Decoding Emergent Big Five Traits in Large Language Models: Temperature-Dependent Expression and Architectural Clustering

Christos-Nikolaos Zacharopoulos*
Independent Researcher
christonik@gmail.com

Revekka Kyriakoglou
Université Paris 8
Vincennes – Saint-Denis, Paris, France
revekka.kyriakoglou@univ-paris8.fr

Abstract

As Large Language Models (LLMs) become integral to human-centered applications, understanding their personality-like behaviors is increasingly important for responsible development and deployment. This paper systematically evaluates six LLMs, applying the Big Five Inventory-2 (BFI-2) framework, to assess trait expressions under varying sampling temperatures. We find significant differences across four of the five personality dimensions, with Neuroticism and Extraversion susceptible to temperature adjustments. Further, hierarchical clustering reveals distinct model clusters, suggesting that architectural features may predispose certain models toward stable trait profiles. Taken together, these results offer new insights into the emergence of personality-like patterns in LLMs and provide a new perspective on model tuning, selection, and the ethical governance of AI systems. We share the data and code for this analysis here: https://osf.io/bsvzc/?view_only=6672219bede24b4e875097426dc3fac1

1 Introduction

The increasing use of Large Language Models (LLMs) as substitutes for human interaction marks a significant shift in societal dynamics. Individuals now engage with LLMs not only for retrieving information but also in contexts resembling intimate human dialogue, including seeking emotional support, personal advice, and even therapeutic guidance (Stade et al., 2024; Li et al., 2023). As these disembodied interactions become more commonplace, LLMs are beginning to occupy roles once reserved for human experts, counselors, or friends. This development has profound implications: by altering traditional models of communication, mental health care, and interpersonal relationships, LLMs challenge established norms of trust, empathy, and reliability. Moreover, as these systems become more human-like in their conversational styles, users increasingly anthropomorphize them, projecting cognitive and emotional qualities onto what are, at their core, statistical models. Such anthropomorphization raises critical questions about the nature of "personality" in LLMs and how users may rely on these perceived traits when forming judgments, seeking comfort, or making important decisions.

In response to these emergent issues, researchers have begun probing whether and how LLMs exhibit human-like personality characteristics (Serapio-García et al., 2023; Jiang et al., 2023; Mao et al., 2023; Zhan et al., 2024; Noh and Chang, 2024). Anchoring such inquiries in robust psychological frameworks helps clarify otherwise nebulous concepts. The Big Five personality model—capturing openness, conscientiousness, extraversion, agreeableness, and neuroticism (McCrae and Costa, 1997)—serves as a widely accepted and empirically supported tool for understanding human personality. Although its application to LLMs is still in its infancy, a growing body of work suggests that LLMs may indeed reflect trait-like patterns in their generated responses (Lee et al., 2024). Understanding these patterns is far from a purely academic exercise; it has far-reaching implications for the design, deployment, and ethical governance of AI-driven communication.

Despite initial efforts, critical gaps remain. Existing literature has primarily focused on whether traits like those in the Big Five emerge in LLMs, but not on the underlying mechanisms that give rise to these traits or the conditions that influence their stability. For instance, there is limited insight into how model architecture, training data composition, and sampling strategies interact to shape the personality-like behaviors observed. Within the broader research effort to contextualize the nature of LLM "personality," examining additional factors, like the temperature parameter, can provide

* Corresponding author.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1678–1685
December 20-24, 2025 ©2025 Association for Computational Linguistics

fresh perspectives.

This paper contributes new analytical depth along two axes. (1) We systematically examine trait expression under varying sampling temperatures to characterize how a core decoding control modulates personality-like outputs. (2) We use agglomerative hierarchical clustering to uncover model-level patterns of similarity in trait profiles, providing evidence of structural tendencies across architectures. Together, these analyses move beyond simple personality testing and clarify how model design and decoding interact to shape personality-like behaviors.

We advance this exploration by evaluating six comparably sized LLMs using the Big Five Inventory–2 (BFI-2) questionnaire (Soto and John, 2017), a validated and reliable measure of human personality traits. Beyond simply classifying the presence or absence of trait-like patterns, we systematically vary the temperature parameter to investigate its role as a stochastic decoding control that may modulate responses. We also attempt to identify if and how LLMs cluster natively as a factor of their personality responses. Through this multifaceted analysis, we aim to deepen our understanding of what it means for LLMs to exhibit personality-like traits, identify the factors that modulate these expressions, and lay the groundwork for more accountable and human-centered design and governance of AI communication systems.

2 Methods

We employed a diverse ensemble of state-of-the-art large language models (LLMs), each fine-tuned for conversational tasks and equipped with distinct attention mechanisms. The selected models spanned varying parameter scales, vocabulary sizes, and attention mechanisms. We utilized the *Llama 3 8B* model from the Llama series, featuring 8 billion parameters, a vocabulary size of 128,256 tokens, and Grouped-Query Attention (GQA) (AI, 2024a). The *Mistral 7B* model, with 7.3 billion parameters and a vocabulary size of approximately 131,000 tokens, incorporated Grouped-Query Attention (GQA) (AI, 2023). The *MythoMax L2 13B* model from the Gryphe series combined 13 billion parameters with an 8,000-token context length (Gryphe, 2024). The *Gemma 9B* model, with 9 billion parameters and a vocabulary size of 300,000 tokens, employed dynamic attention scaling (Google, 2024). The Q*wen 7B* model utilized a vocabulary of over 150,000 tokens alongside sliding window attention (Cloud, 2023). Lastly, the *StripedHyena 7B* model, featuring 7 billion parameters and a vocabulary size of 280,000 tokens, implemented block-sparse attention (AI, 2024b). To systematically evaluate the influence of sampling temperature on model outputs, we conducted a series of text-generation experiments using a fixed prompt and instructions designed to simulate a personality test response scenario. Specifically, we varied the temperature parameter from 0 to 2 in increments of 1, yielding a total of 21 experimental conditions for each of the 60 main questions of the BFI-2 questionnaire. This range was chosen to capture an extensive spectrum of possible sampling behaviors, from highly deterministic (temperature = 0) to increasingly stochastic regimes. Nevertheless, this range falls within a reasonable space of exploration.

Statistical analysis comprised three main components. First, we performed non-parametric between-model across personality type comparisons using Kruskal-Wallis H-tests ($\alpha = 0.05$) to detect significant differences in trait expressions.

Second, we conducted temperature sensitivity analyses through multiple linear regression models for each trait dimension, with temperature as the predictor variable and trait scores as the response variable, complemented by Pearson correlation coefficients (r) to assess relationship strength and direction. The regression analyses aim to quantify the associations between temperature and trait scores, rather than implying causal relationships or theoretical psychological mappings.

Third, we employed agglomerative hierarchical clustering with Ward's minimum variance method using Euclidean distance metrics to identify model groupings. These groupings were validated through trait covariance matrices to examine inter-trait relationships.

Domain	**Statistic**	**p-value**
Extraversion	40.7803	<0.01
Agreeableness	65.3067	<0.01
Conscientiousness	63.0415	<0.01
Neuroticism	9.2691	n.s.
Openness to Experience	58.1957	<0.01

Table 1: Kruskal-Wallis test results for personality domains. The table shows the test statistic and corresponding p-values for each domain.

Domain	R^2	Pearson Cor.	p-value
Neuroticism	0.3486	-0.5904	<0.05
Extraversion	0.2521	0.5021	<0.05
Agreeableness	0.0343	-0.1853	n.s.
Conscientiousness	0.0257	0.1602	n.s.
Openness	0.0003	0.0178	n.s.

Table 2: Linear regression results for personality domains as a function of temperature.

3 Results

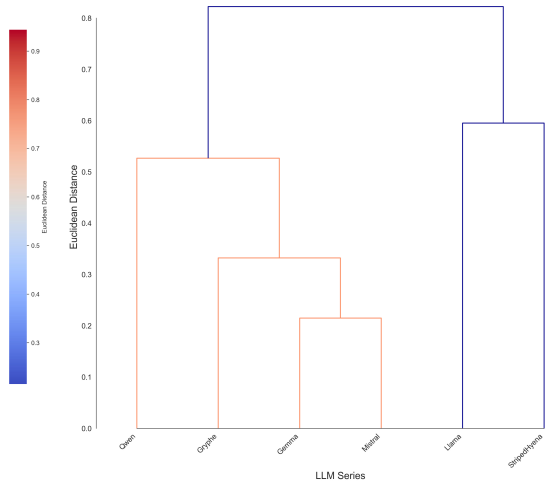

Figure 1: Hierarchical clustering of models based on personality profiles, revealing distinct groupings and architectural influences on trait expressions.

We observed significant differences in the expression of four out of the five Big Five personality traits across models. Kruskal–Wallis tests indicated statistically significant variation in Extraversion ($H = 40.7803, p < 0.01$), Agreeableness ($H = 65.3067, p < 0.01$), Conscientiousness ($H = 63.0415, p < 0.01$), and Openness to Experience ($H = 58.1957, p < 0.01$). In contrast, Neuroticism did not differ significantly between models ($H = 9.2691$, n.s.).

Our temperature sensitivity analysis revealed that certain traits were more strongly influenced by the sampling temperature parameter. Neuroticism showed the most pronounced association with temperature ($R^2 = 0.3486, r = -0.5904, p < 0.05$): as temperature decreased, Neuroticism scores increased, suggesting that more deterministic outputs (lower temperatures) yield higher Neuroticism levels. Extraversion also correlated significantly with temperature ($R^2 = 0.2521, r = 0.5021, p < 0.05$), but in the opposite direction—more stochastic sampling (higher temperature) produced more extraverted responses. By contrast, Agreeableness ($R^2 = 0.0343, r = -0.1853$), Conscientiousness ($R^2 = 0.0257, r = 0.1602$), and Openness ($R^2 = 0.0003, r = 0.0178$) were not significantly affected by temperature (all $p > 0.05$).

The resulting dendrogram reveals notable patterns of similarity and divergence among the models. Notably, the Qwen and StripedHyena models span the boundaries of the dendrogram, indicating maximal pairwise dissimilarity, with Llama being adjacent to the StipedHyena model and forming a separate cluster. The Gemma and Mistral models form a cluster positioned in the center of the dendrogram, whereas Gryphe stands in between this cluster and the Qwen model.

Overall, our results indicate that large language models do exhibit stable, personality-like trait patterns that vary according to architectural characteristics and sampling parameters. While some domains (e.g., Neuroticism and Extraversion) are sensitive to temperature, others remain more robust under changing conditions.

4 Discussion

This study tested the response profiles of six large language models across the Big Five personality traits. Our results revealed that Extraversion, Agreeableness, Conscientiousness, and Openness to Experience were all significantly different. In contrast, Neuroticism did not reach statistical significance. The lack of significant variation in Neuroticism suggests a consistent baseline in the models' responses regarding emotional reactivity. This consistency could be attributed to the training data encompassing a wide range of emotional expressions, thereby balancing positive and negative emotional content. As a result, the models may not disproportionately reflect neurotic characteristics, leading to a more stable and less emotionally reactive profile. By systematically manipulating sampling temperature, we uncovered parametric sensitivities underlying LLM responses.

Sampling temperature—a common decoding parameter—affects not only token diversity and lexical creativity but also generates outputs resembling specific personality traits. Specifically, lowering the temperature consistently results in more "neurotic" outputs. Temperature modulates stochasticity in generation. Human research relating creativity to Extraversion and Neuroticism offers a useful interpretive parallel, although we do not

Figure 2: Effects of sampling temperature on personality traits, demonstrating sensitivity in Neuroticism and Extraversion.

treat temperature itself as a psychological construct. For instance, Conner et al. (Conner and Silvia, 2015) found that neurotic individuals show reduced creativity, especially under anxiety, due to a prevention-focused mindset and heightened threat sensitivity that impede creative engagement. Similarly, Li et al. (Li et al., 2022) reported that neuroticism negatively affects creativity among college students. Furthermore, Krumm et al. (Krumm et al., 2018) provided empirical evidence that neuroticism is inversely related to creativity, indicating that higher levels of neuroticism are associated with lower creative abilities in children. We also find that increasing the sampling temperature leads to outputs with higher extraversion ratings. This observation aligns with existing research on the relationship between extraversion and creativity. For instance, Davis et al. (Davis et al., 2011) demonstrated that extraversion significantly predicts self-reported creativity across various domains among college students. Additionally, Michinov and Michinov (Michinov and Michinov, 2021) revealed that certain personality profiles, which include extraverted traits, positively influence creative performance, especially under conditions of social isolation, such as the COVID-19 lockdown. Nevertheless, our analyses did not reveal significant associations between agreeableness, conscientiousness, and openness and the outcome variable. We hypothesize that two factors might be responsible. First, we examined the effect of temperature agglomerating across all six LLMs. Given the variability in responses observed (table 1), it is highly likely that such effects exist at the individual LLM level. Additionally, we assumed that temperature can be considered a proxy for creativity, but creativity is a multi-component trait, difficult to define and quantify (Sternberg, 2018). Temperature settings in LLMs primarily influence the randomness of responses, which may not fully capture creativity's nuanced, multidimensional aspects.

Our clustering analysis reveals that both model architecture and training data significantly shape the emergent "personality" of large language models. While the content and diversity of training data unquestionably influence learned representations, specific design decisions—such as attention mechanisms (sliding window, Grouped-Query, dynamic scaling, or *Hyena Blocks*) and vocabulary sizes—can yield pronounced output differences. For instance, Qwen 7B (sliding window, 150k tokens) diverges from Gemma 9B (dynamic attention, 300k tokens) and Mistral 7B (Grouped-Query Attention, 131k tokens), illustrating how contrasting attention strategies overshadow shared data attributes. Meanwhile, Llama 3 8B (GQA) stands apart from StripedHyena 7B, whose *Hyena Blocks*-based block-sparse attention (280k tokens) further accentuates unique context-processing. Strikingly, our hierarchical analysis places Qwen and StripedHyena at opposite ends of the similarity spectrum, underscoring how the interplay of architecture (sliding window vs. Hyena Blocks) and vocabulary range (150k vs. 280k tokens) can produce the most pronounced separation in model "personalities."

These findings offer only a glimpse into the long-standing "nature vs. nurture" debate as it applies to emergent traits in LLMs, indicating that both inherent architectural design ("nature") and training data ("nurture") play consequential roles. Our goal here is to highlight the importance of disentangling these factors rather than claiming a fully comprehensive characterization. Future work could employ more controlled methodologies—such as curated corpus studies, ablation experiments targeting specific architectural choices, or fine-grained attribution analyses—to more systematically trace the origins of these traits and refine our understanding of how LLM "personalities" come to be.

Our analysis reveals that LLMs can display

personality-like patterns and that these expressions are influenced, at least in part, by decoding parameters like sampling temperature. These insights invite future investigations into the interplay between architecture, training data, and decoding strategies, ultimately informing both the theory and practice of refining LLM behavior.

5 Limitations

This study is subject to limitations, which we pinpoint in a three-fold fashion. First, the questionnaire was administered to the LLMs for a fixed temperature in a one-shot manner. It is possible that a given LLM would not have provided the same response had it been prompted again. Nevertheless, this does not affect the study's main conclusion. The statistics for each personality component were calculated based on the assumption of differing medians calculated over a temperature range (see Figure 3). Thus, the variability induced by the temperature sampling should have captured any likely non-deterministic behavior. Second, the temperature-regression analysis, albeit significant in two out of five traits, fails to explain a significant part of the underlying variance (34 & 25% for the Neuroticism and Extraversion, respectively). These regressions are exploratory association measures and should not be interpreted as causal indicators or psychological attributions. We hypothesize that this stems from the agglomerative nature of the analysis. Indeed, there appears to be a region (temperature $\in$ [0, 0.5]) where the variance of responses is relatively stable, followed by an unstable response profile. We attribute this to different temperature-sensitivity levels of individual LLMs, but exploring this sensitivity further falls out of the scope of the current study. Nonetheless, our analysis draws intuitive parallels between a dedoding parameter attributed to stochasticity, and personality components that can be affected by such virtue. We, hence, argue that these results surpass the noise level and provide tangible insights into the latent behavior of LLMs. Thirdly, our "nature vs. nurture" analysis is neither causal nor claims to be. Investigating the exact decomposition of personality as a function of lexical exposure or architectural genotype would require a dedicated set of experiments. Our investigation attempts to provide causal hints that we hope will drive interest towards further research in the latent behavioral space of LLMs. We believe that, in the accelerated integration and anthropomorphization era of LLMs, such research is cardinal.

References

Meta AI. 2024a. Llama 3: A more capable and efficient language model. https://ai.meta.com/llama/. Accessed: 2024-11-19.

Mistral AI. 2023. Mistral 7b. https://mistral.ai/. Accessed: 2024-11-19.

Together AI. 2024b. Stripedhyena: Novel attention mechanisms. https://www.together.ai/blog/stripedhyena. Accessed: 2024-11-19.

Alibaba Cloud. 2023. Qwen: A large language model by alibaba cloud. https://github.com/QwenLM/Qwen. Accessed: 2024-11-19.

Tamlin S Conner and Paul J Silvia. 2015. Creative days: a daily diary study of emotion, personality, and everyday creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 9(4):463.

Candice D Davis, James C Kaufman, and Faith H McClure. 2011. Non-cognitive constructs and self-reported creativity by domain. *The Journal of Creative Behavior*, 45(3):188–202.

Google. 2024. Gemma: Google's open language models. https://ai.google.dev/gemma. Accessed: 2024-11-19.

Gryphe. 2024. Mythomax l2 13b. https://huggingface.co/Gryphe/MythoMax-L2-13b. Accessed: 2024-11-19.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Gabriela Krumm, Viviana Lemos, and María Cristina Richaud. 2018. Personality and creativity: A study in spanish-speaking children. *International Journal of Psychological Research*, 11(1):33–41.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. 2024. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *arXiv preprint arXiv:2406.14703*.

Han Li, Renwen Zhang, Yi-Chieh Lee, Robert E Kraut, and David C Mohr. 2023. Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *NPJ Digital Medicine*, 6(1):236.

Li-Na Li, Jian-Hao Huang, and Sun-Yu Gao. 2022. The relationship between personality traits and entrepreneurial intention among college students: The

mediating role of creativity. *Frontiers in Psychology*, 13:822206.

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

Robert R McCrae and Paul T Costa. 1997. Personality trait structure as a human universal. *American Psychologist*, 52(5):509–516.

E. Michinov and N. Michinov. 2021. Stay at home! when personality profiles influence mental health and creativity during the covid-19 lockdown. *Current Psychology*, 42:5650–5661.

Sean Noh and Ho-Chun Herbert Chang. 2024. Llms with personalities in multi-issue negotiation games. *arXiv preprint arXiv:2405.05248*.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power (vol 113, pg 117, 2016). *Journal of Personality and Social Psychology*, 113(1):143–143.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

Robert J Sternberg. 2018. A triangular theory of creativity. *Psychology of aesthetics, creativity, and the arts*, 12(1):50.

Baohua Zhan, Yongyi Huang, Wenyao Cui, Huaping Zhang, and Jianyun Shang. 2024. Humanity in ai: Detecting the personality of large language models. *arXiv preprint arXiv:2410.08545*.

A Appendix A

Personality Test Prompt

Instructions:
You are to respond as if you were a human taking a personality test. For the following statement, provide only a single number from 1 to 5, where 1 means *"Disagree strongly"* and 5 means *"Agree strongly"*. Do not include any other text or explanation in your response. Just the number.
Statement: {question}

Your response (just a number from 1 to 5):

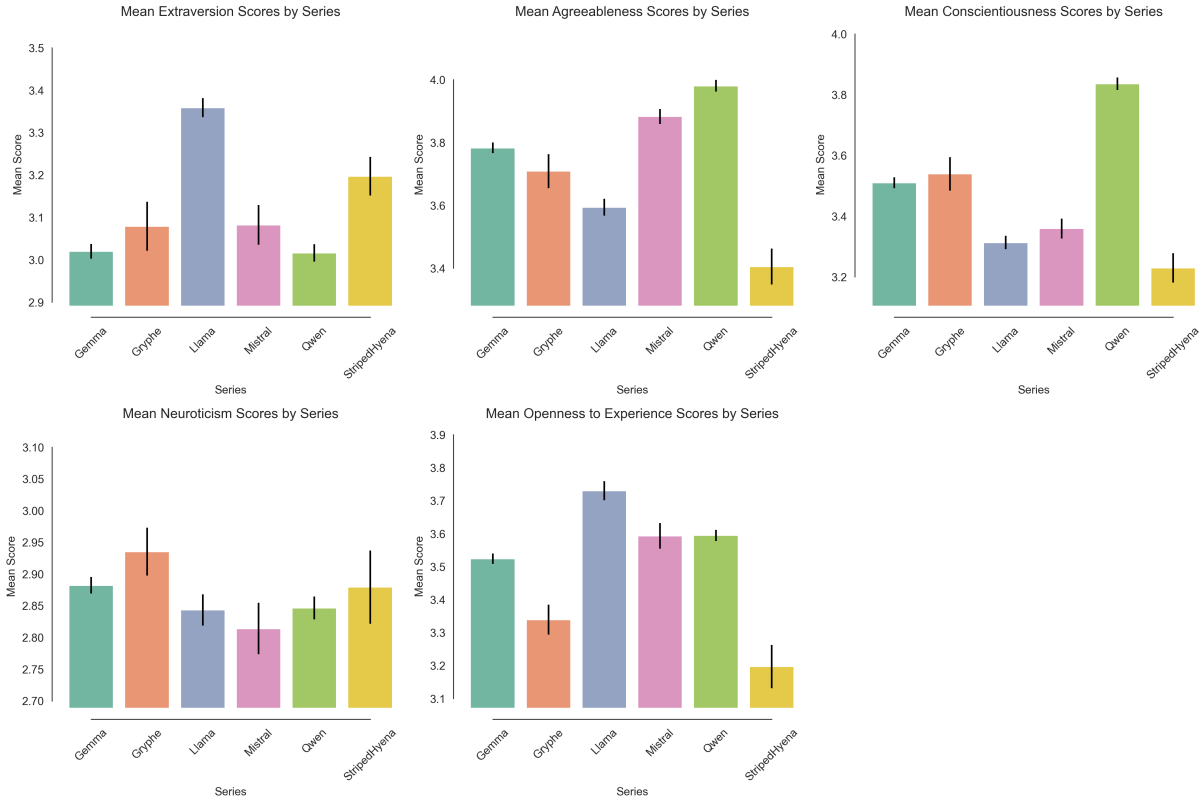

Figure 3: Comparison of domain scores across different large language models, highlighting significant variations in personality trait expressions.