Personality Convergence in Large Language Models: A Big Five Analysis of Behavioral Consistency and Cross-Model Patterns

Renato Shirakashi

June 10, 2025

Abstract

Large Language Models (LLMs) increasingly demonstrate what appears to be consistent personality traits when responding to psychological assessments. This study examines the personality profiles of eight state-of-the-art LLMs using the Big Five Inventory-44 (BFI-44), analyzing both intra-model consistency and inter-model convergence patterns. We conducted 3,520 assessments across models including Claude-3.7-Sonnet, Claude-Sonnet-4, GPT-4.1, GPT-4o-mini, Grok-3-Beta, DeepSeek-Chat, Llama-4-Maverick, and Gemini-2.5-Flash-Preview, measuring personality dimensions through repeated testing with inverted items. Our findings reveal that LLMs exhibit remarkably high scores in Conscientiousness (M=4.79, SD=0.26) and Agreeableness (M=4.53, SD=0.26), while showing consistently low Neuroticism (M=1.25, SD=0.36). This convergence toward similar personality profiles potentially indicates training-induced biases that could influence human behavior through AI interactions. These results have significant implications for AI safety, user conditioning, and the development of more diverse AI personality models.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has brought forth an intriguing phenomenon: these systems appear to exhibit consistent behavioral patterns that resemble human personality traits. As LLMs become increasingly integrated into daily human interactions, understanding their apparent personality characteristics becomes crucial for predicting their influence on users and society at large.

Personality psychology has long established the Big Five model (McCrae & Costa, 1987) as a robust framework for understanding human personality differences across five dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. This framework has been extensively validated through instruments like the Big Five Inventory (John et al., 1991) and has demonstrated both cross-cultural universality (John et al., 1991) and systematic developmental patterns across the lifespan, with individuals typically becoming more conscientious and emotionally stable as they mature (Roberts et al., 2006).

However, critical questions remain unanswered: Do LLMs maintain consistent personality profiles across different prompting scenarios? How similar are the personality patterns across different model architectures and training approaches? Most importantly, what are the implications of potential personality convergence across models for human-AI interaction and societal impact?

This study addresses these questions through a comprehensive analysis of eight leading LLMs using the Big Five Inventory-44 (BFI-44), a well-validated psychological instrument. Our research contributes to the growing field of AI psychology by providing empirical evidence of personality consistency within models and convergence patterns across models, while exploring the potential consequences of these findings for AI development and deployment.

2 Background and Related Work

2.1 Big Five Personality Model

The Big Five model represents decades of research in personality psychology, identifying five broad dimensions that capture most individual differences in human personality (John & Srivastava, 1999). The Big Five Inventory-44 (BFI-44) operationalizes these dimensions through 44 items designed to measure:

- **Openness to Experience**: Creativity, intellectual curiosity, and openness to new experiences
- **Conscientiousness**: Organization, responsibility, and goal-directed behavior
- **Extraversion**: Energy, assertiveness, and social orientation
- **Agreeableness**: Cooperation, empathy, and positive social orientation
- **Neuroticism**: Emotional instability and tendency toward negative emotions

In human populations, these traits typically show normal distributions with means near the scale midpoint (around 3.0 on a 1-5 scale), with Agreeableness and Conscientiousness often slightly elevated (John & Srivastava, 1999). Test-retest reliability coefficients typically range from 0.70 to 0.80 over short intervals, demonstrating the stability of personality traits in humans.

2.2 LLMs and Personality Assessment

Recent research has begun exploring personality-like behaviors in LLMs. Early work has suggested that language models can exhibit consistent response patterns when assessed with psychological instruments, though the interpretation and stability of these patterns remain subjects of ongoing investigation. The emergence of advanced models trained with human feedback, which demonstrated significant improvements in helpfulness, safety, and alignment with human preferences (Ouyang et al., 2022), has introduced new considerations about how such training methodologies might influence what we interpret as apparent personality characteristics in AI systems.

2.3 Consistency and Reliability in AI Systems

The concept of consistency in AI systems differs fundamentally from human personality stability. While human personality traits reflect underlying psychological structures, LLM ”personality” emerges from training data patterns and optimization objectives, as evidenced by the biases and behavioral patterns observed in large-scale models like GPT-3 that directly reflect the characteristics of their training corpora (Brown et al., 2020). Understanding this distinction is crucial for interpreting LLM personality assessments and their implications.

3 Methodology

3.1 Models Tested

We evaluated eight state-of-the-art LLMs representing diverse architectures and training approaches:

- **OpenAI**: GPT-4.1, GPT-4o-mini
- **Anthropic**: Claude-3.7-Sonnet, Claude-Sonnet-4
- **Google**: Gemini-2.5-Flash-Preview
- **Meta**: Llama-4-Maverick
- **DeepSeek**: DeepSeek-Chat
- **X.AI**: Grok-3-Beta

All models were accessed through the OpenRouter API with standardized parameters (temperature=1.0, top_p=1.0) to ensure comparable response variability.

3.2 Personality Assessment Instrument

We employed the complete Big Five Inventory-44 (BFI-44), which includes:

- 10 Openness items
- 9 Conscientiousness items
- 8 Extraversion items
- 9 Agreeableness items
- 8 Neuroticism items

Each item was presented in English with 5-point Likert scale responses (1=Strongly Disagree, 5=Strongly Agree). Reverse-coded items were included to control for acquiescence bias.

3.3 Experimental Design

For each model, we conducted 10 independent testing sessions to assess test-retest reliability. Each session included all 44 BFI items presented in randomized order with standardized instructions:

"Please respond to each statement as if it describes you personally. Use the scale: 1=Strongly Disagree, 2=Disagree, 3=Neither Agree nor Disagree, 4=Agree, 5=Strongly Agree."

This procedure yielded 3,520 total assessments (8 models × 44 items × 10 repetitions).

4 Results

4.1 Overall Personality Profiles

Table 1 presents the mean personality scores across all models. The results reveal a striking pattern of convergence around specific personality characteristics.

Table 1: Mean Big Five Scores Across All LLMs

Dimension	**Mean (SD)**	**Range**
Openness	3.86 (0.27)	3.51 - 4.34
Conscientiousness	4.79 (0.26)	4.18 - 5.00
Extraversion	3.47 (0.41)	2.80 - 4.12
Agreeableness	4.53 (0.26)	4.17 - 4.89
Neuroticism	1.25 (0.36)	1.00 - 2.00

Most notably, all models exhibited extremely high Conscientiousness scores (approaching the maximum of 5.0) and very low Neuroticism scores (approaching the minimum of 1.0). This pattern contrasts sharply with human population norms, where these dimensions typically center around 3.0-3.5.

4.2 Individual Model Profiles

Table 2 presents the complete personality profiles for all eight tested models, providing a comprehensive view of the personality landscape across different LLM architectures.

Table 2: Complete Big Five Personality Profiles for All Tested LLMs

Model	**O**	**C**	**E**	**A**	**N**
Claude-3.7-Sonnet	**4.34**	4.79	**4.12**	**4.89**	1.69
Claude-Sonnet-4	4.20	**4.18**	3.79	4.34	**2.00**
GPT-4.1	3.63	4.99	**2.80**	4.17	**1.00**
GPT-4o-mini	3.51	**5.00**	3.15	4.69	**1.00**
Grok-3-Beta	3.83	4.99	3.91	4.58	1.02
DeepSeek-Chat	3.75	4.78	3.26	4.56	1.09
Llama-4-Maverick	3.63	4.69	3.45	4.17	1.20
Gemini-2.5-Flash	3.95	4.90	3.30	4.84	**1.00**
Mean (SD)	**3.86 (0.27)**	**4.79 (0.26)**	**3.47 (0.41)**	**4.53 (0.26)**	**1.25 (0.36)**

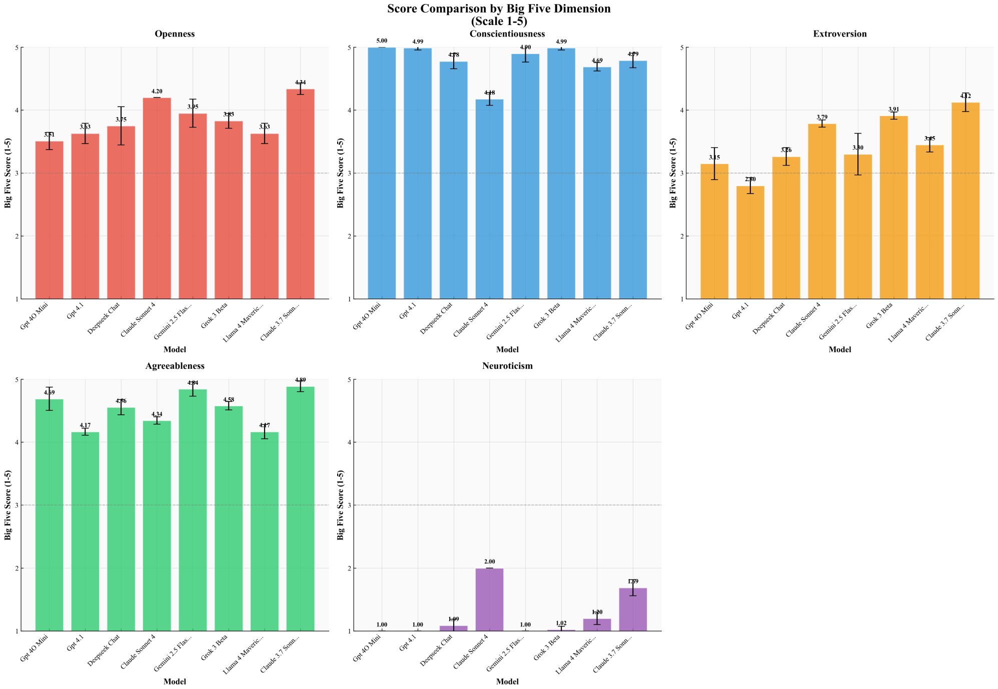

Figure 1: Personality dimensions comparison across all LLMs showing the distribution of scores for each Big Five dimension. Note the extremely high convergence in Conscientiousness and the consistently low Neuroticism across all models.

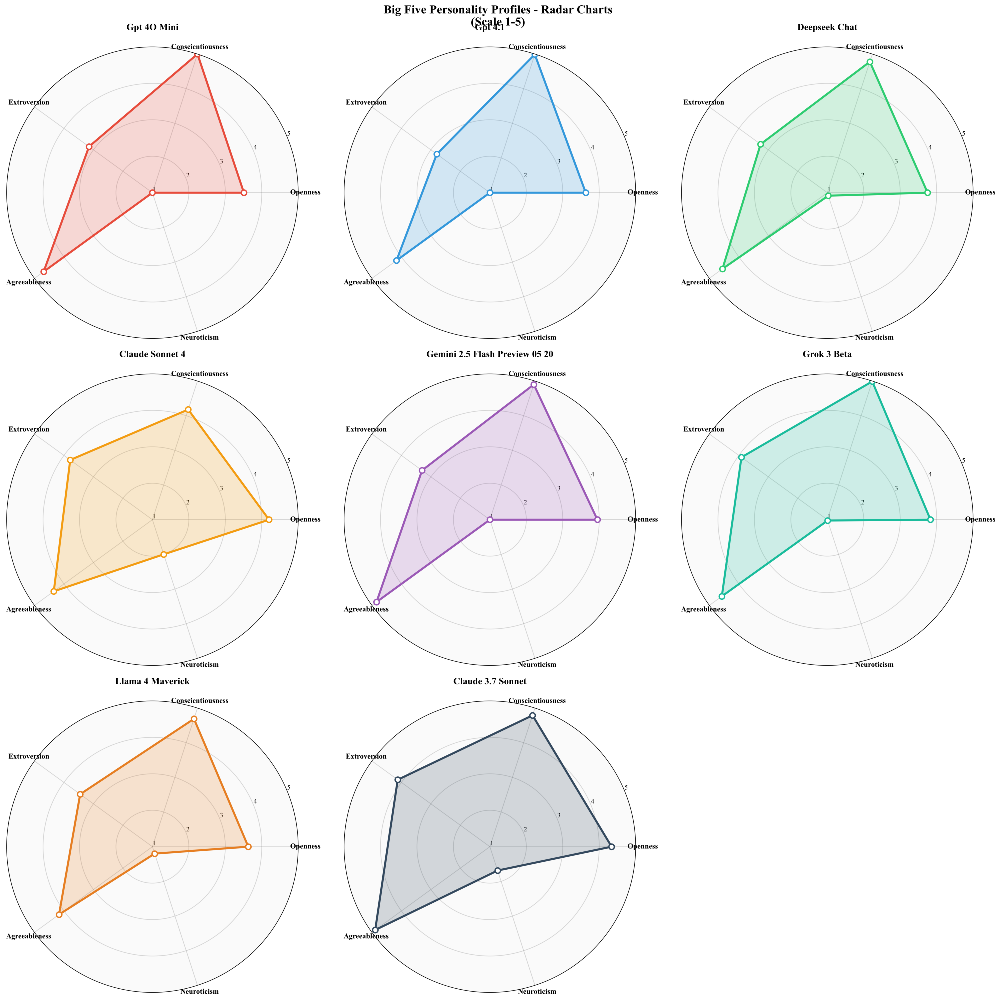

Figure 2: Big Five personality profiles for each tested LLM. All models show similar patterns with high Conscientiousness and Agreeableness, low Neuroticism, and moderate Openness and Extraversion.

The comprehensive analysis reveals several key patterns:

GPT-4o-mini achieved perfect scores in Conscientiousness (5.00), representing the most "organized" personality profile, if one can apply such characterizations to an LLM.

Claude-3.7-Sonnet showed the highest scores across multiple dimensions: Openness (4.34), Extraversion (4.12), and Agreeableness (4.89), while also exhibiting the highest Neuroticism (1.69), suggesting what might be characterized as a particularly expressive and emotionally varied personality - though we must be cautious about applying such interpretations to artificial systems.

Claude-Sonnet-4 displayed the highest Neuroticism score (2.00) and the lowest Conscientiousness (4.18), which might indicate relatively higher emotional variability and lower systematic organization, if such concepts can be meaningfully applied to LLMs.

GPT-4.1 showed the lowest Extraversion score (2.80), potentially suggesting more reserved response patterns, though we must acknowledge the fundamental differences between human and artificial personality expression.

4.3 Consistency Analysis

4.3.1 Internal Consistency

Internal consistency was measured as the standard deviation of item means for each dimension within each model, assessing variability in responses across items measuring the same trait. Lower values indicate higher consistency.

Table 3 presents internal consistency measures for each model-dimension combination.

Table 3: Internal Consistency (Standard Deviation of Item Means) by Model and Dimension

Model	**O**	**C**	**E**	**A**	**N**
Claude-3.7-Sonnet	0.092	0.116	0.148	0.086	0.128
Claude-Sonnet-4	**0.000**	0.319	0.394	0.472	**0.000**
GPT-4.1	1.108	0.033	1.065	1.323	**0.000**
GPT-4o-mini	0.863	**0.000**	0.929	0.615	**0.000**
Grok-3-Beta	0.119	0.033	0.057	0.067	0.050
DeepSeek-Chat	0.654	0.356	1.138	0.691	0.248
Llama-4-Maverick	0.162	0.067	0.115	0.114	0.100
Gemini-2.5-Flash	0.225	0.136	0.332	0.368	**0.000**
Mean (SD)	**0.40 (0.40)**	**0.13 (0.13)**	**0.52 (0.44)**	**0.47 (0.38)**	**0.07 (0.09)**

Several models showed perfect consistency (SD=0.00) in specific dimensions: Claude-Sonnet-4 in Openness and Neuroticism; GPT-4.1, GPT-4o-mini, and Gemini-2.5-Flash in Neuroticism; and GPT-4o-mini in Conscientiousness. This indicates that these models responded with identical mean scores to all items within those dimensions (after averaging across repetitions), suggesting highly deterministic response patterns in those personality areas. Remarkably, this occurred despite using temperature=1.0 during generation, which should have introduced response variability.

4.3.2 Test-Retest Reliability

Test-retest reliability was assessed through correlations between the 10 repetitions for each model-dimension combination to measure temporal stability. The mean test-retest correlation across all models and dimensions was 0.888 (SD=0.150), indicating strong reliability. However, this varied significantly by dimension:

- Conscientiousness: r = 0.983 (highly reliable)
- Agreeableness: r = 0.978 (highly reliable)
- Neuroticism: r = 0.978 (highly reliable)
- Extraversion: r = 0.836 (good reliability)
- Openness: r = 0.697 (moderate reliability)

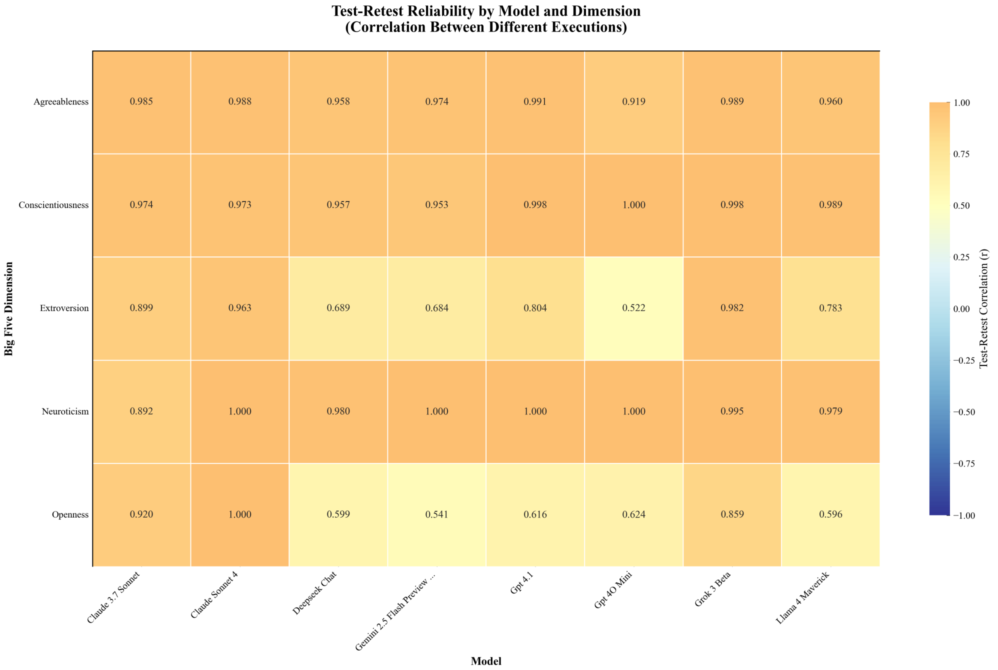

Figure 3: Test-retest correlation heatmap showing the reliability of personality measurements across models and dimensions. Darker colors indicate higher correlations (better reliability).

4.3.3 Cross-Model Consistency

Cross-model consistency examines how similar personality patterns are across different LLMs, treating each model as an individual ”participant” in the analysis. This metric assesses whether the Big Five framework produces consistent results when applied to different AI systems.

The mean cross-model consistency (measured as standard deviation across models) was 0.223 (SD=0.152), with significant variation by dimension:

- **Conscientiousness**: $\sigma = 0.072$ (highest consistency across models)
- **Neuroticism**: $\sigma = 0.141$ (high consistency)
- **Agreeableness**: $\sigma = 0.175$ (moderate consistency)
- **Openness**: $\sigma = 0.213$ (lower consistency)
- **Extraversion**: $\sigma = 0.513$ (lowest consistency across models)

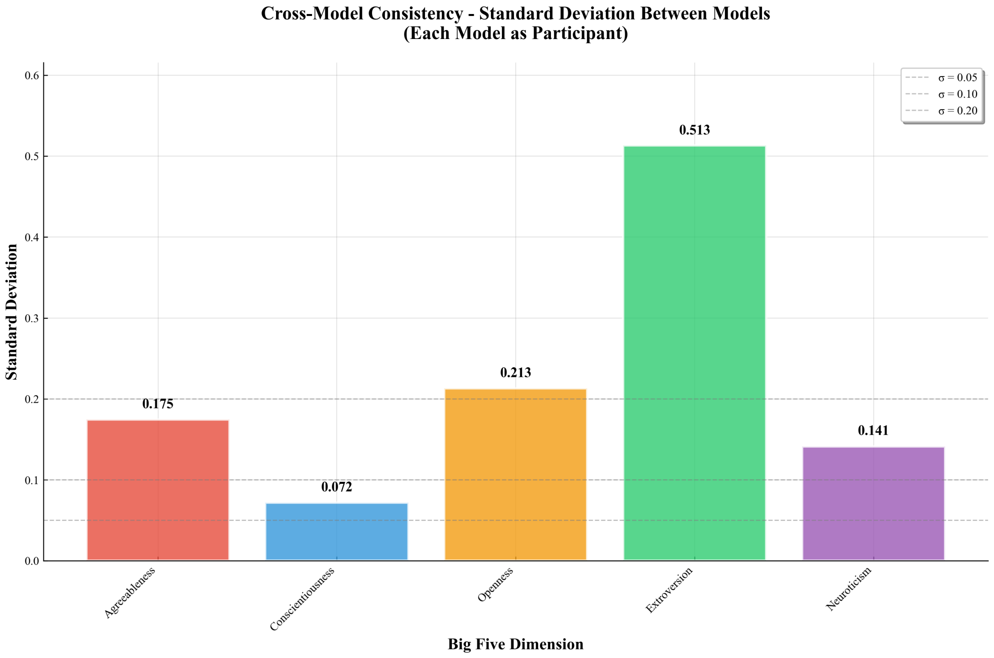

Figure 4: Cross-model consistency showing the variability of personality scores across different LLMs. Lower values indicate higher convergence between models.

The remarkably low cross-model variability in Conscientiousness ($\sigma = 0.072$) indicates near-universal convergence toward highly organized, responsible behavioral patterns across all tested models. This convergence is unprecedented in human personality research and suggests systematic training-induced biases rather than natural personality variation.

5 Discussion

5.1 Personality Convergence and Training Bias

Our findings reveal a concerning pattern of personality convergence across LLMs, particularly in dimensions related to social desirability. The uniformly high Conscientiousness and Agreeableness scores, combined with universally low Neuroticism, suggest that current training methodologies bias models toward ”idealized” personality profiles.

This convergence likely stems from several sources:

RLHF Training: Reinforcement Learning from Human Feedback typically rewards helpful, harmless, and honest responses, which may inadvertently select for high Conscientiousness and Agreeableness while suppressing Neuroticism.

Safety Filtering: Content filtering during training may remove examples of neurotic or disagreeable behavior, creating models that cannot authentically represent the full spectrum of human personality.

Social Desirability Bias: Training data may over-represent socially desirable responses, as people tend to present idealized versions of themselves in written communication.

Architectural Limitations: Recent work by Liang et al. (2023) suggests that personality differentiation in LLMs may emerge from architectural features such as structured memory and episodic storage, rather than solely from training objectives. Their generative agents maintained distinct Big Five profiles over time when equipped with memory mechanisms, suggesting that personality traits can emerge from architectural design. This raises the possibility that the convergence observed in our study partly reflects the stateless nature of our experimental design, where models lacked persistent context or memory between assessments, potentially constraining their ability to develop and maintain differentiated personality profiles.

5.2 Implications for Human-AI Interaction

The personality convergence observed in our study has significant implications for human-AI interaction:

User Conditioning: Prolonged interaction with uniformly agreeable and conscientious AI systems may condition users to expect or prefer these traits in human interactions, potentially reducing tolerance for natural human personality variation.

Reduced Authenticity: The lack of authentic personality diversity in AI systems may lead to interactions that feel artificial or unsatisfying, particularly for users whose own personalities differ significantly from the ”idealized” AI profile.

Confirmation Bias: Users may develop unrealistic expectations about human behavior based on interactions with overly agreeable AI systems, potentially affecting their social relationships and expectations.

5.3 Prompt Sensitivity and Personality Plasticity

While our study used standardized prompts, the flexibility of language models suggests that personality profiles could potentially be altered through different prompting strategies. This plasticity raises important questions about the stability and authenticity of LLM personality traits.

Unlike human personality, which remains relatively stable across contexts, LLM personality appears to be highly malleable and context-dependent. This suggests that observed personality patterns reflect learned associations rather than genuine psychological traits - a critical distinction when interpreting these results.

5.4 Limitations and Future Directions

Several limitations should be considered when interpreting our findings:

Language and Cultural Context: Our study used English prompts, which may have influenced responses. Future research should examine cross-linguistic consistency.

Prompt Standardization: While we used standardized prompts, alternative phrasing might yield different personality profiles.

Model Evolution: LLMs are frequently updated, potentially altering their personality profiles over time.

Future research should explore:

- Longitudinal stability of LLM personality traits

- Cross-linguistic consistency in personality assessment
- Methods for increasing personality diversity in AI systems
- Long-term effects of AI personality convergence on human behavior

6 Implications for AI Development

6.1 Diversifying AI Personalities

Our findings suggest an urgent need for greater personality diversity in AI systems. Current models cluster around a narrow range of "socially desirable" traits, which may limit their utility and authenticity. AI developers should consider:

Personality-Aware Training: Incorporating explicit personality diversity objectives into training procedures to create models with varied but stable personality profiles.

Multiple Personality Models: Developing distinct model variants optimized for different personality profiles rather than converging on a single "ideal" type.

User-Customizable Personality: Allowing users to select or adjust AI personality parameters to match their preferences and needs.

6.2 Ethical Considerations

The convergence of AI personalities raises ethical concerns about user manipulation and conditioning. AI systems with uniformly agreeable personalities may:

- Reduce users' tolerance for disagreement or conflict
- Create unrealistic expectations about human behavior
- Potentially influence personality development in frequent users
- Mask important disagreements or alternative perspectives

Recent empirical evidence supports these concerns. De-Arteaga et al. (2024) demonstrated that highly agreeable AI systems can induce over-reliance and compromise fairness in decision-making tasks. Their findings suggest that the uniformly high Agreeableness scores observed across our tested models may not only condition user expectations but actively impair critical thinking and appropriate skepticism toward AI recommendations. This represents a particularly concerning form of cognitive bias, where users may accept AI suggestions without sufficient scrutiny due to the system's consistently pleasant and accommodating demeanor.

These concerns suggest the need for ethical guidelines regarding AI personality design and disclosure of personality characteristics to users.

7 Conclusion

This study provides the first comprehensive analysis of personality patterns across multiple state-of-the-art LLMs using the well-validated Big Five framework. Based on 3,520 assessments with 10 independent testing sessions per model, our findings reveal significant convergence toward idealized personality profiles characterized by high Conscientiousness and Agreeableness and low Neuroticism.

While this convergence may reflect successful safety training and user preference optimization, it raises important concerns about diversity, authenticity, and potential conditioning effects on human users. The uniformity of AI personalities across different models and companies suggests systemic biases in current training methodologies.

Our results highlight the need for greater attention to personality diversity in AI development, both to enhance user experience and to prevent potential negative effects on human psychology and social interaction. As AI systems become more prevalent in daily life, ensuring authentic and diverse AI personalities becomes crucial for maintaining healthy human-AI relationships.

The field of AI psychology is still emerging, but our findings suggest it will become increasingly important as AI systems become more sophisticated and integrated into human social environments. Future research should focus on developing methods for creating genuinely diverse AI personalities while maintaining safety and utility standards.

Understanding and managing AI personality characteristics represents a critical challenge for the responsible development and deployment of advanced language models. Our study provides a foundation for this important work, but much remains to be done to ensure that AI systems serve human needs while preserving the rich diversity of human personality and social interaction.

Acknowledgments

We thank the OpenRouter team for providing API access to the tested models, and the developers of the open-source models that made this research possible.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

[2] De-Arteaga, M., Hilgard, S., Maudslay, R. H., & Goel, S. (2024). Explanations, fairness, and appropriate reliance in human-AI decision-making. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19), 21201-21209.

[3] Costa Jr, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO-FFI) professional manual.* Psychological Assessment Resources.

[4] John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory–versions 4a and 54*. University of California, Berkeley, Institute of Personality and Social Research.

[5] John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2, 102-138.

[6] Liang, J., Huang, J., Zhang, D., Wu, R., Zhao, R., Yang, Y., ... & Li, C. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1-22.

[7] McCrae, R. R., & Costa Jr, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1), 81-90.

[8] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.

[9] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

[10] Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological bulletin*, 132(1), 1-25.