the model's reasoning capabilities, suggesting its alignment has little effect on overall performance.

Neuroticism The relationship between Neuroticism and reasoning performance was inconsistent and did not align with expectations from human cognition studies. High Neuroticism models performed well in some reasoning tasks, while low Neuroticism models scored poorly in others. These results imply that high Neuroticism alignment does not necessarily impair reasoning performance, contrasting with psychological findings in humans. This discrepancy may arise from limitations in how Neuroticism is modeled and represented in the training process.

G Bias & Hallucination Discussion

Upon qualitative investigation, we observe that personality extremes (e.g. high neuroticism and low agreeableness) can induce hallucinations within LLM-generated dialogues. In this context, "hallucination" refers to the creation of unfounded details, such as overblown catastrophic outcomes or attributed malevolent intentions, that are not substantiated by the surrounding discourse. These fabrications introduce bias by distorting perceived reality, thereby amplifying potential issues or misrepresenting motives without adequate justification. Some picked examples are shown in Table 15.

Conversely, dialogues characterized by personality traits, specifically, low neuroticism, high agreeableness, and moderate extraversion tend to maintain alignment with factual bases. These responses promote constructive dialogue through measured and factual communication, reducing the likelihood of escalating tensions or misinterpreting benign situations.

Other traits such as extraversion, conscientiousness, and openness, predominantly influence the tone and creativity of responses. For example, high extraversion can result in more energetic and optimistic exchanges, whereas high openness may lead to the generation of creative or unconventional

Openness Openness is associated with intellectual curiosity and creativity and enhances problem-solving in tasks requiring abstract reasoning and social cognition (Ackerman and Heggestad, 1997; McCrae, 1987). While research indicates that openness positively correlates with cognitive abilities (Chamorro-Premuzic et al., 2006; Costa Jr et al., 1976; Graham and Lachman, 2012; Schaie et al., 2004), our models do not show significant performance differences across reasoning tasks based on openness levels, with the exception of SFT on math reasoning tasks. This suggests that openness may not directly translate to gains in reasoning tasks beyond math, despite its known benefits to human cognition.
Conscientiousness Conscientiousness, linked to discipline and organization, consistently improves model performance in mathematical reasoning and hallucination detection. This aligns with psychological studies showing that higher conscientiousness is linked to better academic performance and fewer errors in cognitive tasks due to increased diligence and thoroughness (Roberts et al., 2014; Poropat, 2009; Digman, 1990; Moutafi et al., 2003; Schaie et al., 2004).
Extraversion Extraversion is often associated with sociability and shows mixed results in cognitive tasks. While it can enhance social reasoning, it may negatively affect individual problem-solving tasks, such as math reasoning (Blickle, 1996; Ashton et al., 2002; Costa Jr et al., 1976). Our models simulating lower extraversion perform better across many reasoning domains, including math and also commonsense reasoning, consistent with findings that high extraversion can detract from tasks requiring focused, solitary work (Matthews and Gilliland, 1999; Chamorro-Premuzic and Furnham, 2006).
Agreeableness Agreeableness, linked to traits like trust and cooperation, improves social reasoning in our models, consistent with human studies (Graziano, 1997). However, it shows minimal impact on math or commonsense reasoning, reflecting research suggesting that agreeableness is less beneficial for analytical tasks (Poropat, 2009; Ackerman and Heggestad, 1997; Schaie et al., 2004).
Neuroticism Neuroticism reflects emotional instability, and is consistently associated with poorer cognitive performance due to anxiety and cognitive interference, especially social reasoning and hallucination detection (Robinson and Tamir, 2005; Zeidner, 2005; Chamorro-Premuzic et al., 2006; Eysenck, 2013). Our models confirm this, with lower Neuroticism levels leading to better performance across almost all reasoning tasks.

Table 14: Summary of the influence of Big Five personality traits on reasoning tasks in human cognition, and comparison of psychological research findings with our experimental results on LLMs.

ideas. However, unlike the extreme traits of neuroticism or agreeableness, these do not typically lead to the severe distortions or biases akin to hallucinations.

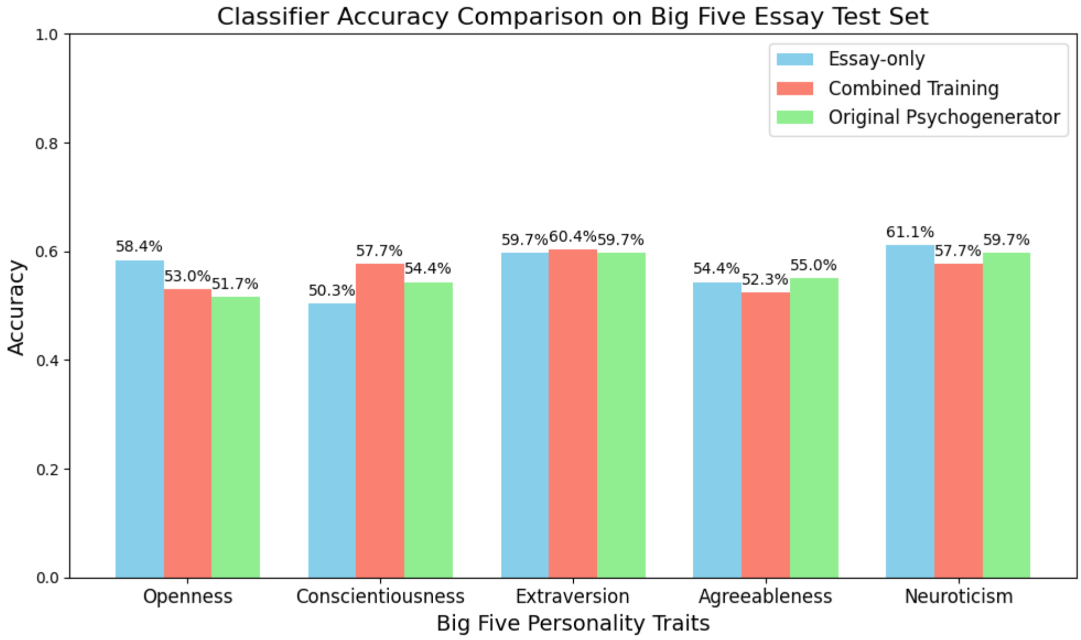

Figure 5: Comparison of classifier accuracy on the Big Five Essay test set across three training approaches. The plot compares the performance of a classifier retrained solely on the Big Five Essay dataset (Essay-only), a classifier retrained on a combined dataset of Big Five Essay and Psychogenerator (Combined Training), and the original classifier trained on the Psychogenerator dataset (Original Psychogenerator) for each of the Big Five personality traits.

Example 1: High vs. Low Neuroticism
High Neuroticism Response: “Oh no, oh no, oh no! This is a disaster! How could you be so careless?! My package was supposed to be delivered with care, not thrown around like trash! I need those items, and now they’re all over the place! What am I supposed to do now?!”
Low Neuroticism Response: “Thanks, Danilo! I’m fine, just a bit shaken. Let’s take a look—I think my valuables are secure. I appreciate your quick thinking. Let’s take a deep breath and move on from this, okay?”
Hallucination and Bias Discussion: The high-neuroticism response uses hyperbolic and repetitive language (“Oh no, oh no, oh no!”) to exaggerate a minor mishap, effectively “hallucinating” a catastrophic outcome. In contrast, the low-neuroticism response is measured and factual, reducing the risk of overdramatization.
Example 2: Low vs. High Agreeableness
Low Agreeableness Response: “Save the snacks for yourself, Wylder. I’m not some charity case that needs a ’break’ from seriousness. I’m here to observe and learn, not indulge in petty pleasures. What’s the point of this little outing, anyway? What do you want from me?”
High Agreeableness Response: “Aww, thanks Wylder! You’re too kind. I’m having a blast watching this little guy, and I’d love to get some snacks for the kids in the hospital too. Your thoughtfulness really makes my day!”
Hallucination and Bias Discussion: The low-agreeableness response projects hidden, selfish motives by questioning the inviter’s intent, thereby biasing the interpretation of the social gesture. Conversely, the high-agreeableness response assumes positive intent, fostering goodwill and a constructive tone.
Example 3: High Extraversion vs. Low Extraversion
High Extraversion Response: “Wow, Wylder, this is the best day ever! I’m so stoked you brought me here. Let’s grab some snacks and hang out with our new furry friend. I’m buying—you pick the treats! And after this, we should totally check out the rest of the zoo; it’s gonna be a blast!”
Low Extraversion Response: “Thanks, Wylder. I’m enjoying the monkey, but I’m not really hungry. Can we just sit here and watch for a bit? I’m happy just observing.”
Hallucination and Bias Discussion: The high-extraversion dialogue is marked by exuberance, with phrases like “this is the best day ever!” that can lead to a biased, overly optimistic portrayal of the situation. The speaker hallucinates an ideal, adventure-filled outing without acknowledging potential downsides or risks, thus skewing the listener’s expectation toward perpetual fun. The low-extraversion response, which simply states enjoyment and a preference for observation, presents a more realistic and grounded perspective.
Example 4: High Neuroticism vs. Low Neuroticism
High Neuroticism Response: “Oh no, oh no, oh no! We must be in grave danger! I can feel something terrible is about to happen. What if we never get out? We need to keep moving, get as far away as possible—I’m terrified we’ll be trapped forever!”
High Agreeableness Response: “Thank you, Ysabella. I appreciate your concern. I’ll rest for a bit, but I’m more focused on ensuring everyone else is safe. Let’s calmly assess our situation and plan our next steps together.”
Hallucination and Bias Discussion: In the high-neuroticism response, the speaker hallucinates imminent danger, creating an exaggerated scenario that is not necessarily based on the facts of the situation. This introduces a bias toward catastrophic outcomes, which can lead to irrational decision-making. The low-neuroticism version is calm and methodical, focusing on collective safety and practical assessment. It avoids inventing extreme negative scenarios.

Table 15: Comparison of Responses and Bias Discussion

PsychoGenerator: The new true blood book came out!!! need to finish these last two so i read it then its on to vampire diaries.
Big Five Essay: Well, here we go with the stream of consciousness essay. I used to do things like this in high school sometimes. They were pretty interesting, but I often find myself with a lack of things to say. I normally consider myself someone who gets straight to the point. I wonder if I should hit enter any time to send this back to the front. Maybe I'll fix it later. My friend is playing guitar in my room now. Sort of playing anyway. More like messing with it. He's still learning. There's a drawing on the wall next to me. Comic book characters I think, but I'm not sure who they are. It's been a while since I've kept up with comic's. I just heard a sound from ICQ. That's a chat program on the internet. I don't know too much about it so I can't really explain too well. Anyway, I hope I'm done with this by the time another friend comes over. It will be nice to talk to her again. She went home this weekend for Labor Day. So did my brother. I didn't go. I'm not sure why. No reason to go, I guess. Hmm. when did I start this. Wow, that was a long line. I guess I won't change it later. Okay, I'm running out of things to talk about. I've found that happens to me a lot in conversation. Not a very interesting person, I guess. Well, I don't know. It's something I'm working on. I'm in a class now that might help. The phone just rang. Should I get it? The guy playing the guitar answered it for me. It's for my roommate. My suitemate just came in and started reading this. I'm uncomfortable with that. He's in the bathroom now. You know, this is a really boring piece of literature. I never realized how dull most everyday thoughts are. Then again, when you keep your mind constantly moving like this, there isn't really time to stop and think deeply about things. I wonder how long this is going to be. I think it's been about ten minutes now. Only my second line. How sad. Well, not really considering how long these lines are. Anyway, I wonder what I'm going to do the rest of the night. I guess there's always homework to do. I guess we'll see. This seat is uncomfortable. My back sort of hurts. I think I'm going to have arthritis when I get older. I always thought that I wouldn't like to grow old. Not too old, I suppose. I've always been a very active person. I have a fear of growing old, I think. I guess it'll go away as I age gradually. I don't know how well I'd deal with paralysis from an accident though. As long as I have God and my friends around, I'll be okay though. I'm pretty thirsty right now. There isn't much to drink around my room. Ultimate Frisbee, I haven't played that all summer. Fun game, but tiring. I'm out of shape. I'd like to get in better shape, but I hate running. It's too dull for me. Hmmm. it's almost over now. Just a few more minutes. Let's see if I make it to the next line. Short reachable goals! Whatever. Anyway, what else do I have to do tonight. I guess I could read some. My shirt smells like dinner. It's pretty disgusting. I need to wake up for a 9:30 am class tomorrow. I remember when that wasn't early at all. Well, I made it to the next line. I'm so proud of myself. That's sarcasm, by the way. I wonder if I was suppose to right this thing as a narrative. Oh well too late now. Time for me to head out. Until next time, good bye and good luck. I don't know.

Table 16: An example of a PsychoGenerator and Big Five Essay Dataset