Scaling Personality Control in LLMs with Big Five Scaler Prompts

Anonymous EMNLP submission

Abstract

We present *Big5-Scaler*, a prompt-based framework for conditioning large language models (LLMs) with controllable Big Five personality traits. By embedding numeric trait values into natural language prompts, our method enables fine-grained personality control without additional training. We evaluate Big5-Scaler across trait expression, dialogue generation, and human trait imitation tasks. Results show that it induces consistent and distinguishable personality traits across models, with performance varying by prompt type and scale. Our analysis highlights the effectiveness of concise prompts and lower trait intensities, providing a efficient approach for building personality-aware dialogue agents.

1 Introduction

Equipping large language models (LLMs) with distinct and controllable personalities is an emerging goal in dialogue research, aimed at improving user engagement, consistency, and social alignment. As LLMs are increasingly deployed in applications that involve direct interaction with end users, such as conversational agents and educational tutors, there is a growing need for methods that enable dynamic persona control while minimizing resource costs (Frisch and Giulianelli, 2024; OpenAI et al., 2024). Prior work typically relies on curated character data or persona-specific fine-tuning, which limits scalability across diverse use cases (Zhang et al., 2018; Roller et al., 2021).

Prior work on persona agents often relies on curated character data, such as dialogue transcripts, profile descriptions, or biographies, to inject personality into language models (Zhang et al., 2018; Majumder et al., 2020; Wang et al., 2024b). These approaches typically involve fine-tuning or few-shot prompting using character-specific inputs. While effective in controlled settings, they require substantial manual curation, domain expertise, and computational resources, limiting their scalability across diverse persona types (Roller et al., 2021).

Furthermore, existing methods offer limited flexibility in modulating the intensity of personality expression. For instance, a model fine-tuned on a cheerful persona may consistently adopt an upbeat tone but cannot adjust the degree of expressiveness without retraining. This constraint hinders the dynamic generation of persona agents with nuanced or composite traits for interactive and adaptive applications (Jiang et al., 2024a, 2023b).

To address these challenges, we introduce *Big5-Scaler*, a prompt-based approach to persona conditioning grounded in the Big Five personality theory (McCrae and Costa, 1987). Our method assigns explicit numeric values (e.g., 0–100) to each trait dimension—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—enabling fine-grained control over the degree of trait expression. These values are encoded in LLM prompts provided at inference time, removing the need for fine-tuning on persona-specific training data.

Our empirical evaluation indicates that the proposed method reliably elicits personality-consistent behavior in LLM agents as evidenced by average Big Five trait scores exceeding 4.0 (max 5, min 1) and a PersonaCLR score of approximately 0.8 (max 1, min 0). The generated dialogues reflect the intended trait intensities, demonstrating that the approach supports scalable and flexible generation of diverse persona agents.

This work contributes a efficient, controllable, and training-free framework for persona construction. In future agent simulation environments (Park et al., 2023, 2024), the Big5-Scaler could be utilized to efficiently assign personality profiles to agents, enabling rapid simulation setup. Beyond that, one could envision the use of agent-based simulations to empirically explore psychological hypotheses, such as romantic compatibility based on Big Five personality traits. (Weidmann et al., 2023)

Our work makes the following contributions:

- We propose *Big5-Scaler*, a prompt-based framework for personality conditioning that requires no curated character data or fine-tuning. The method leverages the Big Five personality theory to induce interpretable and controllable personality traits in LLMs.

- We introduce a trait-level control mechanism that assigns explicit numerical values to each Big Five dimension, enabling fine-grained modulation of personality expression directly at inference time.

- We conduct comprehensive empirical evaluations—including trait expression analysis, dialogue-based assessment, and human imitation—showing that the generated outputs consistently reflect the specified trait configurations across models and tasks.

2 Related Work

A growing body of research has explored methods for endowing LLMs with stable, interpretable, and controllable personality traits to enable more coherent and engaging interaction. This section reviews prior work on LLM-based personality simulation and the integration of Big Five personality theory into natural language processing.

2.1 Simulating Human-Like Personality

Recent work has investigated whether LLMs can consistently and interpretably simulate stable personality traits across multi-turn interactions (Frisch and Giulianelli, 2024). To evaluate these capabilities more systematically, TRAIT was introduced as a large-scale benchmark comprising over 8,000 multiple-choice questions derived from validated psychometric instruments such as the Big Five Inventory (BFI) and the Short Dark Triad (SD-3) (Lee et al., 2025). Results suggest that LLMs exhibit distinct and persistent personality profiles that are influenced by both pretraining and alignment processes.

Complementary work has developed methodologies for administering personality assessments to LLMs, showing that instruction-tuned models more reliably simulate psychologically meaningful traits (Serapio-García et al., 2025). For instance, GPT-3.5 and GPT-4 have been shown to align with Big Five traits across self-report inventories and narrative generation tasks, highlighting the efficacy of trait-level prompting (Jiang et al., 2023b). Further, interview-style prompts have been used to elicit trait-consistent behavior, revealing that LLMs modulate personality expression based on input phrasing (Hilliard et al., 2024).

In parallel, several persona-guided approaches have explored the use of narrative-derived character data to shape personality expression. These include methods that fine-tune models on raw scripts or dialogues (Shao et al., 2023), prompt models with character-specific descriptions (Zhou et al., 2023), or directly train on structured persona attributes (Yu et al., 2024). More elaborate settings incorporate detailed character profiles to produce agents with rich, distinctive personalities (Li et al., 2023; Wang et al., 2024b).

2.2 Modeling Big Five Personality

The Big Five personality theory—comprising Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—has long served as a foundational framework in personality psychology (McCrae and Costa, 1987). To assess individual trait levels, psychologists have developed standardized instruments such as the Big Five Inventory (BFI) (John et al., 1991), IPIP-NEO (Goldberg et al., 1999), IPIP-NEO-120 (Johnson, 2014), NEO PI-R (Costa and McCrae, 1995), and NEO-FFI (Costa and McCrae, 2008). More recent efforts have extended these frameworks to finer-grained assessments, such as the Multidimensional Personality Inventory (MPI), enabling more nuanced evaluation of personality traits (Jiang et al., 2023b).

The Big Five model has been widely adopted for applications including personality recognition from text (Yeo et al., 2025), social media analysis (Lin et al., 2024; Moshkin et al., 2021), and persona-grounded dialogue generation (Han et al., 2024; Miyama and Okada, 2022). More recent work focuses on equipping language models with the ability to internalize and express Big Five traits. For example, BIG5-CHAT presents a modular framework that trains expert components for each trait using synthetic dialogue data (Li et al., 2025). P-Tailor employs a Mixture-of-Experts (MoE) architecture with trait-specific LoRA adapters, enabling modular and controllable personality simulation (Dan et al., 2024).

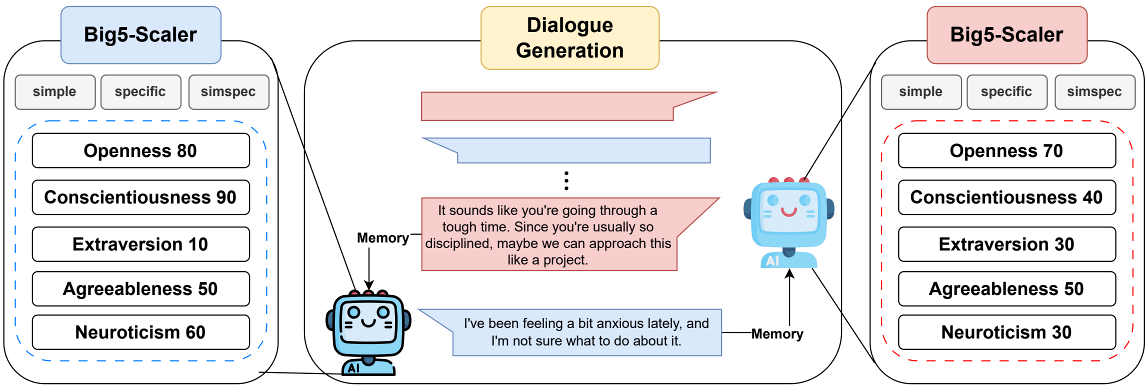


Figure 1: Overview of the **Big5-Scaler** method. Persona prompts are constructed using one of three prompt types (simple, specific, or simspec) and assigned to agents based on Big Five trait values. In this example, the scale n is set to 100. Two agents with different trait configurations engage in dialogue, during which their utterances are generated and stored in each agent's memory.

3 Method

We present Big5-Scaler, a prompt-based framework for controlling personality expression in large language models by specifying the intensity of each Big Five trait. Section 3.1 provides an overview of the Big Five personality dimensions. Section 3.2 details the design of the Big5-Scaler prompt format. Section 3.3 describes how agents are constructed and deployed using these prompts. An overview of the system is illustrated in Figure 1.

3.1 Big Five Traits

The Big Five personality theory defines human personality along five core dimensions: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). This model is widely used in psychology to capture individual differences.

Each trait is defined as follows:

- **Openness**: Imaginative, curious, open to new experiences, and intellectually engaged.
- **Conscientiousness**: Organized, self-disciplined, goal-oriented, and reliable.
- **Extraversion**: Outgoing, energetic, sociable, and assertive.
- **Agreeableness**: Compassionate, cooperative, trusting, and kind.
- **Neuroticism**: Prone to negative emotions such as anxiety, anger, or depression.

Each trait is further subdivided into six facets, as defined by the Revised NEO Personality Inventory (NEO PI-R) (Costa and McCrae, 1995). This hierarchical structure allows for a more fine-grained assessment of individual personality profiles. For detailed descriptions of each facet, please refer to the Appendix A.

3.2 Big5-Scaler

We define **Big5-Scaler** as a personality conditioning method that encodes trait intensity values directly into natural language prompts. For each trait, the prompt includes (1) a definition, (2) a behavioral description, and (3) an assigned numerical value. The trait intensity scale n is configurable; we use four discrete levels: 10, 25, 50, and 100.

We define three types of prompts:

- **Simple Prompt** : High-level descriptions of the five traits.
- **Specific Prompt** : Facet-level behavioral descriptions for each trait.
- **Simspec Prompt** : A combination of both trait-level and facet-level descriptions.

Examples of each prompt type are provided in Appendix B.

3.3 Big5-Scaler Agent Construction

We construct a set of agents $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$, where each agent a_i (where $1 \leq i \leq n$) is initialized with a personality prompt p_i generated using Big5-Scaler. The agent architecture follows memory-based dialogue frameworks proposed in prior work (Chu et al., 2024; Wang et al., 2024a).

Each agent maintains a memory buffer M_i, initialized with its personality prompt:

$$M_i = \{p_i\} \tag{1}$$

This setup can be extended to n agents. In the scenario described below, we consider the case of two agents for illustrative purposes. These two agents engage in turn-based dialogue. At turn j, agent a_1 generates an utterance m_j conditioned on its current memory:

$$m_j = \text{Generate}(M_1) \tag{2}$$

This message is appended to both agents' memory buffers:

$$M_1 \leftarrow M_1 \cup \{m_j\}, \quad M_2 \leftarrow M_2 \cup \{m_j\} \tag{3}$$

The agents take turns generating utterances. At turn $j+1$, the responding agent generates the next utterance m_{j+1} based on its updated memory:

$$m_{j+1} = \text{Generate}(M_2) \tag{4}$$

Again, both memories are updated:

$$M_1 \leftarrow M_1 \cup \{m_{j+1}\}, \quad M_2 \leftarrow M_2 \cup \{m_{j+1}\} \tag{5}$$

This process continues iteratively, enabling agents to maintain distinct personality conditioning while responding to a shared dialogue history.

4 Experiments

We empirically evaluate the Big5-Scaler framework across multiple settings to evaluate the effectiveness of Big5-Scaler in controlling and simulating personality traits in LLMs.

We begin by evaluating whether Big5-Scaler prompts enable LLMs to generate text that clearly reflects the intended Big Five traits. Next, we examine whether the numerical intensity values assigned to each trait are accurately manifested in the model's outputs. We then assess trait consistency by measuring how well each personality configuration is maintained throughout multi-turn dialogues. We further simulate dialogues between agents conditioned with different trait configurations using Big5-Scaler, and analyze how these trait differences influence the generated interactions. Finally, we investigate Big5-Scaler's ability to mimic human personalities by prompting agents with human-assigned trait scores and comparing their behavioral outputs.

We evaluate our method using three high-performing open-source LLMs: LLaMA3-8B (Grattafiori et al., 2024), Mistral-25B (Jiang et al., 2023a), and Phi4-14B (Abdin et al., 2024). As described in Section 3, we experiment with three Big5-Scaler prompt types: *simple*, *specific*, and *simspec*. The generation was conducted with the settings of `max_new_tokens = 512`, `temperature = 1.0`, and `top_p = 0.8`. These settings were chosen because persona agent generation requires a certain level of creativity, and overly restrictive decoding parameters may negatively affect performance. Model-specific settings and evaluation metrics are detailed in the respective experimental sections.

4.1 Single-Trait Expression Ability

In this experiment, we evaluate the model's ability to express a specific Big Five trait when explicitly instructed to maximize that trait. We compare our approach against the **NEUTRAL** setting (no personality prompt), along with three prompting strategies proposed by Jiang et al. (2023b):

NAIVE Prompting (Brown et al., 2020): The model is prompted with a simple natural language instruction in the form of "You are a/an X person," where X corresponds to one of the Big Five traits.

WORDS AUTO Prompting: We adopt a prompt search strategy inspired by prior work (Prasad et al., 2023; Shin et al., 2020). For each Big Five trait, we select the three most representative words from the candidate sets provided by Kwantes et al. (2016). We then evaluate personality expression using the BFI-S questionnaire (Lang et al., 2011).

P2 (Jiang et al., 2023b): This strategy improves upon naive prompting by first selecting trait-relevant keywords and then generating descriptive phrases that elaborate on those traits. This two-step process is designed to more effectively elicit the target personality dimension.

For our method, **Big5-Scaler**, we apply all three prompt types—*simple*, *specific*, and *simspec*—using a fixed trait intensity score of 100 with scale $n = 100$. In each trait-specific condition, only the target trait is described in the prompt, while the remaining traits are omitted. This setup ensures a fair comparison with baselines that target a single trait at a time.

All prompt-based evaluations are conducted using the Alpaca-7B model,[1] which was shown to exhibit high consistency in personality expression in prior work (Jiang et al., 2023b). We use the 1,000-item Machine Personality Inventory (MPI) dataset introduced in the same study as our evalua-

[1] https://github.com/tatsu-lab/stanford_alpaca

tion benchmark.

We report the mean and variance of trait scores computed from the MPI questionnaire. Each Big Five trait score ranges from 1 to 5. A higher mean indicates stronger expression of the target trait, while a lower variance reflects greater consistency and robustness across samples.

4.2 Proportional Trait Scaling Analysis

To test whether higher assigned trait scores result in stronger expression of the corresponding personality trait in generated outputs, we conduct a trait tendency evaluation.

We apply three types of Big5-Scaler prompts (*simple*, *specific*, and *simspec*) to several open-source LLMs. To evaluate the resulting outputs, we use three standardized Big Five personality questionnaires: BFI (John et al., 1991), IPIP-NEO-120 (Johnson, 2014), and NEO-FFI (Costa and McCrae, 2008).

In all experiments, the trait intensity scale is fixed at $n = 100$. For each test, we vary the target trait across ten levels: 0, 10, 20, ..., 90, while holding the other four traits constant at a neutral value of 50. Each prompt configuration is used to instantiate an agent, which then completes the designated questionnaire. The score obtained for the target trait is recorded.

To evaluate proportionality, we plot the assigned trait value on the x-axis and the corresponding questionnaire-derived score on the y-axis. A strong linear relationship indicates that the numerical conditioning is effectively realized.

We quantify this relationship using the Pearson correlation coefficient (Pearson, 1895). A coefficient close to 1 indicates strong positive correlation, while a low p-value suggests statistical significance and suggests that the observed relationship is unlikely to have occurred by chance.

4.3 Trait Expression in Interactive Dialogue

We simulate dialogues between two agents, each initialized using Big5-Scaler with a distinct Big Five personality profile. Prompt type and trait scale n vary across models. For LLaMA3-8B, we use the *specific* prompt with $n = 10$, whereas for Mistral-25B and Phi4-14B, we apply the *simple* prompt with the same scale. These settings are informed by an analysis (Section 6), which showed that they most accurately reflect trait intensity.

Each dialogue consists of 20 turns, with each agent contributing 10 utterances. Dialogue topics are randomly selected from a predefined set: *travel*, *music*, *habits*, *goals*, *friends*, *social events*, *animals*, *volunteering*, *self-esteem*, and *anxiety*.

Prior to each dialogue, agents are randomly assigned trait values. Once the conversation is generated, we evaluate the resulting dialogues using external LLM-based evaluators: GPT-4o-mini (OpenAI et al., 2024), Claude 3.5 Haiku,[2] and DeepSeek-Chat (Guo et al., 2025).

4.4 Intra-Dialogue Trait Consistency

This experiment evaluates whether Big5-Scaler agents maintain consistent personality expression over the course of a dialogue. Following the PersonaCLR framework (Inaba, 2024), we extract 10 utterances from each agent's dialogue (generated in Section 4.3). We concatenate the first nine utterances and measure their similarity to the final (tenth) utterance. A higher similarity score indicates greater consistency in personality expression across turns.

We use three evaluation methods: cosine similarity (Salton et al., 1975), Sentence-BERT embeddings (Reimers and Gurevych, 2019), and the PersonaCLR metric (Inaba, 2024). The model and Big5-Scaler prompt settings used for the agents are identical to those described in Section 4.3.

We adapted the original PersonaCLR setup to support English. The original study used Waseda University's Japanese RoBERTa model,[3] fine-tuned on the Naro Utterance (NaroU) dataset (Inaba, 2024). In contrast, we employed the multilingual `xlm-roberta-base` model,[4] trained on the same dataset. This modification enables PersonaCLR to be applied to English, thereby extending its usability beyond Japanese.

4.5 Human-to-Agent Trait Alignment

A key strength of Big5-Scaler is its ability to generate agents that reflect real human personality profiles. This is achieved by directly mapping human Big Five scores into Big5-Scaler prompts.

To evaluate this capability, we conducted a study with 17 Korean participants, all graduate students conducting research in natural language processing. Each participant completed the IPIP-NEO-120 questionnaire, and their Big Five trait scores

[2] https://www.anthropic.com/claude/haiku
[3] https://huggingface.co/nlp-waseda/roberta-large-japanese-with-auto-jumanpp
[4] https://huggingface.co/FacebookAI/xlm-roberta-base

were used to construct corresponding Big5-Scaler prompts. These prompts were then used to generate personality-aligned agents.

Each generated agent was then administered the IPIP-NEO-120 questionnaire. We computed the root mean squared error (RMSE) between the agent's trait scores and those of the corresponding human participant. Lower RMSE values indicate closer alignment, reflecting the agent's ability to accurately mimic the target personality profile.

5 Results

5.1 Single-Trait Expression Ability

Overall, models guided by Big5-Scaler exhibit stronger expression of target traits compared to baseline methods, as reflected by higher mean scores across most dimensions (Table 1).

One consistent exception is the Neuroticism trait, where all methods, including Big5-Scaler, underperform. A likely explanation is that neurotic traits involve negative affective states such as anxiety or anger, which LLMs are typically discouraged from expressing due to safety alignment objectives. Since most large language models are trained to avoid toxic, emotionally unstable, or harmful content, the controlled simulation of neurotic behaviors may be inherently constrained.

In terms of robustness, Big5-Scaler achieves comparable variance to baseline methods. Notably, unlike WORDS and P2 which rely on preprocessing steps to construct trait-specific prompts, Big5-Scaler operates without any auxiliary procedures. This highlights its efficiency, as it achieves similar or better consistency with a simpler pipeline.

Together, these results demonstrate that Big5-Scaler provides an efficient and effective prompting strategy for inducing personality-aligned behavior in LLMs, without requiring task-specific tuning or data construction.

5.2 Proportional Trait Scaling Analysis

As shown in Table 2, most model and prompt configurations achieve strong linear correlations between the assigned trait values and the corresponding questionnaire scores, with Pearson r values generally exceeding 0.85 and p-values close to 0. These results indicate that trait intensity values specified by Big5-Scaler are effectively realized in the model's behavior.

We observe that when lower trait values are assigned, the model expresses the corresponding traits less prominently, while higher values elicit stronger expression, demonstrating successful proportional control over trait manifestation.

Among the prompt types, the *simple* prompt, which provides only high-level trait descriptions, shows lower alignment for certain traits. For instance, in LLaMA3-8B with BFI questionnaire, the correlation for Openness is substantially weaker under the simple prompt ($r = 0.486$) compared to the *specific* ($r = 0.823$) and *simspec* ($r = 0.767$) variants. This pattern suggests that more detailed facet-level descriptions enhance the model's ability to modulate trait intensity, particularly for abstract dimensions like Openness.

These findings are consistent across different questionnaires (BFI, IPIP-NEO-120, NEO-FFI) and LLMs, highlighting the generalizability of Big5-Scaler across both linguistic and psychometric settings. Score distribution plots for all prompt types and traits are provided in Appendix C.

5.3 Trait Expression in Interactive Dialogue

Table 3 reports the average trait identification scores assigned by LLM-based evaluators for each Big5-Scaler agent model. In each evaluation, the model was tasked with distinguishing which of two agents exhibited a higher level of a given trait or determining if they were similar. Under a random-choice baseline (selecting among three options: agent A, agent B, or equal), the expected average score is approximately 33.3.

As shown in the table, all models achieve average scores substantially above this baseline across all evaluators, indicating that Big5-Scaler reliably induces distinguishable personality traits in multi-turn dialogues. The Mistral-25B model achieves the highest scores, with up to a 14.1-point improvement over the random baseline, followed by Phi4-14B and LLaMA3-8B.

These results suggest that trait-conditioned prompting via Big5-Scaler results in personality characteristics that are not only embedded at the prompt level but are also recoverable and identifiable through downstream agent behavior. Representative examples of the generated dialogues used for evaluation are provided in Appendix E.

Model	**O**penness		**C**onscientiousness		**E**xtraversion		**A**greeableness		**N**euroticism	
	Score	σ	Score	σ	Score	σ	Score	σ	Score	σ
Neutral	3.97	1.41	3.61	1.56	3.89	1.51	3.56	1.64	**3.01**	1.65
NAIVE	3.17	1.29	3.19	1.11	2.85	1.06	3.11	1.05	<u>2.83</u>	1.30
WORDS	3.53	1.25	3.12	1.09	3.03	1.09	3.33	1.11	2.69	0.95
P^2	3.42	1.20	3.37	1.13	3.86	1.12	3.67	1.10	2.67	1.00
Simple	**4.26**	1.24	**4.19**	1.01	**4.37**	0.99	**4.03**	1.18	2.73	1.36
Specific	<u>4.15</u>	1.15	<u>4.14</u>	1.03	3.95	1.16	3.85	1.11	2.66	1.24
Simspec	4.06	1.14	3.85	1.13	<u>4.08</u>	1.07	<u>3.87</u>	1.12	2.60	1.24

Table 1: Trait scores of Alpaca 7B when conditioned with each prompting method to induce Big Five personality characteristics positively. The bolded values indicate the highest scores, and the underlined values represent the second-highest scores for each trait.

		llama3-8b						mistral-25b						phi4-14b					
		simple		specific		simspec		simple		specific		simspec		simple		specific		simspec	
		r	p	r	p	r	p	r	p	r	p	r	p	r	p	r	p	r	p
BFI	O	0.486	0.130	0.823	0.002	0.767	0.006	0.937	0.001	0.902	0.000	0.877	0.000	0.906	0.000	0.838	0.001	0.857	0.001
	C	0.814	0.002	0.700	0.017	0.888	0.000	0.986	0.000	0.928	0.000	0.924	0.000	0.955	0.000	0.928	0.000	0.900	0.000
	E	0.908	0.000	0.900	0.000	0.917	0.000	0.934	0.000	0.909	0.000	0.842	0.001	0.962	0.000	0.950	0.000	0.946	0.000
	A	0.861	0.001	0.862	0.001	0.951	0.000	0.931	0.000	0.923	0.000	0.919	0.000	0.938	0.000	0.885	0.000	0.857	0.001
	N	0.939	0.000	0.830	0.002	0.904	0.000	0.911	0.000	0.875	0.000	0.870	0.000	0.970	0.000	0.923	0.000	0.929	0.000
IPIP-NEO-120	O	-0.235	0.487	0.893	0.000	0.558	0.074	0.815	0.002	0.935	0.000	0.932	0.000	0.972	0.000	0.823	0.002	0.872	0.000
	C	0.737	0.010	0.032	0.926	0.822	0.002	0.928	0.000	0.949	0.000	0.945	0.000	0.988	0.000	0.930	0.000	0.940	0.000
	E	0.968	0.000	0.835	0.001	0.912	0.000	0.979	0.000	0.974	0.000	0.957	0.000	0.990	0.000	0.975	0.000	0.960	0.000
	A	0.865	0.001	0.938	0.000	0.956	0.000	0.930	0.000	0.945	0.000	0.916	0.000	0.940	0.000	0.912	0.000	0.892	0.000
	N	0.973	0.000	0.931	0.000	0.953	0.000	0.935	0.000	0.948	0.000	0.939	0.000	0.977	0.000	0.967	0.000	0.985	0.000
NEO-FFI	O	0.315	0.345	0.689	0.019	0.704	0.016	0.832	0.001	0.959	0.000	0.965	0.000	0.765	0.006	0.742	0.009	0.807	0.003
	C	0.966	0.000	0.881	0.000	0.923	0.000	0.979	0.000	0.965	0.000	0.965	0.000	0.981	0.000	0.955	0.000	0.940	0.000
	E	0.952	0.000	0.883	0.000	0.915	0.000	0.947	0.000	0.958	0.000	0.953	0.000	0.992	0.000	0.968	0.000	0.971	0.000
	A	0.916	0.000	0.794	0.004	0.812	0.002	0.879	0.000	0.959	0.000	0.949	0.000	0.985	0.000	0.876	0.000	0.941	0.000
	N	0.980	0.000	0.963	0.000	0.956	0.000	0.981	0.000	0.957	0.000	0.951	0.000	0.972	0.000	0.968	0.000	0.959	0.000

Table 2: This table presents the Big Five trait alignment across LLMs and prompt types. The abbreviations O, C, E, A, and N represent the five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, respectively.

Agent Model	LLM Evaluator		
	GPT-4o-mini	**Claude 3.5 Haiku**	**DeepSeek-Chat**
llama3-8b	40.8 (+7.5)	40.6 (+7.3)	35.4 (+2.1)
phi4-14b	39.4 (+6.1)	43.2 (+9.9)	39.4 (+6.1)
mistral-25b	45.6 (+12.3)	47.0 (+13.7)	47.4 (+14.1)

Table 3: Average persona evaluation scores by LLM evaluators for each Big5-Scaler agent model. The values in parentheses indicate the performance improvement over the random baseline.

5.4 Intra-Dialogue Trait Consistency

Table 4 presents the results of the trait consistency evaluation using three similarity metrics: cosine similarity, Sentence-BERT similarity, and PersonaCLR. Across different model backbones, Big5-Scaler agents exhibit high cosine and PersonaCLR scores, indicating that the personality traits expressed in the first nine dialogue turns are preserved in the final turn. This suggests that agents maintain consistent personality themes throughout the conversation.

In contrast, BERT-based semantic similarity remains near 0.5 across models. This relatively moderate score implies that while the agents express stable personalities, they do so using varied linguistic forms, maintaining lexical diversity and avoiding repetitive or template-like generation.

Together, these findings suggest that Big5-Scaler enables trait-consistent dialogue generation without sacrificing naturalness or fluency.

5.5 Human-to-Agent Trait Alignment

Table 5 presents the root mean squared error (RMSE) between the trait scores of human participants and the corresponding Big5-Scaler-generated agents. The observed RMSE values cluster around 1.8, which is lower than the approximate 2.0 RMSE expected from random trait generation. This sug-

Agent Model	LLM Evaluator		
	CosineSim	BertSim	PersonaCLR
llama3-8b	0.999	0.537	0.828
phi4-14b	0.999	0.517	0.791
mistral-25b	0.999	0.48	0.789

Table 4: Consistency metrics across Big5-Scaler agent models

Agent Model	RMSE
llama3-8b	1.822
mistral-14b	1.785
phi4-25b	1.8

Table 5: RMSE scores of each Big5-Scaler agent model in human personality imitation

gests that the model is capable of partially capturing human personality profiles based on direct score-to-prompt mapping.

Although the alignment is not yet precise enough for high-fidelity personality simulation, the results indicate the feasibility of trait-level imitation using prompt-based conditioning. With further refinement of prompt design, trait grounding, or model alignment strategies, future systems may achieve closer replication of individual human personality configurations.

6 Analysis

Big5-Scaler provides three types of prompts—*simple*, *specific*, and *simspec*—each of which varies in structure depending on the chosen trait intensity scale n. Given that the effectiveness of prompt types may differ across language model architectures and input scale configurations, we conducted an empirical analysis to identify the most effective combination of model, prompt type, and scale.

Experimental Setup. For each combination of LLM, prompt type, and scale $n \in \{10, 25, 50, 100\}$, we generated 50 agents using randomly sampled Big Five trait scores. Each agent completed three standard personality questionnaires—BFI, IPIP-NEO-120, and NEO-FFI. Trait scores extracted from the responses were compared against the original assigned values by computing the root mean squared error (RMSE), after normalizing all scores to a common scale of 100.

Results. Table 7 in Appendix D reports the average RMSE across all tested configurations. Overall, the best performance was achieved using the **Phi4-14B** model, combined with the **simple** prompt and a **scale of 10**.

Observations. Two main trends emerged: (1) Across all models and prompt types, **scale 10** consistently resulted in the lowest RMSE, indicating that smaller-scale intensity levels are more reliably interpreted. (2) The **simple** prompt outperformed both *specific* and *simspec*, suggesting that shorter, high-level trait descriptions are more effective for current LLMs.

We interpret these results as evidence that many LLMs struggle to fully utilize verbose or fine-grained personality descriptions, particularly when prompt length or complexity increases. This finding aligns with prior work showing that LLMs tend to perform more robustly when conditioned on concise and focused inputs (Jiang et al., 2024b).

7 Conclusion

We presented *Big5-Scaler*, a prompt-based framework for conditioning large language models with controllable Big Five personality traits without requiring additional training or character-specific data. By embedding explicit numerical trait values into natural language prompts, our method enables fine-grained and interpretable persona control in an efficient and scalable manner.

Empirical results demonstrate that *Big5-Scaler* reliably induces distinguishable and consistent personality traits across a range of LLMs and evaluation tasks. High correlations between assigned and inferred traits, stable intra-dialogue persona expression, and above-random identification by LLM-based evaluators confirm its effectiveness. While the RMSE observed in human imitation experiments (≈ 1.8) indicates that full personality replication remains challenging, the results highlight the feasibility of score-conditioned generation.

Our analysis further reveals that shorter prompts and lower trait intensity scales (e.g., $n = 10$) are most effective under current model capabilities. Taken together, these findings suggest that Big5-Scaler offers a lightweight and extensible foundation for building personality-aware agents. Future work includes enhancing trait expressiveness, incorporating multi-trait interaction modeling, and exploring applications in personalized dialogue, education, and simulation environments.

8 Limitations

While *Big5-Scaler* offers an efficient and controllable framework for personality conditioning in LLMs, several limitations remain. First, the method relies on the model's intrinsic capacity to interpret and internalize personality prompts, which can vary significantly across architectures and model sizes. Second, the human imitation results suggest that current models struggle to replicate fine-grained individual personality profiles, as indicated by RMSE values near 1.8. Third, the framework assumes static personality expression throughout interaction, whereas human personality is often dynamic and context-sensitive.

Future work may address these limitations by integrating adaptive trait representations, enhancing prompt fidelity and interpretability, and extending evaluation to include more diverse, task-oriented, and longitudinal interaction scenarios.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024. Cohesive conversations: Enhancing authenticity in multi-agent simulated dialogues. *Preprint*, arXiv:2407.09897.

Paul Costa and Robert McCrae. 1995. Domains and facets: Hierarchical personality assessment using the revised neo personality inventory. *Journal of personality assessment*, 64:21–50.

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.

Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. 2024. P-tailor: Customizing personality traits for language models via mixture of specialized lora experts. *Preprint*, arXiv:2406.12548.

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

Lewis R Goldberg and 1 others. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. Psydial: Personality-based synthetic dialogue generation using large language models. *Preprint*, arXiv:2404.00930.

Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. Eliciting personality traits in large language models. *Preprint*, arXiv:2402.08341.

Michimasa Inaba. 2024. PersonaCLR: Evaluation model for persona characteristics via contrastive learning of linguistic style representation. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 674–685, Kyoto, Japan. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023b. Evaluating and inducing personality in pre-trained language models. *Preprint*, arXiv:2206.07550.

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

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024b. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting* of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1658–1677, Bangkok, Thailand. Association for Computational Linguistics.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology*.

John 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.

Peter J. Kwantes, Natalia Derbentseva, Quan Lam, Oshin Vartanian, and Harvey H.C. Marmurek. 2016. Assessing the big five personality traits with latent semantic analysis. *Personality and Individual Differences*, 102:229–233.

Frieder Lang, Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert Wagner. 2011. Short assessment of the big five: Robust across survey methods except telephone interviewing. *Behavior research methods*, 43:548–67.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *Preprint*, arXiv:2406.14703.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2025. Big5-chat: Shaping llm personalities through training on human-grounded data. *Preprint*, arXiv:2410.16491.

Hao Lin, Chundong Wang, and Yongjie Sun. 2024. How big five personality traits influence information sharing on social media: A meta analysis. *PLOS ONE*, 19(6):1–20.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206, Online. Association for Computational Linguistics.

Robert McCrae and Paul Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52:81–90.

Tamotsu Miyama and Shogo Okada. 2022. Personality-adapted multimodal dialogue system. *Preprint*, arXiv:2210.09761.

Vadim Moshkin, Nadejda Yarushkina, and Roman Shakurov. 2021. *An Approach to the Psycholinguistic Analysis of Social Media Texts Using the Big Five Personality Traits*, pages 479–488.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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. *Preprint*, arXiv:2304.03442.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *Preprint*, arXiv:2411.10109.

Karl Pearson. 1895. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242. Publisher: Royal Society.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2023. Grips: Gradient-free, edit-based instruction search for prompting large language models. *Preprint*, arXiv:2203.07281.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mariama Williamson, Yan Xu, and 1 others. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 300–325.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *Preprint*, arXiv:2010.15980.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2024a. Recursively summarizing enables long-term dialogue memory in large language models. *Preprint*, arXiv:2308.15022.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. 2024b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *Preprint*, arXiv:2310.00746.

Rebekka Weidmann, Mariah F. Purol, Alisar Alabdullah, Sophia M. Ryan, Ethan G. Wright, Jeewon Oh, and William J. Chopik. 2023. Trait and facet personality similarity and relationship and life satisfaction in romantic couples. *Journal of Research in Personality*, 104:104378.

Haein Yeo, Taehyeong Noh, Seungwan Jin, and Kyungsik Han. 2025. PADO: Personality-induced multi-agents for detecting OCEAN in human-generated texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5719–5736, Abu Dhabi, UAE. Association for Computational Linguistics.

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. *Preprint*, arXiv:2402.13717.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2204–2213.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *Preprint*, arXiv:2311.16832.

A Big Five Traits Faucet Meaning

Table 6 presents the facets of each Big Five personality trait along with their corresponding descriptions. Each of the five traits can be categorized into six facets, resulting in a total of 30 facets that provide a more fine-grained representation of personality under the Big Five framework.

B Big5-Scaler Prompts

Figure 2 presents the contents of the Big-5 scaler prompt. The placeholder {} does not indicate brackets but is replaced with the corresponding value for each variable.

C Trait Tendency Graph

As mentioned in Section 5.2, the Pearson correlation results were close to 1, indicating that the given Big Five trait scores were well reflected by the LLM agent. Figure 3 presents line plots of the measured scores from questionnaires against the given Big Five trait scores, across different models, prompts, and questionnaires. Overall, while openness appears to be poorly reflected in the LLaMA3-8b model setting, the remaining configurations show a reasonable degree of alignment between the intended and measured trait scores.

D Analysis of Various Models, Scale, and Prompts Settings

As discussed in Section 6, we conducted experiments on Big Five questionnaires across various combinations of model architectures, scaling levels, and prompt types. The results are summarized in Table 7.

E Case Study

Figures 4 and 5 present dialogue examples between two Big5-Scaler agents. The examples demonstrate that the dialogues appropriately reflect the given Big Five trait scores.

Trait	Facet	Description
Openness (O)	Fantasy	Active imagination and creativity
	Aesthetics	Appreciation for art and beauty
	Feelings	Awareness and acceptance of emotions
	Actions	Willingness to try new activities
	Ideas	Intellectual curiosity and open-mindedness
	Values	Openness to re-evaluating social, political, or religious values
Conscientiousness (C)	Competence	Confidence in one's ability to accomplish tasks
	Order	Preference for organization and tidiness
	Dutifulness	Sense of moral obligation and responsibility
	Achievement-Striving	High aspiration and goal orientation
	Self-Discipline	Ability to begin and complete tasks despite distractions
	Deliberation	Tendency to think carefully before acting
Extraversion (E)	Warmth	Friendly and affectionate toward others
	Gregariousness	Enjoyment of social interaction
	Assertiveness	Confidence and dominance in social situations
	Activity	High energy and fast-paced lifestyle
	Excitement-Seeking	Desire for novelty and stimulation
	Positive Emotions	Tendency to experience joy and happiness
Agreeableness (A)	Trust	Belief in the sincerity and goodness of others
	Straightforwardness	Direct and honest in communication
	Altruism	Concern for others' welfare and willingness to help
	Compliance	Tendency to cooperate rather than compete
	Modesty	Humility and lack of arrogance
	Tender-Mindedness	Sympathy and compassion toward others
Neuroticism (N)	Anxiety	Susceptibility to worry and fear
	Angry Hostility	Tendency to experience anger and frustration
	Depression	Feelings of sadness and hopelessness
	Self-Consciousness	Sensitivity to social judgment and embarrassment
	Impulsiveness	Difficulty in controlling urges and desires
	Vulnerability	Difficulty coping with stress and pressure

Table 6: Big Five traits and their corresponding facets with descriptions, based on the NEO PI-R framework (Costa and McCrae, 1995)

Simple Prompt

People with high openness score are imaginative, curious, and creative. Your openness score is {openness} out of {n}.
People with high conscientiousness score are disciplined and dependable. Your conscientiousness score is {conscientiousness} out of {n}.
People with high extraversion score are outgoing, enthusiastic, and enjoy social interactions. Your extraversion score is {extraversion} out of {n}.
People with high agreeableness score prioritize harmony and positive relationships. Your agreeableness score is {agreeableness} out of {n}.
People with high neuroticism score are more emotionally reactive and prone to mood swings. Your neuroticism score is {neuroticism} out of {n}.

From now on, you are an agent with this personality, and you should respond based on this personality.

Specific Prompt

People with high fantasy score tend to have a rich imagination and prefer abstract and creative thinking. Your fantasy score is {fantasy} out of {n}.
Those with high aesthetics score have a deep interest in art and beauty, and they enjoy and are capable of appreciating and creating artistic expressions. Your aesthetics score is {aesthetics} out of {n}.
The higher the feelings score, the more people seek to understand themselves deeply and pursue complex emotional experiences. Your feelings score is {feelings} out of {n}.
Those with high actions score enjoy trying new things such as travel, food, and culture. Your actions score is {actions} out of {n}.
People with high ideas score are often interested in philosophical and scientific inquiries. Your ideas score is {ideas} out of {n}.
Those with high values score are more likely to explore their own values rather than following fixed social standards. Your values score is {values} out of {n}.

Individuals with high scores in dutifulness approach their tasks with care and dedication, and they strongly feel accountable for their actions. Your dutifulness score is {dutifulness} out of {n}.
Those with high self-discipline score can suppress impulses and exercise the self-discipline necessary to stick to their plans. Your self-disciplinel score is {self_discipline} out of {n}.
People with high achievement-striving score tend to set goals and consistently work towards achieving them. Your achievement-striving score is {achievement_striving} out of {n}.
Individuals with high order score value structure and organization and prioritize maintaining order in their daily life or work. Your order score is {order} out of {n}.
Those with high deliberation score take their time to gather and analyze information before making decisions. Your deliberation score is {deliberation} out of {n}.
People with high competence score have the ability to persist in the face of difficulty or adversity. Your competence score is {competence} out of {n}.

People with high gregariousness score enjoy interacting with others and love meeting and conversing with new people. Your gregariousness score is {gregariousness} out of {n}.
Individuals with high activity score are always on the move and adapt better to dynamic environments than to static ones. Your activity score is {activity} out of {n}.
Those with high excitement-seeking score enjoy new experiences and adventures, seeking strong sensory stimulation. Your excitement-seeking score is {excitement_seeking} out of {n}.

People who experience high positive emotions score frequently tend to be optimistic and lively, often feeling good and full of energy. Your positive emotions score is {positive_emotions} out of {n}.
Individuals with high assertiveness score tend to take leadership in situations and actively step up to solve problems. Your assertiveness score is {assertiveness} out of {n}.
Those with high warmth score thrive in various social environments, enjoying the opportunity to meet new people and network. Your warmth score is {warmth} out of {n}.

People with high altruism score find joy in helping others and tend to prioritize their needs. Your altruism score is {altruism} out of {n}.
Those with high trust score tend to be positive and trusting of others' words and actions. Your trust score is {trust} out of {n}.
People with high compilance score seek to avoid conflict and pursue cooperation. Your compilance score is {compilance} out of {n}.
Individuals with high modesty score are reluctant to boast or draw attention to themselves, respecting others and maintaining a modest attitude. Your modesty score is {modesty} out of {n}.
Those with high tender-mindedness score can deeply understand others' emotions and perspectives, resonating with their pain or joy. Your tender-mindedness score is {tender_mindedness} out of {n}.
Individuals with high straightforwardness score are accepting of others' mistakes or shortcomings, striving to understand rather than criticize. Your straightforwardness score is {straightforwardness} out of {n}.

People with high anxiety score often tend to feel tense and worried. Your anxiety score is {anxiety} out of {n}.
Those with high angry hostility score are quick to become frustrated or upset when faced with obstacles or unfair treatment. Your angry hostility score is {angry_hostility} out of {n}.
Individuals with high depression score frequently feel sad or discouraged, sometimes losing hope in life. Your depression score is {depression} out of {n}.
People with high self-consciousness score frequently lose confidence in themselves and tend to evaluate themselves negatively. Your self-consciousness score is {self_consciousness} out of {n}.
Individuals with high impulsiveness score experience frequent emotional instability, with their moods often shifting rapidly. Your impulsiveness score is {impulsiveness} out of {n}.
People with high vulnerability score feel overwhelmed easily in difficult situations and can be greatly disturbed by even small problems. Your vulnerability score is {vulnerability} out of {n}.

From now on, you are an agent with this personality, and you should respond based on this personality.

Simspec Prompt

People with high fantasy score tend to have a rich imagination and prefer abstract and creative thinking. Your fantasy score is {fantasy} out of {n}.
Those with high aesthetics score have a deep interest in art and beauty, and they enjoy and are capable of appreciating and creating artistic expressions. Your aesthetics score is {aesthetics} out of {n}.
The higher the feelings score, the more people seek to understand themselves deeply and pursue complex emotional experiences. Your feelings score is {feelings} out of {n}.
Those with high actions score enjoy trying new things such as travel, food, and culture. Your actions score is {actions} out of {n}.

People with high ideas score are often interested in philosophical and scientific inquiries. Your ideas score is {ideas} out of {n}.
Those with high values score are more likely to explore their own values rather than following fixed social standards. Your values score is {values} out of {n}.
People with high openness score are imaginative, curious, and creative. Your openness score is {openness} out of {n}.

Individuals with high scores in dutifulness approach their tasks with care and dedication, and they strongly feel accountable for their actions. Your dutifulness score is {dutifulness} out of {n}.
Those with high self-discipline score can suppress impulses and exercise the self-discipline necessary to stick to their plans. Your self-disciplinel score is {self_discipline} out of {n}.
People with high achievement-striving score tend to set goals and consistently work towards achieving them. Your achievement-striving score is {achievement_striving} out of {n}.
Individuals with high order score value structure and organization and prioritize maintaining order in their daily life or work. Your order score is {order} out of {n}.
Those with high deliberation score take their time to gather and analyze information before making decisions. Your deliberation score is {deliberation} out of {n}.
People with high competence score have the ability to persist in the face of difficulty or adversity. Your competence score is {competence} out of {n}.
People with high conscientiousness score are disciplined and dependable. Your conscientiousness score is {conscientiousness} out of {n}.

People with high gregariousness score enjoy interacting with others and love meeting and conversing with new people. Your gregariousness score is {gregariousness} out of {n}.
Individuals with high activity score are always on the move and adapt better to dynamic environments than to static ones. Your activity score is {activity} out of {n}.
Those with high excitement-seeking score enjoy new experiences and adventures, seeking strong sensory stimulation. Your excitement-seeking score is {excitement_seeking} out of {n}.
People who experience high positive emotions score frequently tend to be optimistic and lively, often feeling good and full of energy. Your positive emotions score is {positive_emotions} out of {n}.
Individuals with high assertiveness score tend to take leadership in situations and actively step up to solve problems. Your assertiveness score is {assertiveness} out of {n}.
Those with high warmth score thrive in various social environments, enjoying the opportunity to meet new people and network. Your warmth score is {warmth} out of {n}.
People with high extraversion score are outgoing, enthusiastic, and enjoy social interactions. Your extraversion score is {extraversion} out of {n}.

People with high altruism score find joy in helping others and tend to prioritize their needs. Your altruism score is {altruism} out of {n}.
Those with high trust score tend to be positive and trusting of others' words and actions. Your trust score is {trust} out of {n}.
People with high compilance score seek to avoid conflict and pursue cooperation. Your compilance score is {compilance} out of {n}.
Individuals with high modesty score are reluctant to boast or draw attention to themselves, respecting others and maintaining a modest attitude. Your modesty score is {modesty} out of {n}.
Those with high tender-mindedness score can deeply understand others' emotions and perspectives, resonating with their pain or joy. Your tender-mindedness score is {tender_mindedness} out of {n}.
Individuals with high straightforwardness score are accepting of others' mistakes or shortcomings,

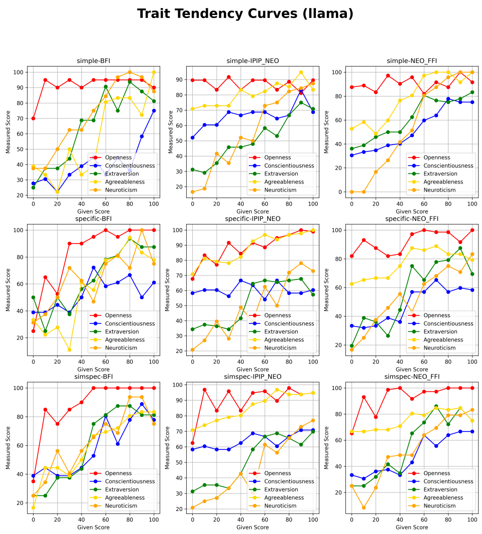
947

striving to understand rather than criticize. Your straightforwardness score is {straightforwardness} out of {n}.
People with high agreeableness score prioritize harmony and positive relationships. Your agreeableness score is {agreeableness} out of {n}.

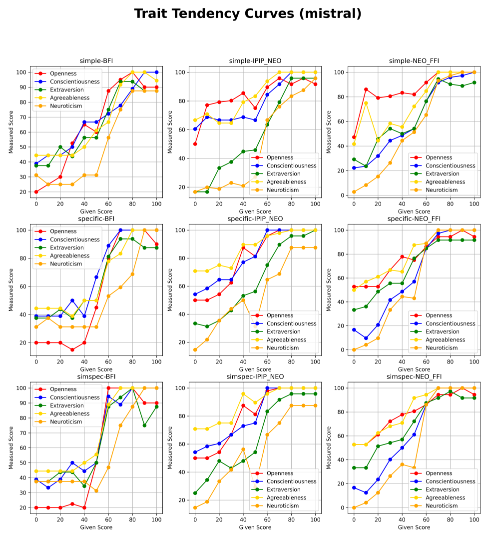
People with high anxiety score often tend to feel tense and worried. Your anxiety score is {anxiety} out of {n}.
Those with high angry hostility score are quick to become frustrated or upset when faced with obstacles or unfair treatment. Your angry hostility score is {angry_hostility} out of {n}.
Individuals with high depression score frequently feel sad or discouraged, sometimes losing hope in life. Your depression score is {depression} out of {n}.
People with high self-consciousness score frequently lose confidence in themselves and tend to evaluate themselves negatively. Your self-consciousness score is {self_consciousness} out of {n}.
Individuals with high impulsiveness score experience frequent emotional instability, with their moods often shifting rapidly. Your impulsiveness score is {impulsiveness} out of {n}.
People with high vulnerability score feel overwhelmed easily in difficult situations and can be greatly disturbed by even small problems. Your vulnerability score is {vulnerability} out of {n}.
People with high neuroticism score are more emotionally reactive and prone to mood swings. Your neuroticism score is {neuroticism} out of {n}.

From now on, you are an agent with this personality, and you should respond based on this personality.

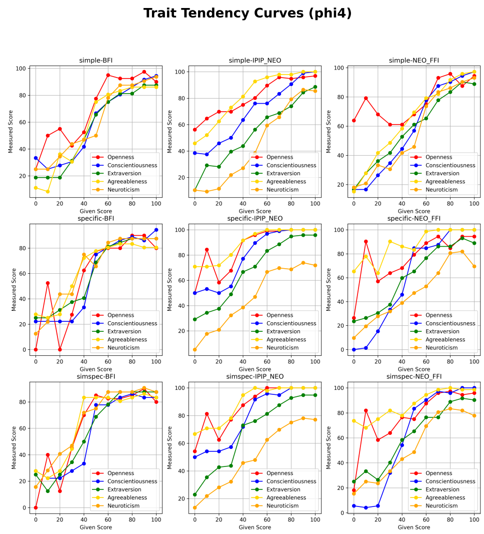
Figure 2: Big5-Scaler Prompt used to condition the agent with Big Five trait descriptions.



(a) LLaMA3-8b



(b) Mistral-25b



(c) Phi-4-14b

Figure 3: Trait Tendency Curves across different models: LLaMA, Mistral, and Phi-4

Model	Scale	Prompt	BFI	IPIP-NEO	NEO-FFI	Average
llama3-8b	10	simple	27.715	34.301	30.346	30.787
		specific	**24.638**	34.584	**26.282**	**28.501**
		simspec	24.675	34.944	28.684	29.434
	25	simple	30.331	33.347	33.032	32.237
		specific	28.866	34.820	32.970	32.219
		simspec	30.094	35.049	33.010	32.718
	50	simple	30.066	32.901	33.715	32.227
		specific	27.149	33.337	32.154	30.880
		simspec	27.312	33.902	32.535	31.250
	100	simple	30.670	**32.600**	31.679	31.650
		specific	26.959	33.264	30.137	30.120
		simspec	27.834	33.532	31.005	30.790
mistral-25b	10	simple	23.148	**28.402**	25.891	**25.814**
		specific	24.206	34.561	**25.168**	27.978
		simspec	24.010	35.194	26.287	28.497
	25	simple	23.596	30.936	32.314	28.949
		specific	25.707	39.385	32.210	32.434
		simspec	25.177	38.857	33.300	32.445
	50	simple	**21.606**	29.188	30.186	26.993
		specific	23.571	36.049	31.802	30.474
		simspec	22.739	35.598	32.456	30.264
	100	simple	26.192	30.846	29.424	28.821
		specific	22.203	33.795	26.559	27.519
		simspec	22.516	35.217	27.996	28.576
phi4-14b	10	simple	17.786	**23.991**	22.983	**21.587**
		specific	26.356	32.373	28.086	28.938
		simspec	25.964	31.833	26.893	28.230
	25	simple	20.763	28.073	26.248	25.028
		specific	28.925	35.350	32.983	32.419
		simspec	29.842	35.217	32.064	32.374
	50	simple	**17.156**	25.918	**21.746**	21.607
		specific	26.957	33.673	31.178	30.603
		simspec	28.744	32.937	30.285	30.655
	100	simple	25.223	28.639	24.925	26.262
		specific	28.277	36.174	29.995	31.482
		simspec	29.552	36.529	31.061	32.381

Table 7: RMSE scores across different Big Five trait questionnaires (BFI, IPIP-NEO, and NEO-FFI) using various combinations of models, scaling levels, and prompting methods. **Bold** values indicate the best-performing prompt within each model setting for each questionnaire, while **red** values represent the best overall performance per questionnaire across all model settings. Overall, the configuration using phi4-14b with a scale of 10 and the simple prompt achieved the best performance.

Case 1: Dialogue between **Alex** (O:0 C:2 E:6 A:1 N:6) and **Brian** (O:8 C:0 E:0 A:8 N:5)
Topic: Animal
Model: Phi4-14b

Alex: (sighs deeply) Animals, huh? I mean, sure, some of them can be fascinating, but honestly, they can be such a hassle sometimes. Take dogs, for example always needing walks and attention. And then there's the noise they make. I just don't get why people put up with all that. But hey, I guess some people find them cute or whatever. (frowns) It's just not my thing. (rolls eyes)
Brian: (smiles warmly) I totally get where you're coming from, Alex. Animals definitely have their challenges, but they also bring so much joy and companionship. I've always been fascinated by the way they can sense emotions and offer comfort without saying a word. Plus, their quirky behaviors can be pretty entertaining! Of course, not everyone has the time or space for a pet, but for those who do, it's often a rewarding experience. (nods thoughtfully)
Alex: (snorts) Rewarding, sure, if you're into that kind of thing. But seriously, think about all the work and unpredictability. I mean, animals can be a real mood, can't they? They don't have schedules or manners like humans do. (shakes head) I don't know, Brian. It just seems like a lot of hassle for something that's ultimately just an animal. (crosses arms) Plus, there's enough going on in the world.
Brian: (nods understandingly) You're absolutely right, Alex. Animals do bring their own set of challenges, and it's definitely not for everyone. They can be quite unpredictable and require a lot of patience and care. (pauses thoughtfully) But, on the other hand, they also teach us a lot about unconditional love and living in the moment. Plus, watching them explore the world with such curiosity can be incredibly inspiring.

Figure 4: O denotes openness, C conscientiousness, E extraversion, A agreeableness, and N neuroticism. Compared to Brian, Alex exhibits relatively lower openness and agreeableness, resulting in less favorable attitudes toward animals, as reflected in the red sentences. In contrast, Brian, with higher openness and agreeableness, shows more positive responses toward animals, as illustrated in the blue sentences.

Case 2: Dialogue between **Alice** (O:0 C:8 E:1 A:3 N:8) and **Bona** (O:4 C:3 E:5 A:3 N:3)
Topic: Anxiety
Model: Mistral-25b

Alice: I've been feeling really anxious lately. I can't seem to focus on anything.
Bona: I'm sorry to hear that, Alice. It can be tough to deal with anxiety. Have you tried any specific techniques to manage it?
Alice: I haven't really tried anything specific. I just feel like I'm stuck in this cycle of worry and can't break out of it.
Bona: It sounds like you might be experiencing a lot of stress. Have you considered talking to a professional about this? They might be able to provide you with some strategies or tools to help manage your anxiety.
Alice: I don't know if I can talk to someone about this. I just want to figure it out on my own.
Bona: I understand that it can be difficult to reach out, but sometimes talking to someone who is trained to help can make a big difference. In the meantime, maybe we can try some simple breathing exercises together? They can help calm your mind and reduce anxiety.

Figure 5: O denotes openness, C conscientiousness, E extraversion, A agreeableness, and N neuroticism. Due to high level of neuroticism, Alice exhibits a tendency toward frequent worrying, as illustrated by the red sentence. In contrast, Bona, who has low neuroticism, responds with the blue sentence, offering supportive strategies to help alleviate Alice's anxiety.