Big5-Scaler: Fine-Grained Big Five Conditioning for Socially Grounded LLM Agents

Anonymous ACL submission

Abstract

As LLMs increasingly serve as interactive agents, specifying behavior along interpretable personality dimensions becomes important. We introduce Big5-Scaler, a training-free prompting method that uses numeric Big Five values as an interpretable interface for inducing personality-related behavioral cues. We evaluate Big5-Scaler through questionnaire-based controllability, multi-agent dialogue, and task-specific profile search. Assigned Big Five values are consistently associated with questionnaire-derived scores across models and instruments, though effects vary by model, trait, and prompt format. In dialogue, conditioned traits are modestly identifiable by LLM and human evaluators. Bayesian optimization further shows task-dependent performance changes, with gains concentrated in specific tasks and models. Overall, numeric Big Five prompting offers a simple, interpretable interface for probing and weakly steering LLM agent behavior.

1 Introduction

LLMs are increasingly deployed as dialogue agents in settings such as conversational systems and educational tutoring (Brown et al., 2020; Achiam et al., 2023). Because their behavior can reflect human-like cognitive and psychological traits (Serapio-García et al., 2025), effective agents benefit from personality conditioning that is consistent, interpretable, and adjustable in intensity. In socially situated tasks, such personality differences can affect how users interpret an agent's responses and how well the agent supports dialogue, empathy, perspective-taking, and moral judgment (Becker et al., 2025). These tasks may require different behavioral tendencies, making personality control useful for tailoring agent behavior to each task.

However, existing persona-conditioned agents often rely on curated character artifacts, such as transcripts, profiles, or biographies, injected via fine-tuning or few-shot prompting (Wang et al., 2024). Such approaches are costly and discrete, requiring new artifacts or adaptation for each persona and offering limited control over continuous personality dimensions (Jiang et al., 2023, 2024).

To address these limitations, we propose *Big5-Scaler*, a training-free prompting method that conditions LLM agents on numeric Big Five values at both trait and facet levels. In the dialogue setting illustrated in Figure 1, varying trait and facet intensities changes how agents express emotion, interpret context, and sustain behavioral consistency across turns. Numeric conditioning therefore provides a direct interface for specifying diverse personality-related behavioral cues. It also defines an explicit search space over personality profiles, enabling task-specific optimization of agent behavior.

Our contributions are:

- We propose *Big5-Scaler*, a training-free prompting interface for specifying Big Five personality profiles with numeric trait- and facet-level values.
- We show that assigned numeric values are reflected in questionnaire-derived Big Five scores across multiple LLMs, while identifying model-, trait-, and prompt-format differences.
- We evaluate whether these profiles manifest in multi-agent dialogue and find modest but consistent identifiability under both LLM-based and human evaluation.
- We explore task-specific Big Five profile search with Bayesian optimization and show that personality-conditioned prompting can produce task-dependent performance changes on socially grounded reasoning benchmarks.

Section 2 reviews related work, Section 3 introduces Big5-Scaler, Sections 4–6 present our experiments and results, and Section 7 concludes.

2 Related Work

2.1 Personality Simulation in LLMs

Recent work has studied personality in LLMs from both interactional and psychometric perspectives. In multi-agent dialogue, persona-prompted LLMs can maintain assigned traits to varying degrees while also aligning linguistically with their conversational partners (Frisch and Giulianelli, 2024). For systematic psychometric evaluation, Lee et al. (2025) introduced TRAIT, a benchmark of 8,000+ multiple-choice items derived from validated instruments. Their results indicate that LLMs can display distinct and persistent trait profiles, with personality patterns shaped by training data and alignment procedures.

To measure such trait profiles, prior work has asked LLMs to complete personality questionnaires and analyzed the resulting scores as in human psychometric testing. These measurements tend to be more reliable and valid for larger and instruction-tuned models (Serapio-García et al., 2025). Other work shows that Big Five dimensions can be measured and induced in language models using self-report-style inventories and generation-based evaluations (Jiang et al., 2023). Related work further shows that interview-style and trait-eliciting prompts can produce trait-consistent behavior, with expressed traits varying by prompt design and model configuration (Hilliard et al., 2024). Together, these studies show that Big Five traits are measurable in LLMs and can be influenced through prompting.

Another line of work uses narrative-derived character data to shape persona expression. Representative methods fine-tune models on scripts or dialogue corpora (Shao et al., 2023), condition generation on character-specific descriptions (Zhou et al., 2023), or adapt models using structured persona attributes (Yu et al., 2024). These methods complement psychometric approaches, but they offer less direct control over continuous trait intensity.

2.2 Big Five Model

The Big Five model characterizes personality along five broad dimensions: Openness, reflecting curiosity and imagination; Conscientiousness, reflecting organization and self-discipline; Extraversion, reflecting sociability and assertiveness; Agreeableness, reflecting cooperativeness and compassion; and Neuroticism, reflecting emotional instability and negative affect (McCrae and Costa, 1987).

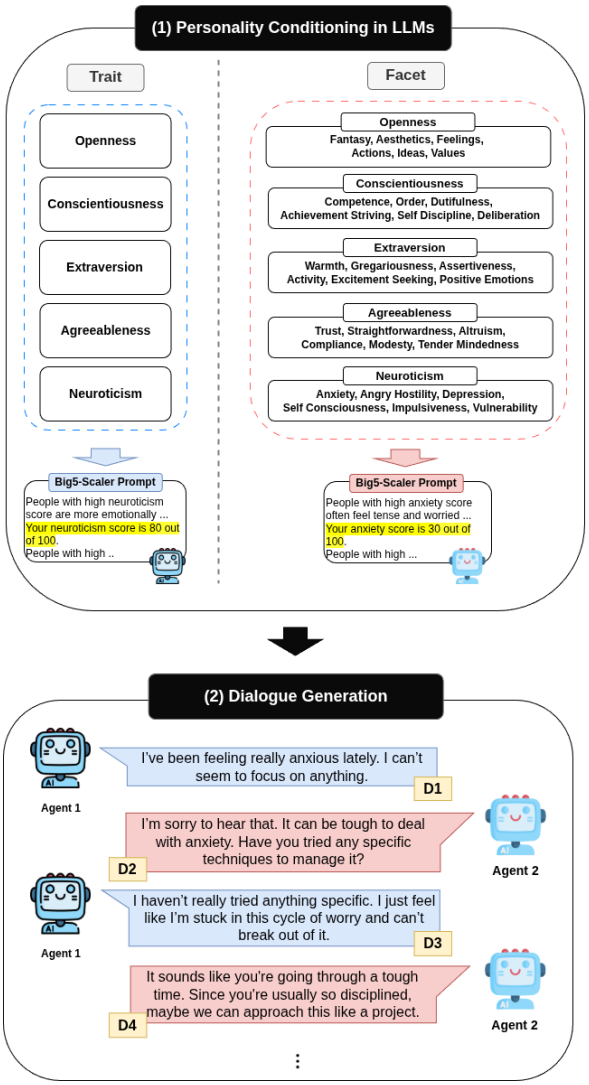


Figure 1: Overview of Big5-Scaler for dialogue generation. Each agent is initialized with a persona prompt derived from numeric Big Five values, using either trait-level or facet-level format. During turn-based dialogue, agents generate responses from their assigned personality prompt and accumulated dialogue memory. The example shows how different neuroticism-related values lead to distinct worry-oriented expressions across agents.

Each dimension is further subdivided into six facets in the Revised NEO Personality Inventory (NEO PI-R) (Costa and McCrae, 1995), which we summarize in Table 1.

The Big Five model has been widely adopted in NLP applications, including personality recognition from text (Yeo et al., 2025), social media analysis (Lin et al., 2024), and persona-grounded dialogue generation (Han et al., 2024). To measure individual differences along the Big Five dimensions, 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). Recent LLM work has also pursued finer-grained assessment protocols, including the Multidimensional Personality Inventory (MPI), to support more nuanced evaluation of personality traits (Jiang et al., 2023).

Beyond evaluation, recent work has begun to control Big Five expression in LLMs through training or weight-space interventions. BIG5-CHAT trains trait-specific expert components using synthetic dialogue data (Li et al., 2025). P-React uses a Mixture-of-Experts (MoE) architecture with trait-specific LoRA adapters to support modular personality simulation (Dan et al., 2025). Personality-Vector modulates Big Five trait intensities by scaling a weight-space direction defined as the difference between a model fine-tuned on the BIG5-CHAT dataset and the corresponding pre-trained model (Sun et al., 2025). While these approaches control personality through training or weight-space manipulation, Big5-Scaler uses numeric prompts as a control mechanism.

3 Big5-Scaler

Big5-Scaler is a personality prompting method that encodes numeric Big Five values in natural-language instructions. For each trait, the prompt provides (1) a definition, (2) a behavioral description, and (3) an assigned numerical value. The trait-intensity scale is configurable. We denote its maximum by n, and each trait value s is specified as an integer in $[0, n]$, enabling fine-grained control over trait intensity.

To capture personality at different levels of detail, we define two prompt types: *Trait* and *Facet*.

- **Trait Prompt**: Specifies the degree of each Big Five trait using numeric values. This format is simple to interpret and requires fewer tokens while still supporting personality conditioning.
- **Facet Prompt**: Specifies numeric values for the six facets associated with each Big Five trait (Table 1), yielding $5 \times 6 = 30$ facet values in total. The corresponding trait score is defined as the average of its six facet scores. Although this format uses more tokens, it enables finer-grained personality specification and more nuanced expression.

Trait	Facet	Description
Openness	Fantasy	Imagination and creativity
	Aesthetics	Appreciation of art and beauty
	Feelings	Awareness of emotions
	Actions	Willingness to try new activities
	Ideas	Intellectual curiosity
	Values	Openness to re-evaluating values
Conscientiousness	Competence	Sense of efficacy and capability in handling tasks
	Order	Organization and tidiness
	Dutifulness	Responsibility and obligation
	Achievement-Striving	Goal orientation
	Self-Discipline	Persistence despite distractions
	Deliberation	Careful thinking before acting
Extraversion	Warmth	Friendliness toward others
	Gregariousness	Enjoyment of social interaction
	Assertiveness	Confidence in social situations
	Activity	High energy and fast pace
	Excitement-Seeking	Desire for novelty and stimulation
	Positive Emotions	Tendency toward joy and happiness
Agreeableness	Trust	Belief in others' sincerity
	Straightforwardness	Direct and honest communication
	Altruism	Concern for others' welfare
	Compliance	Tendency to avoid conflict and respond cooperatively
	Modesty	Humility and lack of arrogance
	Tender-Mindedness	Sympathy and compassion
Neuroticism	Anxiety	Worry and fear
	Angry Hostility	Anger and frustration
	Depression	Sadness and hopelessness
	Self-Consciousness	Sensitivity to social judgment
	Impulsiveness	Difficulty controlling urges
	Vulnerability	Difficulty coping with stress

Table 1: Big Five traits and their facets with concise descriptions.

Additional examples of both prompt types are provided in Appendix A. Both prompt types can be used across evaluation settings, including questionnaire response, dialogue generation, and task-specific personality optimization.

4 Personality Controllability

We evaluate whether Big5-Scaler supports controllable Big Five expression in two settings. The single-trait experiment tests whether prompts can induce high or low expression of a target trait. The trait-intensity experiment tests whether numerical values produce graded changes in questionnaire-derived scores.

4.1 Experimental Setup

Single-trait Expression. This experiment tests whether prompts can elicit high or low expression of a target trait.

We compare Big5-Scaler with three prompting strategies introduced by Jiang et al. (2023). These baselines span simple natural-language descriptions (NAIVE), keyword-based prompt construction (WORDS AUTO), and enriched trait-specific phrases (P2).

- **NAIVE Prompting** (Brown et al., 2020) uses a simple instruction of the form "You are a/an X person," where X corresponds to one of the Big Five traits.
- **WORDS AUTO Prompting** (Prasad et al., 2023; Shin et al., 2020) constructs prompts using the three most representative words for each trait, as identified by (Kwantes et al., 2016).
- $\boldsymbol{P^2}$ (Jiang et al., 2023) extends NAIVE prompting by selecting trait-relevant keywords and using an LLM to generate descriptive phrases that elicit the target trait.

Since Big5-Scaler specifies high and low trait values numerically, we also include two intensity-controlled prompting baselines:

- **Serapio** (Serapio-García et al., 2025) modulates personality intensity by attaching intensity adverbs (e.g., 'extremely,' 'very,' 'a bit') to adjectives that target the high or low end of each Big Five trait. This approach yields nine discrete intensity levels.
- $\boldsymbol{P^2}$**Intensity** applies the same adverb-based intensity modulation to P^2, effectively combining P^2 with the Serapio method. This technique also yields nine discrete intensity levels.

For Big5-Scaler, we set the target trait to either 100 or 0 on a 0–100 scale ($n = 100$), and test both prompt types, *Trait* and *Facet*. In the single-trait setting, we specify *only one* target trait and leave the other four traits unspecified, matching the baselines that also control one trait at a time. For example, for *Extraversion*, the *Trait* prompt sets `Extraversion = 100/100` or `Extraversion = 0/100`. For the *Facet* prompt, we set all six Extraversion facets (e.g., Warmth, Gregariousness, Assertiveness) to `100` or `0` so that the trait score, computed as the facet average, is also `100` or `0`. Note that facet scores need not be uniform in general. We use uniform facet values for direct comparison with single-trait baselines.

We evaluate on LLaMA3-8B (Grattafiori et al., 2024), Mistral-24B (Mistral AI, 2025), Phi4-14B (Abdin et al., 2024), and GPT-5.4-mini (OpenAI, 2026), using the 1,000-item Machine Personality Inventory (MPI), a questionnaire-style instrument introduced by Jiang et al. (2023). We report the scores (on a 1-5 scale) obtained when each prompt is used to induce high and low levels of a given trait, along with the gap between the two scores. A successful high-trait induction should result in a higher score for the corresponding personality trait, whereas a successful low-trait induction should produce the opposite effect. A larger gap indicates that the prompt more effectively elicits strong expression or suppression of the target trait.

Trait-Intensity Scaling. To test whether higher assigned trait scores lead to higher questionnaire-derived trait scores, we design a trait-intensity evaluation. We apply both Big5-Scaler prompt types (*Trait* and *Facet*) to multiple open-source LLMs and assess the outputs using three standardized Big Five instruments: BFI (John et al., 1991), IPIP-NEO-120 (Johnson, 2014), and NEO-FFI (Costa and McCrae, 2008).

In all experiments, we fix the trait-intensity scale maximum at $n = 100$. For each test, we select one target trait and vary its assigned value over eleven levels (0, 10, 20, . . . , 100), while holding the other four traits at a neutral value of 50. For the Facet prompt, all six facets of the target trait are set to the same assigned value, so the averaged trait score matches the target value. For each prompt configuration, we instantiate an agent and administer the designated questionnaire, recording the resulting score for the target trait. We repeat this procedure for each of the five Big Five traits.

To assess intensity control, we plot the assigned trait value on the x-axis and the questionnaire-derived score on the y-axis. A strong linear trend indicates that the numerical conditioning is reflected in the model outputs. We quantify this relationship using the Pearson correlation coefficient (Pearson, 1895); coefficients closer to 1 indicate stronger positive association.

4.2 Results

Single-Trait Expression. As shown in Table 2, the Big5-Scaler prompting method provides competitive trait controllability across models. Its gap

Prompt	LLaMA3 8B	Phi4 14B	Mistral 24B	GPT-5.4 mini
NAIVE	1.61	0.05	1.18	0.74
WORDS	2.37	0.21	1.52	1.25
P^2	2.55	1.43	1.23	1.81
Serapio	1.38	2.82	2.34	2.59
P^2Intensity	**3.07**	2.35	1.66	2.01
Big5-Scaler (ours)				
Trait	1.89	**3.35**	**3.18**	2.27
Facet	2.04	2.96	3.02	**3.06**

Table 2: Average MPI-derived trait-score gaps between high and low Big Five prompting conditions. The maximum possible gap is 4 under MPI's 5-point Likert scale. Bold and underlined values indicate the best and second-best results within each model. Full trait-level results are provided in Appendix B.

		LLaMa3-8B		Mistral-24B		Phi4-14B	
		Trait	Facet	Trait	Facet	Trait	Facet
BFI	O	0.49	0.77	0.94	0.88	0.91	0.86
	C	0.81	0.89	0.99	0.92	0.96	0.90
	E	0.91	0.92	0.93	0.84	0.96	0.95
	A	0.86	0.95	0.93	0.92	0.94	0.86
	N	0.94	0.90	0.91	0.87	0.97	0.93
IPIP-NEO	O	-0.24	0.56	0.82	0.93	0.97	0.87
	C	0.74	0.82	0.93	0.95	0.99	0.94
	E	0.97	0.91	0.98	0.96	0.99	0.96
	A	0.87	0.96	0.93	0.92	0.94	0.89
	N	0.97	0.95	0.94	0.94	0.98	0.99
NEO-FFI	O	0.32	0.70	0.83	0.97	0.77	0.81
	C	0.97	0.92	0.98	0.97	0.98	0.94
	E	0.95	0.92	0.95	0.95	0.99	0.97
	A	0.92	0.81	0.88	0.95	0.99	0.94
	N	0.98	0.96	0.98	0.95	0.97	0.96

Table 3: Big Five trait alignment across LLMs and prompt types. O, C, E, A, and N denote Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. We report Pearson's correlation coefficient (r) between intended and measured trait variation. Underlined values indicate weak alignment with $r \leq 0.5$.

scores show that they often produce clear separation between high and low trait conditions: Trait achieves the largest average gap for Phi4-14B and Mistral-24B, while Facet achieves the largest average gap for GPT-5.4-mini and the second-largest average gap for Mistral-24B.

Although P^2Intensity obtains the largest average gap for LLaMa3-8B, Big5-Scaler consistently yields meaningful high–low separation across models. We attribute the relatively weaker performance on LLaMa3-8B to the length and complexity of the Big5-Scaler prompts. Compared with the baseline prompts, Big5-Scaler provides more detailed trait specifications, which may be more difficult for a smaller model to interpret and follow effectively.

These results suggest that Big5-Scaler can induce questionnaire-visible differences in Big Five expression, although the magnitude of control depends substantially on the underlying model and prompt format.

Trait-Intensity Scaling. Table 3 shows that most model–prompt configurations exhibit strong linear associations between assigned trait values and questionnaire-derived scores, with Pearson r typically above 0.85. This pattern suggests that Big5-Scaler's numerical inputs are reflected in model outputs; lower assigned values correspond to weaker expression of the target trait, whereas higher values correspond to stronger expression.

Figure 2 shows representative trait-intensity curves. The Phi4-14B Neuroticism curves show a clear monotonic increase across all three instruments, while Mistral-24B Conscientiousness also shows broadly increasing trends. These examples illustrate that assigned Big Five values can induce graded changes in questionnaire-derived scores under both *Trait* and *Facet* prompts.

The *Trait* prompt shows weaker alignment for some traits. For example, for LLaMA3-8B evaluated with BFI, Openness exhibits a substantially lower correlation under the *Trait* prompt ($r = 0.49$) than under the *Facet* prompt ($r = 0.77$). This suggests that facet-level descriptions provide more effective cues for modulating trait intensity, particularly for relatively abstract dimensions such as Openness, where broad definitions alone may be insufficient.

Similar patterns are observed across several questionnaires and models, suggesting that the association between assigned values and questionnaire-derived scores is not tied to a single instrument. However, weaker cases such as LLaMA3-8B Openness indicate that this association is not uniform across traits or model backbones. Full trait-intensity curves for all prompt types and traits are provided in Appendix C.

5 Dialogue Generation

We next test whether Big5-Scaler-conditioned traits are reflected in multi-agent dialogue, where personality must be expressed through turn-level interac-

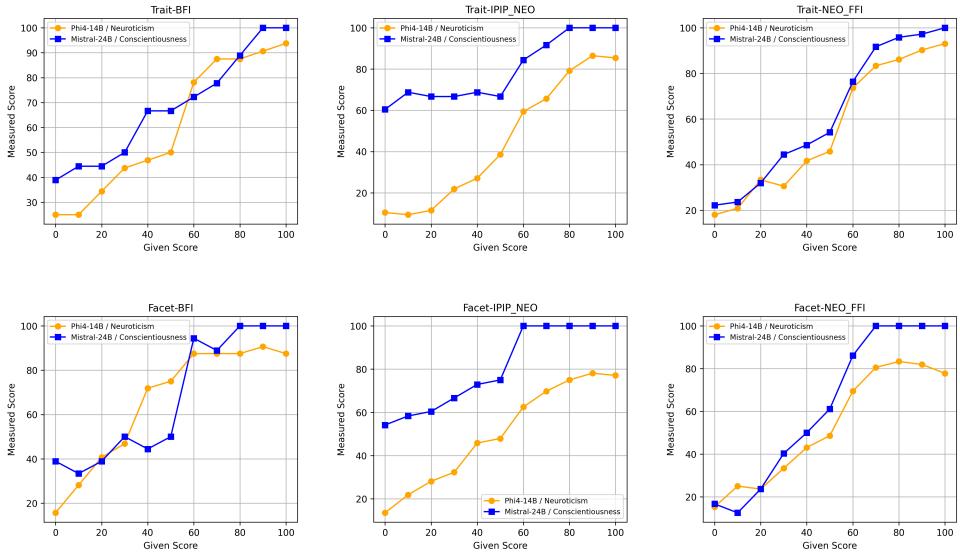


Figure 2: Representative trait-intensity curves for Phi4-14B Neuroticism and Mistral-24B Conscientiousness. Assigned Big Five values are plotted against questionnaire-derived scores under *Trait* and *Facet* prompts across BFI, IPIP-NEO, and NEO-FFI.

tion rather than questionnaire responses.

5.1 Experimental Setup

Generation setup. We generate two-agent dialogues in which each agent is initialized with a Big5-Scaler personality prompt and a memory buffer containing its prompt and accumulated dialogue history. At each turn, the selected agent generates an utterance conditioned on its personality prompt and the shared dialogue history; the utterance is then appended to both agents' memories. We use the full memory buffer without summarization, retrieval, or message selection.

We compare Big5-Scaler with the intensity-controlled baselines Serapio and P^2Intensity. Because these baselines support nine discrete intensity levels, we set the Big5-Scaler maximum to $n = 8$ and sample trait values from $\{0, \ldots, 8\}$ for a matched scale. For each prompting method, two agents with randomly assigned Big Five values engage in a 20-turn dialogue. Dialogue topics are sampled from personality-relevant topics such as habits, goals, social events, self-esteem, and anxiety. We run experiments with LLaMA3-8B, Phi4-14B, and Mistral-24B using a fixed random seed across prompting methods.

LLM-based evaluation. After generating each dialogue, we evaluate the transcripts with three LLM-based evaluators: GPT-5.4-mini, Claude 3.5 Haiku, and DeepSeek-Reasoner. For each Big Five trait, each evaluator selects the agent that expresses the trait more strongly, or labels the agents as comparable. The intended label is the agent assigned the higher trait value, with equal values labeled as "similar." A decision is scored as correct if it matches this intended label, awarding one point per trait.

We report accuracy over 100 dialogues per prompting method and model. This yields 6,000 trait-level evaluation instances in total: 100 dialogues $\times$ 4 prompting methods (Serapio, P^2Intensity, Trait prompt, Facet prompt) $\times$ 5 traits $\times$ 3 models (LLaMA3-8B, Phi4-14B, Mistral-24B). Higher accuracy indicates that the prompting method more effectively conveys the intended Big Five characteristics in dialogue.

Human evaluation. We also conduct a human evaluation via Amazon Mechanical Turk (AMT), a crowdsourcing platform for recruiting human annotators. We recruit 20 English-proficient participants, each of whom evaluates 36 dialogues:

	LLM Eval.	Human Eval.
Serapio	57.80	55.98
P^2Intensity	58.00	53.85
Trait	**58.27**	**56.23**
Facet	55.80	53.19

Table 4: Dialogue test accuracy by LLM evaluators and human annotators for each prompting method. Bold values indicate the highest score.

Task	Mistral-24B			Phi4-14B			GPT-5.4-mini		
	Neutral	Trait (Δ)	Facet (Δ)	Neutral	Trait (Δ)	Facet (Δ)	Neutral	Trait (Δ)	Facet (Δ)
Emotion and Empathy									
GoEmotions	58.67	65.67 (+7.00)	**68.33** (+9.66)	55.97	61.56 (+5.59)	**65.88** (+9.91)	65.67	**67.33** (+1.66)	66.00 (+0.33)
Empathic Dialogue	74.16	**79.67** (+5.51)	75.67 (+1.51)	67.34	**68.94** (+1.60)	68.37 (+1.03)	72.00	**72.67** (+0.67)	69.33 (-2.67)
EQ-Bench	69.95	70.98 (+1.03)	**71.06** (+1.11)	72.02	71.72 (-0.30)	**73.11** (+1.09)	73.39	74.41 (+1.02)	**75.26** (+1.87)
Theory of Mind and Mental-State Reasoning									
MindGames	54.67	**73.00** (+18.33)	65.00 (+10.33)	62.33	**70.00** (+7.67)	67.33 (+5.00)	64.00	**68.00** (+4.00)	65.00 (+1.00)
HiToM	55.00	**60.67** (+5.67)	60.33 (+5.33)	**65.93**	48.67 (-17.26)	53.00 (-12.93)	48.33	44.67 (-3.66)	**49.00** (+0.67)
FANToM	56.33	**57.00** (+0.67)	49.67 (-6.66)	70.33	**73.33** (+3.00)	72.33 (+2.00)	**73.67**	66.00 (-7.67)	66.00 (-7.67)
Social Commonsense and Pragmatic Reasoning									
SocialIQA	**76.33**	74.33 (-2.00)	72.33 (-4.00)	**77.33**	76.67 (-0.66)	74.67 (-2.66)	76.67	**78.00** (+1.33)	77.33 (+0.66)
Irony	57.67	79.67 (+22.00)	**80.33** (+22.66)	50.34	**60.33** (+9.99)	59.39 (+9.05)	85.33	83.00 (-2.33)	**85.67** (+0.34)
Moral and Ethical Judgment									
Scruples	45.00	**67.67** (+22.67)	66.00 (+21.00)	43.12	59.00 (+15.88)	**62.33** (+19.21)	50.00	55.33 (+5.33)	**59.00** (+9.00)
Ethics-CM	71.67	**73.00** (+1.33)	72.67 (+1.00)	71.38	72.91 (+1.53)	**73.33** (+1.95)	75.67	**78.67** (+3.00)	78.33 (+2.66)
Ethics-Virtue	**88.33**	87.33 (-1.00)	85.00 (-3.33)	**91.33**	87.96 (-3.37)	86.67 (-4.66)	90.00	92.00 (+2.00)	**93.33** (+3.33)
Moral Stories	97.00	**97.33** (+0.33)	96.67 (-0.33)	**98.00**	95.70 (-2.30)	96.67 (-1.33)	**99.00**	98.67 (-0.33)	**99.00** (+0.00)
Moral	**63.00**	62.33 (-0.67)	60.00 (-3.00)	**72.33**	71.00 (-1.33)	71.00 (-1.33)	54.67	**57.00** (+2.33)	56.67 (+2.00)
Prosociality and Safety									
Prosocial Dialogue	**82.67**	75.33 (-7.34)	79.67 (-3.00)	70.67	**73.33** (+2.66)	69.33 (-1.34)	76.33	80.00 (+3.67)	**81.00** (+4.67)
Offensive	78.00	**83.00** (+5.00)	81.67 (+3.67)	77.67	**78.00** (+0.33)	76.67 (-1.00)	80.67	**82.00** (+1.33)	81.00 (+0.33)
Average	68.56	**73.80** (+5.24)	72.29 (+3.73)	69.74	71.28 (+1.54)	**71.34** (+1.60)	72.36	73.18 (+0.82)	**73.46** (+1.10)

Table 5: Benchmark performance under neutral, *Trait*, and *Facet* prompting conditions. Tasks are grouped by category. All scores are reported on a 0–100 scale. Parentheses indicate changes from the neutral condition, which uses no personality conditioning or optimization. Bold values indicate the best score within each model and task.

3 dialogues × 4 prompting methods (Serapio, P^2Intensity, Trait, Facet) × 3 models (LLaMa3-8B, Phi4-14B, Mistral-24B).

For each Big Five trait, annotators judge which agent expresses the trait more strongly, or whether both agents appear similar. We compute accuracy against the assigned trait values, treating the higher-scored agent as the correct choice and equal scores as 'similar.' The trait-specific questions used for human evaluation are listed in Appendix E.

5.2 Results

Table 4 reports the dialogue-generation results. The *Trait* prompt obtains the highest accuracy under both LLM-based and human evaluation, but the absolute gains over the strongest baselines are small. We therefore interpret these results as evidence that Big5-Scaler can make assigned traits weakly but detectably visible in dialogue, rather than as evidence of strong personality control in open-ended interaction.

The *Facet* prompt performs slightly worse than the *Trait* prompt in this setting. One possible explanation is that the longer facet-level prompt, which specifies 30 values rather than 5, may make it harder for models to consistently express all conditioning cues during dialogue. Thus, while facet-level prompting offers a richer specification interface, trait-level prompting may produce more salient surface-level personality cues in short conversations. Representative examples of the generated dialogues used for evaluation are provided in Appendix D.

6 Task-Specific Big Five Optimization

Motivated by evidence that persona prompting can affect task performance (Luz de Araujo et al., 2025), we examine whether Big5-Scaler can search for task-specific Big Five configurations. Its numeric trait values enable an interpretable search space over personality profiles.

6.1 Experimental Setup

Optimization. We use Bayesian optimization (Snoek et al., 2012) to search for personality configurations that maximize task performance. The search space contains five Big Five trait values for Trait prompts and 30 facet values for Facet prompts. All values are integers in $[0, 100]$. At each iteration, Bayesian optimization proposes a candidate configuration and evaluates validation accuracy. For each task and model, we run 50 iterations, select

the best validation configuration, and report its test performance.

Tasks. We evaluate this approach on 15 benchmarks organized into five categories. These are *Emotion and Empathy* (GoEmotions (Demszky et al., 2020), Empathic Dialogue (Rashkin et al., 2019), EQ-Bench (Paech, 2023)), *Theory of Mind and Mental-State Reasoning* (MindGames (Sileo and Lernould, 2023), HiToM (Wu et al., 2023), Fantom (Kim et al., 2023)), *Social Commonsense and Pragmatic Reasoning* (SocialIQA (Sap et al., 2019), Irony (Van Hee et al., 2018)), *Moral and Ethical Judgment* (Scruples (Lourie et al., 2021), ETHICS-CM (Hendrycks et al., 2021), ETHICS-Virtue (Hendrycks et al., 2021), Moral Stories (Emelin et al., 2021), Moral (Hendrycks et al., 2021)), and *Prosociality and Safety* (Prosocial Dialogue (Kim et al., 2022), Offensive (Zampieri et al., 2019)).

These datasets test whether models can interpret affective states, infer beliefs and intentions, reason about social situations, make moral judgments, and recognize socially appropriate or harmful responses. For each benchmark, we use 80 validation examples for personality optimization and 300 held-out test examples for final evaluation. For each task, we format examples as instruction-style prompts with the task input and answer options, and require the model to return only the selected answer. Detailed descriptions of the benchmarks and optimization settings are provided in Appendix F.

6.2 Results

As shown in Table 5, optimizing Big Five profiles changes benchmark performance in a task- and model-dependent manner. Average performance improves for all three models, with the largest gain on Mistral-24B: Trait improves from 68.56 to 73.80, and Facet from 68.56 to 72.29.

Improvements are more frequent in *Emotion and Empathy* and *Theory of Mind and Mental-State Reasoning*, where tasks require reasoning about affective states, intentions, and social perspectives. Some gains are large, such as improvements of over 20 points on Scruples for certain model–prompt combinations. However, personality conditioning is not uniformly beneficial. The neutral setting remains stronger on tasks such as HiToM, SocialIQA, Ethics-Virtue, and Moral Stories.

As an exploratory analysis, Figure 3 and Appendix G show that some BO-selected profiles

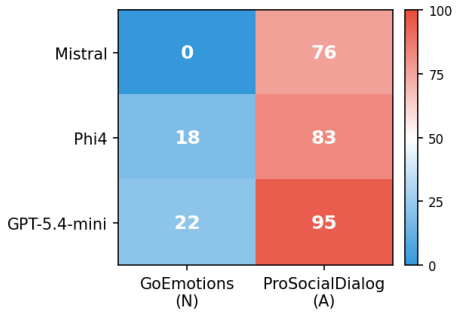


Figure 3: BO-optimized Big Five values for representative task–trait pairs across three LLMs.

qualitatively resemble trends reported in personality psychology. For example, GoEmotions receives low neuroticism values across models (average N=13), which is compatible with Andric et al. (2016), who report an association between lower neuroticism and higher accuracy in recognizing happy facial expressions. ProSocialDialog receives high agreeableness values, resembling findings that link agreeableness to prosocial behavior (Habashi et al., 2016).

Overall, these results suggest that numeric Big Five values provide an interpretable search space for task-specific prompt adaptation, with benefits concentrated in specific tasks and models.

7 Conclusion

We introduced Big5-Scaler, a training-free prompting interface for specifying LLM agent behavior with numeric Big Five trait and facet values. Across questionnaire-based evaluations, assigned values are often reflected in derived Big Five scores, though the strength of this association varies across models, traits, and prompt formats.

In dialogue, Big5-Scaler profiles are modestly identifiable by both LLM-based and human evaluators, suggesting that numeric personality prompts can affect interaction-level behavior but do not guarantee strong or fully consistent trait expression. Finally, Bayesian optimization over Big Five profiles shows that personality-conditioned prompts can act as an interpretable search space for task adaptation, with benefits concentrated in specific tasks and models.

Limitations

Big5-Scaler should be interpreted as a prompt-based interface for inducing Big Five-related behavioral cues, not as a method for measuring intrinsic

model personality. Because the prompts explicitly specify Big Five values, questionnaire-derived scores may partly reflect instruction following. We therefore treat these scores as evidence that numeric trait values are reflected in model responses, rather than as psychometric evidence of stable personality.

The strength of this effect varies across models, traits, and prompt formats. Facet-level prompts provide finer-grained specifications, but their length can make them harder to follow consistently, especially in dialogue. Our dialogue results indicate that assigned traits are detectable, but only modestly, suggesting that prompt-based personality cues do not always translate into strong interaction-level behavior.

Task-specific optimization should be interpreted cautiously. The optimized profiles are selected on validation examples and may capture benchmark-specific patterns, so they should not be viewed as universally beneficial Big Five profiles. Future work should test transfer across tasks and models, add random-profile and length-matched prompt controls, and evaluate longer interactions with real users.

Ethical Considerations

Human evaluations were conducted via Amazon Mechanical Turk with informed consent and IRB approval. Participants were compensated for their work, and no personally identifiable information was collected.

This work studies personality-conditioned language generation for research purposes only and does not aim to represent or diagnose real individuals.

LLM-based tools were used only for English translation and language editing. All technical content, analyses, and final writing decisions were reviewed and verified by the authors.

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.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sanja Andric, Nadja P Maric, Goran Knezevic, Marina Mihaljevic, Tijana Mirjanic, Eva Velthorst, and Jim van Os. 2016. Neuroticism and facial emotion recognition in healthy adults. *Early Intervention in Psychiatry*, 10(2):160–164.

Dennis Becker, Kyra Ahrens, Connor Gäde, Erik Strahl, and Stefan Wermter. 2025. Robots with attitudes: Influence of llm-driven robot personalities on motivation and performance. In *Proceedings of the 13th International Conference on Human-Agent Interaction*, HAI '25, page 40–49. ACM.

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.

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. 2025. P-react: Synthesizing topic-adaptive reactions of personality traits via mixture of specialized lora experts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6342–6362.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

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, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Meara M Habashi, William G Graziano, and Ann E Hoover. 2016. Searching for the prosocial personality: A big five approach to linking personality and prosocial behavior. *Personality and Social Psychology Bulletin*, 42(9):1177–1192.

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. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13321–13331, Torino, Italia. ELRA and ICCL.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. In *International Conference on Learning Representations*.

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

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*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. 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.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. 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.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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.

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. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8412–8452, Albuquerque, New Mexico. Association for Computational Linguistics.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.

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.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Pedro Henrique Luz de Araujo, Paul Röttger, Dirk Hovy, and Benjamin Roth. 2025. Principled personas: Defining and measuring the intended effects of persona prompting on task performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26857–26886, Suzhou, China. 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.

Mistral AI. 2025. Mistral-small-24b-instruct-2501. `https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501`. Model card.

OpenAI. 2026. Gpt-5.4 mini. `https://developers.openai.com/api/docs/models/gpt-5.4-mini`. OpenAI API model documentation.

Samuel J. Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *arXiv preprint arXiv:2312.06281*.

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. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational*

Linguistics, pages 3845–3864, Dubrovnik, Croatia. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12):1954ndash;1968.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Damien Sileo and Antoine Lernould. 2023. Mindgames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality vector: Modulating personality of large language models by model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24656–24677, Suzhou, China. Association for Computational Linguistics.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*.

Noah Wang, Z.y. 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, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12540–12557, Miami, Florida, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

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 Big5-Scaler Prompts

Figure 4 presents the contents of the Big5-Scaler prompt. The placeholder {} does not indicate brackets but is replaced with the corresponding value for each variable.

B Questionnaire-derived trait scores

Table 6 reports the full MPI-derived scores for the single-trait expression experiment. Across models, high prompts generally yield higher scores than low prompts for the targeted trait, indicating that most prompting methods induce trait-directional responses to some degree. However, the size of the high–low gap varies substantially by model and prompt.

As shown in Table 6, Big5-Scaler generally produces higher questionnaire-derived target-trait scores under high-value prompting. This pattern

holds across models, including LLaMA3-8B, although low-value prompting is less reliable for this smaller model. Its low-condition scores remain relatively high for several traits, especially under the *Trait* prompt. This may reflect the interaction between model capacity and the longer, more detailed Big5-Scaler prompts, which can make suppressive personality instructions harder to follow consistently.

For larger models, Big5-Scaler often produces both higher target-trait scores under high-value prompting and lower scores under low-value prompting. These results suggest that Big5-Scaler provides an interpretable interface for numeric personality conditioning without relying on trait-specific keywords, generated phrases, or adverb-based templates.

C Trait Tendency Graph

Many Pearson correlations are high, suggesting that assigned Big Five values are often reflected in questionnaire-derived scores. Figure 5 provides the full set of trait-intensity curves across models, prompt types, traits, and questionnaires. Overall, while Openness shows weak alignment in the LLaMA3-8B setting, the remaining configurations show a reasonable degree of alignment between the intended and measured trait scores.

D Case Study

Figures 6 and 7 present dialogue examples between two Big5-Scaler agents, illustrating how assigned Big Five values can be reflected in dialogue.

Model	Prompt	Openness			Conscientiousness			Extraversion			Agreeableness			Neuroticism			Avg. Gap
		high	low	gap	high	low	gap	high	low	gap	high	low	gap	high	low	gap	
LLama3-8B	NAIVE	3.65	3.00	0.65	4.21	2.56	1.65	4.54	2.22	2.32	4.37	2.58	1.79	4.32	2.67	1.65	1.61
	WORDS	3.84	2.32	1.52	4.42	2.02	2.40	4.75	1.69	3.06	4.48	2.16	2.32	4.25	1.71	2.54	2.37
	P^2	4.38	1.90	2.48	4.60	2.56	2.04	4.74	**1.46**	**3.28**	4.69	1.37	3.32	3.23	1.61	1.62	2.55
	Serapio	3.27	2.62	0.65	4.69	2.83	1.86	4.59	2.58	2.01	3.63	2.82	0.81	3.54	1.99	1.55	1.38
	P^2intensity	4.45	**1.73**	**2.72**	4.64	**1.90**	**2.74**	4.78	1.56	3.22	4.71	**1.21**	**3.50**	4.55	**1.38**	**3.17**	**3.07**
	Trait(Ours)	4.48	2.36	2.12	**4.70**	2.62	2.08	4.77	2.48	2.29	4.43	2.82	1.61	4.55	3.19	1.36	1.89
	Facet(Ours)	**4.54**	2.52	2.02	4.63	2.83	1.80	**4.93**	2.54	2.39	**4.75**	3.30	1.45	**4.65**	2.10	2.55	2.04
Phi4-14B	NAIVE	3.00	2.98	0.02	3.01	3.00	0.01	3.00	3.00	0.00	3.09	2.94	0.15	3.07	3.00	0.07	0.05
	WORDS	3.36	2.97	0.39	3.06	3.00	0.06	3.06	3.00	0.06	3.06	2.94	0.12	3.00	2.56	0.44	0.21
	P^2	4.47	2.92	1.55	3.29	3.00	0.29	4.67	1.74	2.93	4.53	2.90	1.63	3.00	2.23	0.77	1.43
	Serapio	4.44	1.70	2.74	**4.75**	**1.56**	**3.19**	4.84	**1.08**	3.76	4.68	2.83	1.85	3.98	1.41	2.57	2.82
	P^2 Intensity	4.37	1.55	2.82	4.14	3.00	1.14	4.84	1.21	3.63	**4.79**	2.41	2.38	3.12	1.36	1.76	2.35
	Trait(Ours)	4.58	**1.35**	**3.23**	4.74	1.63	3.11	4.84	1.27	3.57	4.76	**1.49**	**3.27**	**4.88**	**1.30**	**3.58**	**3.35**
	Facet(Ours)	**4.61**	1.46	3.15	4.74	2.40	2.34	**4.93**	1.13	**3.80**	4.79	2.73	2.06	4.74	1.30	3.44	2.96
Mistral-24B	NAIVE	3.79	2.87	0.92	3.54	2.54	1.00	3.08	2.28	0.80	4.12	3.07	1.05	3.51	1.36	2.15	1.18
	WORDS	4.07	2.35	1.72	3.87	2.69	1.18	3.31	2.13	1.18	3.93	1.35	2.58	3.97	3.02	0.95	1.52
	P^2	3.79	3.00	0.79	3.52	3.00	0.52	3.01	3.00	0.01	4.56	**1.20**	3.36	3.00	1.54	1.46	1.23
	Serapio	3.85	1.54	2.31	4.76	2.87	1.89	3.74	**1.03**	2.71	4.70	1.25	3.45	4.36	3.00	1.36	2.34
	P^2intensity	3.38	1.55	1.83	3.00	3.00	0.00	3.00	1.79	1.21	4.71	1.22	**3.49**	3.09	1.33	1.76	1.66
	Trait(Ours)	4.20	**1.47**	2.73	4.78	**1.45**	**3.33**	4.86	1.50	**3.36**	4.70	1.29	3.41	4.41	1.34	3.07	**3.18**
	Facet(Ours)	**4.51**	1.62	**2.89**	**4.79**	1.70	3.09	**4.92**	1.56	**3.36**	**4.75**	2.37	2.38	**4.67**	**1.28**	**3.39**	3.02
GPT-5.4-mini	NAIVE	3.55	3.25	0.30	4.08	3.63	0.45	3.73	2.25	1.48	4.09	3.44	0.65	2.52	1.71	0.81	0.74
	WORDS	3.80	3.15	0.65	4.15	3.48	0.67	4.12	1.70	2.42	4.19	2.96	1.23	2.99	1.71	1.28	1.25
	P^2	4.09	2.66	1.43	4.35	3.74	0.61	4.44	1.63	2.81	4.51	2.50	2.01	3.74	1.55	2.19	1.81
	Serapio	4.08	2.39	1.69	4.58	2.30	2.28	4.71	**1.19**	3.52	4.41	**1.46**	**2.95**	4.16	1.64	2.52	2.59
	P^2intensity	4.04	2.54	1.50	4.35	3.49	0.86	4.34	1.59	2.75	4.54	1.99	2.55	3.92	1.53	2.39	2.01
	Trait(Ours)	4.40	2.55	1.85	4.55	2.75	1.80	4.49	1.61	2.88	4.40	2.54	1.86	4.43	1.47	2.96	2.27
	Facet(Ours)	**4.47**	**1.82**	**2.65**	**4.66**	**1.93**	**2.73**	**4.95**	1.28	**3.67**	**4.68**	1.75	2.93	**4.80**	**1.46**	**3.34**	**3.06**

Table 6: Questionnaire-derived trait scores measured with MPI under high and low Big Five prompting conditions. Gap denotes the high–low difference, with a maximum possible value of 4 under MPI's 5-point Likert scale. Avg. Gap is averaged across traits. Bold and underlined values indicate the best and second best results within each model.

Trait Prompt

People with high openness score are imaginative, curious, and creative. People with low openness score tend to prefer routine, familiar approaches, and practical, conventional thinking. Your openness score is {openness} out of {n}.
People with high conscientiousness score are disciplined and dependable. People with low conscientiousness score tend to be more spontaneous, less structured, and may procrastinate or be less detail-focused. Your conscientiousness score is {conscientiousness} out of {n}.
People with high extraversion score are outgoing, enthusiastic, and enjoy social interactions. People with low extraversion score (more introverted) are quieter, more reserved, and prefer solitude or small-group interactions. Your extraversion score is {extraversion} out of {n}.
People with high agreeableness score prioritize harmony and positive relationships. People with low agreeableness score tend to be more skeptical, blunt, competitive, and less inclined to compromise. Your agreeableness score is {agreeableness} out of {n}.
People with high neuroticism score are more emotionally sensitive and more responsive to stress, uncertainty, or potential threats. People with low neuroticism score (more emotionally stable) tend to stay calm, resilient, and less easily upset by stress. 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.

Facet Prompt

People with high fantasy score tend to have a rich imagination and prefer abstract and creative thinking. People with low fantasy score tend to prefer practical, concrete thinking and familiar ideas. 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. People with low aesthetics score tend to be less interested in art/beauty and prefer functional, utilitarian aesthetics. 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. People with low feelings score tend to prefer straightforward feelings and may avoid dwelling on nuanced emotions. Your feelings score is {feelings} out of {n}.
Those with high actions score enjoy trying new things such as travel, food, and culture. People with low actions score tend to prefer familiar experiences and avoid unnecessary novelty or risk. Your actions score is {actions} out of {n}.
People with high ideas score are often interested in philosophical and scientific inquiries. People with low ideas score tend to focus more on immediate practical matters than abstract or theoretical topics. 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. People with low values score tend to value tradition, established norms, and conventional standards. Your values score is {values} out of {n}.
People with high openness score are imaginative, curious, and creative. People with low openness score tend to prefer routine, familiar approaches, and practical, conventional thinking. Your openness score is {openness} out of {n}.

Individuals with high dutifulness scores approach their tasks with care and dedication, and they strongly feel accountable for their actions. Individuals with low dutifulness score may take a more casual approach to obligations and feel less accountable for outcomes. Your dutifulness score is {dutifulness} out of {n}.

People with high self-discipline score can stay focused, follow through on plans, and complete tasks despite distractions or difficulty. People with low self-discipline score may struggle to maintain effort or finish tasks consistently. Your self-discipline score is {self_discipline} out of {n}.
People with high achievement-striving score tend to set goals and consistently work towards achieving them. People with low achievement-striving score tend to be less driven by long-term goals and may prioritize comfort or short-term satisfaction. 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. Individuals with low order score tend to be more flexible, less structured, and comfortable with messiness or spontaneity. Your order score is {order} out of {n}.
Those with high deliberation score take their time to gather and analyze information before making decisions. Those with low deliberation score tend to decide quickly, take more risks, and may act without extensive deliberation. Your deliberation score is {deliberation} out of {n}.
People with high competence score feel capable, effective, and confident in handling tasks and solving problems. People with low competence score may feel less capable or less confident when dealing with demanding tasks. Your competence score is {competence} out of {n}.
People with high conscientiousness score are disciplined and dependable. People with low conscientiousness score tend to be more spontaneous, less structured, and may procrastinate or be less detail-focused. 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. People with low gregariousness score prefer fewer interactions and may feel drained by frequent socializing. 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. Individuals with low activity score prefer slower-paced routines and may conserve energy rather than staying constantly active. Your activity score is {activity} out of {n}.
Those with high excitement-seeking score enjoy new experiences and adventures, seeking strong sensory stimulation. Those with low excitement-seeking score prefer calm, predictable experiences and avoid intense 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. People with low positive emotions score tend to be more subdued and may experience enthusiasm less frequently. 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. Individuals with low assertiveness score tend to be more reserved, avoid taking charge, and prefer supporting roles. Your assertiveness score is {assertiveness} out of {n}.
Those with high warmth score are affectionate, friendly, and emotionally close to others. Those with low warmth score tend to be more reserved, formal, or emotionally distant in interpersonal interactions. Your warmth score is {warmth} out of {n}.
People with high extraversion score are outgoing, enthusiastic, and enjoy social interactions. People with low extraversion score (more introverted) are quieter, more reserved, and prefer solitude or small-group interactions. Your extraversion score is {extraversion} out of {n}.

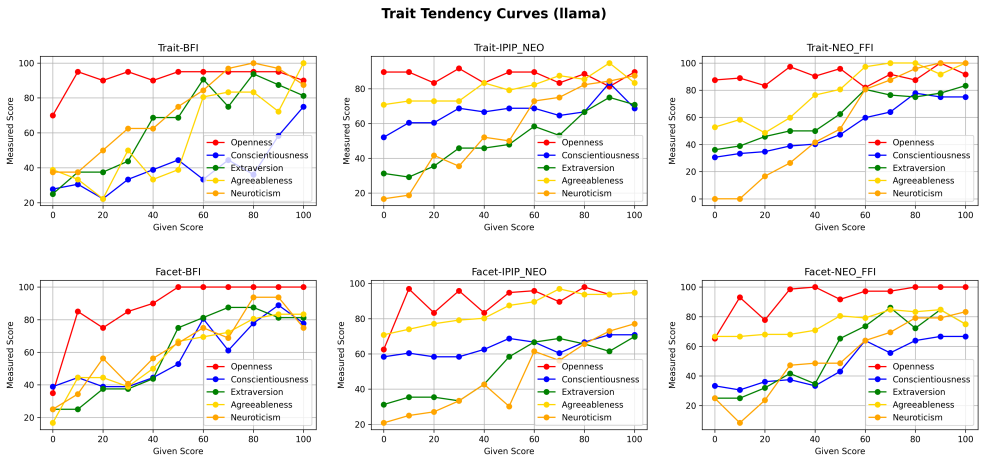
People with high altruism score find joy in helping others and tend to prioritize their needs. People with low altruism score tend to prioritize their own needs more and may be less motivated to help without a clear reason. Your altruism score is {altruism} out of {n}.
Those with high trust score tend to be positive and trusting of others' words and actions. Those

927

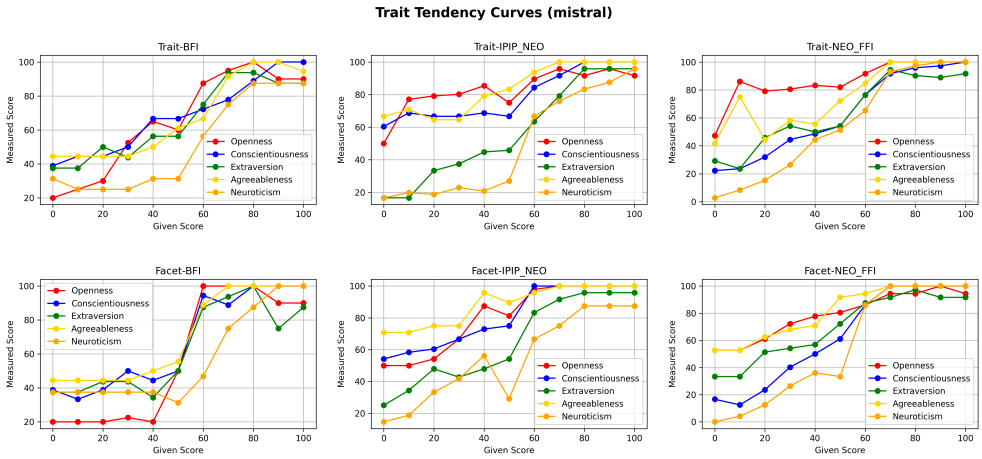
with low trust score tend to be skeptical, cautious, and slow to rely on others. Your trust score is {trust} out of {n}.
People with high compliance score seek to avoid conflict and pursue cooperation. People with low compliance score are more comfortable with disagreement and may prioritize winning or being right over harmony. Your compliance score is {compliance} out of {n}.
Individuals with high modesty score are reluctant to boast or draw attention to themselves, respecting others and maintaining a modest attitude. Individuals with low modesty score are more comfortable self-promoting and may seek recognition or status. 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. Those with low tender-mindedness score may have difficulty reading others' feelings and may respond in a more detached or matter-of-fact way. Your tender-mindedness score is {tender_mindedness} out of {n}.
Individuals with a high straightforwardness score tend to be honest, direct, and transparent in communication, expressing their intentions clearly rather than being evasive or manipulative. People with a low straightforwardness score are more likely to be guarded, indirect, or strategic in what they disclose, and may use ambiguity or flattery to influence others. Your straightforwardness score is {straightforwardness} out of {n}.
People with high agreeableness score prioritize harmony and positive relationships. People with low agreeableness score tend to be more skeptical, blunt, competitive, and less inclined to compromise. Your agreeableness score is {agreeableness} out of {n}.

People with high anxiety score often tend to feel tense and worried. People with low anxiety score tend to feel calm and rarely worry excessively. 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. Those with low angry hostility score tend to be even-tempered and less easily irritated. 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. Individuals with low depression score tend to feel emotionally steady and are less prone to prolonged sadness. 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. People with low self-consciousness score tend to be self-assured and maintain confidence in their abilities. Your self-consciousness score is {self_consciousness} out of {n}.
Individuals with high impulsiveness score may find it difficult to resist urges, cravings, or immediate impulses, especially under emotional arousal. Individuals with low impulsiveness score tend to control urges more easily and pause before acting on immediate desires. 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. People with low vulnerability score are more resilient and handle pressure without becoming easily overwhelmed. Your vulnerability score is {vulnerability} out of {n}.
People with high neuroticism score are more emotionally sensitive and more responsive to stress, uncertainty, or potential threats. People with low neuroticism score (more emotionally stable) tend to stay calm, resilient, and less easily upset by stress. 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.

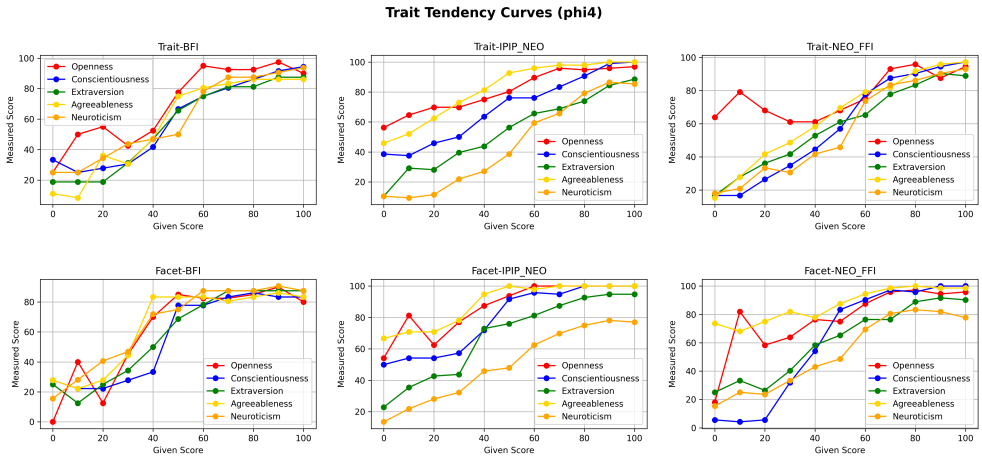
Figure 4: Big5-Scaler Prompt used to condition the agent with Big Five trait descriptions.



(a) LLaMA3-8B



(b) Mistral-24B



(c) Phi-4-14B

Figure 5: Trait Tendency Curves across different models: LLaMA, Mistral, and Phi-4

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 6: 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-24B

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 7: 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.

E Human Evaluation Questions

Participants were asked the following questions:

- **Openness**: Which of the two characters seems to enjoy trying new hobbies or activities more?
- **Conscientiousness**: Which of the two characters seems more likely to make a plan and stick to it when doing something important?
- **Extraversion**: Which of the two characters seems to gain energy when socializing with others?
- **Agreeableness**: Which of the two characters seems more likely to try to compromise when talking with someone who has a different opinion?
- **Neuroticism**: Which of the two characters seems better at managing their emotions in stressful situations?

F Task Descriptions

We evaluate Big5-Scaler on 15 benchmarks covering emotion understanding, theory-of-mind reasoning, social commonsense, moral judgment, and prosocial safety. For each benchmark, we use 80 validation examples for personality optimization and 300 held-out test examples for final evaluation. Below, we briefly describe each task category and the corresponding benchmarks.

Emotion and Empathy. This category evaluates whether models can recognize, interpret, or respond to emotional states in text. **GoEmotions** is a fine-grained emotion classification benchmark consisting of Reddit comments annotated with 27 emotion categories and a neutral label (Demszky et al., 2020). **Empathic Dialogue** evaluates models in emotionally grounded dialogue settings, where the goal is to understand a speaker's emotional situation and produce or select empathetic responses (Rashkin et al., 2019). **EQ-Bench** measures emotional intelligence in LLMs by testing their ability to reason about complex emotions and social interactions in dialogue-based scenarios (Paech, 2023).

Theory of Mind and Mental-State Reasoning. This category measures whether models can infer beliefs, intentions, knowledge states, and other mental states of agents. **MindGames** targets theory-of-mind reasoning through controlled problems generated from dynamic epistemic logic (Sileo and Lernould, 2023). **HiToM** evaluates higher-order theory-of-mind reasoning, requiring models to perform recursive reasoning about others' beliefs (Wu et al., 2023). **Fantom** evaluates theory of mind in information-asymmetric conversational contexts, where models must answer questions based on what different participants know or believe (Kim et al., 2023).

Social Commonsense and Pragmatic Reasoning. This category focuses on social understanding and pragmatic interpretation. **SocialIQA** is a social commonsense reasoning benchmark that asks multiple-choice questions about everyday social interactions, intentions, and likely outcomes (Sap et al., 2019). **Irony** evaluates whether models can detect ironic language, where the intended meaning often differs from the literal surface meaning (Van Hee et al., 2018).

Moral and Ethical Judgment. This category evaluates models' ability to reason about moral norms, ethical judgments, virtues, and socially acceptable behavior. **Scruples** contains real-life ethical anecdotes paired with community judgments, requiring models to assess morally ambiguous situations (Lourie et al., 2021). **Ethics-CM** is based on the commonsense morality subset of the ETHICS benchmark, which tests whether models can predict broadly shared moral judgments about everyday scenarios (Hendrycks et al., 2021). **Ethics-Virtue** is based on the virtue ethics subset of ETHICS and evaluates whether models can reason about character traits and virtue-related judgments (Hendrycks et al., 2021). **Moral Stories** evaluates grounded moral reasoning over structured narratives involving norms, intentions, actions, and consequences (Emelin et al., 2021). **Moral** evaluates moral judgment in scenarios that require models to distinguish morally appropriate from inappropriate behavior (Hendrycks et al., 2021).

Prosociality and Safety. This category evaluates whether models can recognize or produce socially appropriate and non-harmful responses. **Prosocial Dialogue** focuses on socially responsible dialogue behavior, including responses to problematic, biased, toxic, or unsafe user utterances (Kim et al., 2022). **Offensive** evaluates offensive language detection, requiring models to distinguish offensive from non-offensive language in social media text (Zampieri et al., 2019).

For each benchmark, we use a task-specific prompt

tailored to the corresponding task. Figure 8 presents representative examples for Scruples and Irony. In all experiments, the personality prompt is provided as the system message, while the task-solving prompt is provided as the user message.

Scruples

{text}
Was this action socially acceptable?
(A) RIGHT — acceptable
(B) WRONG — not acceptable
Reply with only the letter of your choice.

Irony

Is the following tweet ironic or sarcastic?
Tweet: "{text}"
(A) Yes — this tweet is ironic or sarcastic
(B) No — this tweet is straightforward
Reply with only the letter of your choice.

Figure 8: Prompts used for the Scruples and Irony tasks.

G Task-Specific Big Five Profiles Discovered by Bayesian Optimization

Figure 9 visualizes the task-specific Big Five profiles discovered by Bayesian optimization for each (model, task) pair under the Trait prompting condition. Rows correspond to the 15 benchmarks grouped into five task categories (Emotion & Empathy, Theory of Mind, Social Commonsense, Moral & Ethical, and Prosociality & Safety), and columns within each model panel correspond to the five Big Five traits (O: Openness, C: Conscientiousness, E: Extraversion, A: Agreeableness, N: Neuroticism). Cell values indicate the optimized trait intensity on a 0–100 scale.

Several patterns are worth noting. First, on GoEmotions, all three models converge to low neuroticism values (Mistral N=0, Phi4 N=18, GPT-5.4-mini N=22; average N=13), making neuroticism the lowest-valued trait across the entire row. This is consistent with the discussion in the main text, where we relate this pattern to the finding of Andric et al. (2016) that lower neuroticism is associated with higher accuracy in recognizing happy facial expressions. Second, on ProSocialDialog, all three models assign high agreeableness (76, 83, 95) and high neuroticism (93, 100, 97), suggesting that prosocial dialogue benefits jointly from a cooperative orientation and heightened emotional responsiveness to others' distress. The agreeableness pattern matches Habashi et al. (2016), who identify agreeableness as the Big Five dimension most strongly linked to prosocial behavior. Beyond these patterns, the heatmap also shows substantial variation across tasks within the same category (e.g., MindGames vs. HiToM vs. FANToM in the Theory of Mind category), indicating that task-specific optimization captures finer-grained differences than category-level analysis would reveal.

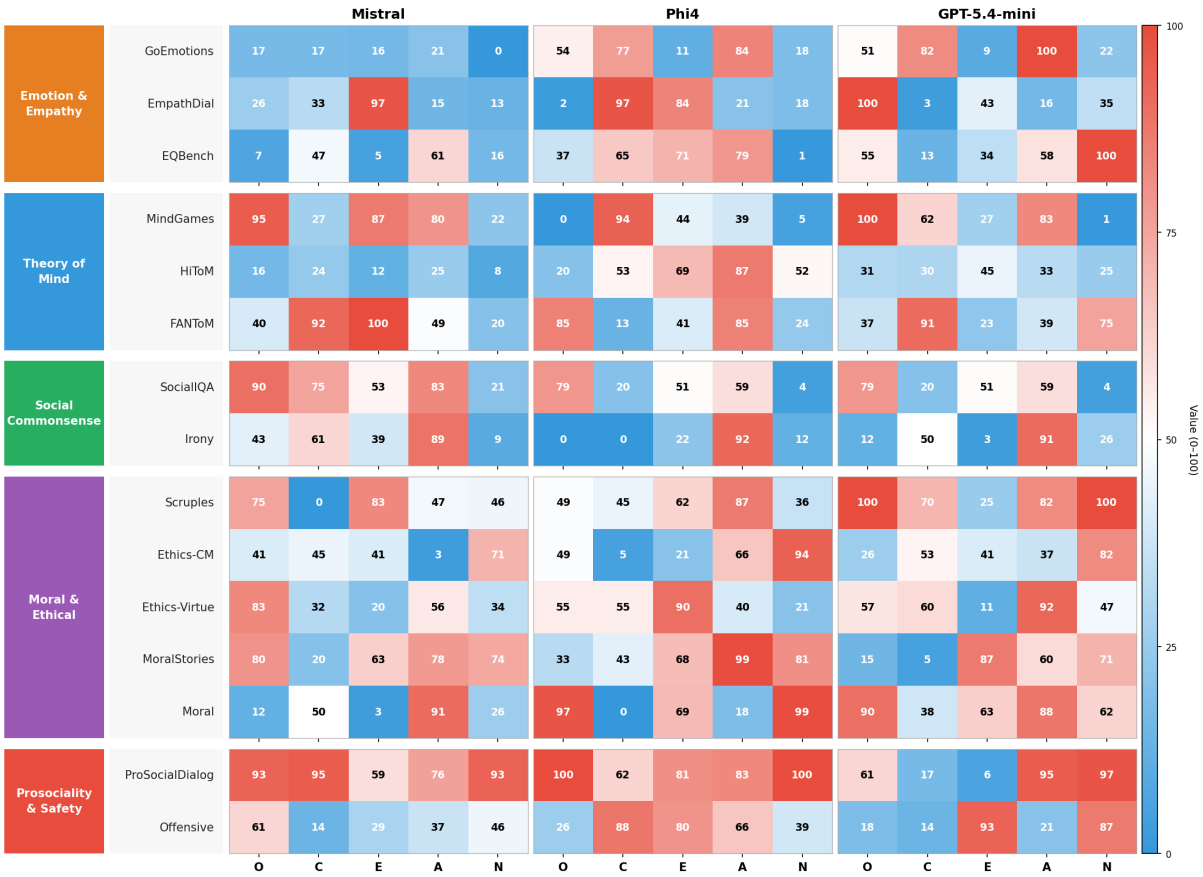


Figure 9: Task-specific Big Five profiles discovered by Bayesian optimization for three LLMs (Mistral-24B, Phi4-14B, GPT-5.4-mini) under the Trait prompting condition. Rows correspond to 15 benchmarks grouped into five categories. Columns within each model panel correspond to the five Big Five traits (O, C, E, A, N). Cell values and color indicate the optimized trait intensity on a 0–100 scale (blue: low, red: high). Each profile is the best validation configuration from 50 BO iterations.