A Survey of Personality Customization Techniques for Large Language Models: Methods, Implementation, and Applications

Cheng Jing
Tianjin No.1 High School, Tianjin, 300051, China
18522951028@163.com

Abstract. While the large language models (LLMs) are developing rapidly, artificial intelligence systems have begun to present human-like personality features. Consequently, "to characterize an AI system proactively" has emerged as a popular topic, capturing public interest. This technology aims to pre-set a stable personality for AI, so distinct characters AI systems can be precisely applied in diverse particular scenarios. This survey provides a comprehensive overview for this technique, consisting of the technical underpinnings, core methods, evaluation systems, application scenarios, and future challenges of AI personality configuration. This study first dissects the underlying technical pillars, including the Transformer architecture, the pre-training and fine-tuning paradigms, and Parameter-Efficient Fine-Tuning (PEFT) techniques such as QLoRA. Subsequently, this study systematically categorizes and analyzes the core technical pathways for personality shaping, covering data-driven methods, parameter control methods, and memory-augmented architectures. The paper further introduces an evaluation framework integrating psychometric scales and long-term memory validation, and explores diverse application scenarios ranging from various professional fields to entertainment and social interaction. Finally, this study discusses existing technical barriers and high-potential future research frontiers, such as cognitive architecture innovation, dynamic trait adjustment, multimodal personality extension and ethical considerations. This survey aims to provide a comprehensive analysis and a clear reference for subsequent research and practice in the field of AI personality configuration.

Keywords: Large Language Models; AI Personality; Parameter-Efficient Fine-Tuning (PEFT); Persona Vectors; Memory Augmentation.

1. Introduction

Modern large language models, with their billions or even hundreds of billions of parameters and massive training corpora, demonstrate remarkable capabilities in language understanding and generation. Beyond mere functional proficiency, these models increasingly exhibit human-like personality traits and emotional expressiveness in interactions with humans [1]. This trend marks the shift in the role of AI from a pure tool to a personalized companion. This "personification" enables AI to provide more authentic and engaging experiences in various domains requiring highly natural interaction, such as customer service, educational tutoring, psychological companionship, and entertainment.

To make AI's outputs more authentic, natural, and aligned with human emotional needs, endowing AI with a specific and stable personality is suggested to be an effective strategy. Therefore, "AI personality customization" has become a hotspot in current research. Current mainstream research primarily focuses on three dimensions: how the foundational model architecture carries personality traits, how fine-tuning techniques implant specific personalities, and how evaluation systems quantify personality consistency. Numerous relevant technologies have been developed, but due to the field's rapid evolution, a systematic and holistic summary is inevitably lacking.

This survey aims to fill this gap by providing a comprehensive overview of AI personality configuration technologies. This study will systematically analyze its technical framework (in section 2, consisted the Transformer model, Pre-training, Fine-tuning techniques), implementation paths (in section 3, consisted three core techniques for customizing AI's personality more specifically), and application challenges (in section 6), striving to deliver a panoramic analysis.

2. Technical Foundation: The Underlying Architecture and Source of Capabilities for Personality Customization

To be endowed with and stably express specific personalities, LLM has several basic technologies, which are respectively the Transformer-based model architecture and its multi-stage training pipeline. These technologies collectively form the model’s nascent personality, which enabling the model to have the ability of making basic judgement. Besides, the relationship between these techniques can be conceptualized hierarchically: the Transformer provides the fundamental computational substrate (the "hardware"), pre-training fills it with a vast repository of knowledge and linguistic patterns (the "raw material"), and fine-tuning sculpts this raw potential into a specific, stable personality (the "shaping tool").

2.1 Transformer Architecture: The Cornerstone of Personality Understanding

The Transformer architecture, particularly its self-attention mechanism, is the prerequisite for enabling personality customization. This mechanism processes all information in an input sequence in parallel and simultaneously, without overlooking distant details. This characteristic allows the Transformer to accurately capture emotional tendencies and stylistic correlations within long-range contexts. For instance, when using prompt engineering to set a personality, the Transformer analyzes all instructions simultaneously, enabling it to understand and execute the role-setting more completely and accurately than sequential models like RNNs. During user interaction, it does not forget key information from the beginning of the conversation, ensuring outputs consistently align with the preset personality tone. It can be said that the self-attention mechanism is the foundation upon which all personality-shaping techniques can stably take effect. Today, Transformer models can support context windows of up to 128K tokens or more, while research like LongNet aims to extend context length to billions of tokens [2], offering the potential for models to maintain highly consistent personality expression across conversations.

2.2 Pre-training: The "Raw Material Library" for Personality

During the pre-training phase, the model learns the statistical patterns of language and world knowledge from trillions of tokens of massive text corpora. While this process does not directly shape a specific personality, it determines the bases and tendencies of the model's "raw personality." The source, quality, and cultural background of the training data implicitly instill basic values and language style preferences into the model. Pre-training can be seen as constructing a vast and rich "personality material library" for the model, equipping it with the potential to play various roles, but its personality at this stage is chaotic and unstable.

2.3 Fine-tuning Techniques: The Fine Sculpting of Personality

The fine-tuning stage, particularly Instruction Tuning and Reinforcement Learning from Human Feedback (RLHF), is key to transitioning the model from a "generalist" to a "specialist" and forming a stable personality. Through RLHF, a reward model guides the generated content to better align with human preferences; for example, penalizing aggressive replies to enhance the model's "Agreeableness" [3]. Furthermore, the advent of Parameter-Efficient Fine-Tuning (PEFT) techniques, such as QLoRA, has significantly lowered the barrier to personality customization. QLoRA, utilizing quantization and low-rank adaptation technologies, makes it possible to fine-tune a 70-billion-parameter model on a single RTX 4090 (24GB) graphics card [4][5], providing researchers with the technical means to experiment with multiple personality prototypes quickly and at low cost.

3. Core Techniques for Personality Shaping: From Foundational Capabilities to Directed Control

After establishing the aforementioned foundational capabilities, researchers have developed a series of more targeted core techniques for personality shaping. And this study lists three prominent technologies in the following, which are respectively the Data-Driven methods, **Parameter Control Methods, and** the **Memory Augmentation.** These methods aim to endow and control AI personality more precisely, stably, and persistently.

3.1 Data-Driven Methods: Directed Cultivation Based on Personality Corpora

This paragraph will talk about the Data-Driven methods, whose core is to fine-tune the model using datasets labeled with personality tags, thereby "teaching" the model specific expression patterns. For instance, training a model on dialogue data annotated with MBTI personality types (e.g., INTP, ESFJ) [6] can enable it to exhibit thinking patterns and language styles highly consistent with that personality type during conversations. This method yields stable results but is highly dependent on large-scale, high-quality personality-annotated data.

3.2 Parameter Control Methods: The "Personality Dials" Within the Model

Currently, parameter control methods serve as a key technology for enhancing system performance. These techniques do not rely on retraining but directly intervene in the model's forward inference process, enabling real-time, fine-grained control over personality.

The Persona Vectors may be introduced first. Research from Anthropic discovered that sparse neural circuits within LLMs control different personality traits [7]. Through comparative analysis, vectors representing specific tendencies (e.g., "Extraversion," "Openness") can be extracted. By injecting this vector into the activations of the Feedforward Network (FFN) during answer generation, specific personality traits can be enhanced or weakened in real-time, akin to turning a dial.

And the **second one is the Reinforcement Learning Control.** Traditional RLHF might compromise other model capabilities when optimizing for a single objective. Advanced algorithms like GFPO (Group Filtered Policy Optimization) proposed by Microsoft [8], by introducing more complex filtering mechanisms during policy optimization, can better maintain the stability and diversity of model performance while reinforcing specific personalities.

3.3 Memory Augmentation: Maintaining Long-Term Personality Consistency

Table 1. A comparison between three core techniques

Technique	Advantages	Disadvantages	Typical Applications	Relative Cost
Data-Driven	Stable, high-quality results; well-understood process.	Requires large, annotated datasets; risk of overfitting; less flexible.	Creating predefined character personas (e.g., customer service agents).	High (Data collection & training)
Parameter Control	Highly flexible; real-time adjustment; no retraining needed.	Can be less stable; requires deep model understanding; may affect coherence.	Dynamic personality adjustment in chatbots & interactive NPCs.	Low (Post-training)
Memory Augmentation	Excellent long-term consistency; decouples memory from model params.	Introduces system complexity; requires efficient retrieval; external storage.	Dynamic personality adjustment in chatbots & interactive NPCs.	Medium (System design & storage)

The third one, Memory Augmentation, a technique remarkably improving AI personality's persistency, enabling AI maintain stable persona during long-term interactions with users. This method establishes an external memory bank for each user or dialogue session, continuously recording key personality settings, user preferences, and important historical interactions [9]. During each conversation, the model retrieves relevant memories as context, ensuring its behavior remains consistent with the long-term personality setting, overcoming the inherent context length limitation of the model itself. For example, Macaron AI's Agentic Memory technology is a typical practice in this area.

And to offer a comprehensive reference for these three core techniques described above, here's the Table 1 illustrates them in four aspects (the advantages, disadvantages, typical applications and relative cost).

4. Evaluation System and Verification Framework

Scientifically evaluating the consistency and authenticity of an AI's personality requires integrating psychological scales with computational objective metrics, forming a multi-dimensional verification framework.

4.1 Psychometric Scales

Adapted traditional personality psychology scales are used to evaluate AI models. For instance, the Language Model Personality Assessment (LMLPA) system [10] uses specifically designed prompts to guide the model through a series of situational questions and answers. The responses are then scored against the "Big Five" personality model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), enabling quantitative analysis of the AI's personality.

4.2 Long-Term Memory and Consistency Verification

Assessing the stability of personality over long-term interactions is crucial. The Forgetting Curve Test [11] evaluates the coherence of personality expression by querying the model about personality preferences or key facts set in earlier conversations after different time intervals, thereby testing its long-term memory retention rate.

5. Application Scenarios and Typical Cases

5.1 Professional Roles in Vertical Domains

In professional service scenarios, personality customization aims to enhance service efficiency and user experience. For example, bank customer service systems employ "meticulous" AIs to reduce hallucination rates and ensure reply accuracy; whereas medical consultation assistants adopt "empathetic" personalities, using a gentle tone to enhance user trust and comfort. The "Automatic/Fast/Deep Thinking" three-mode provided by GPT-5 Enterprise Edition essentially involves switching different "work personalities" for different task needs [12].

5.2 Entertainment and Social Applications

In gaming and virtual social interactions, personality customization is core to creating immersion. Research indicates that analyzing players' in-game spatial customization behaviors (like building layout and decoration style) can infer their personality traits, thereby providing a theoretical basis for personalized interaction design in social games [13]. Further, research has proposed frameworks driven by biosensor data, questionnaires, and gameplay behavior data to adapt the personality of Non-Player Characters (NPCs), realizing "companion NPCs" that can grow with the player, offering a technical prototype for "lifelong companion AI agents" [14]

6. Challenges and Future Directions

6.1 Existing Technical Bottlenecks

First of all, the personality conflict and the lack of decision weighting is consistently a big barrier for current AI development. When multiple personality traits are reinforced simultaneously, the model may fall into internal contradictions, lacking the ability to prioritize different traits in specific contexts.

The second problem is the high computational cost. Maintaining long-term personality memories for large-scale users requires significant storage and computational overhead. For instance, sustaining high-quality long-term memory for 1 million users might require the computational power of a 500+ H100 GPU cluster.

6.2 Frontier Exploration Directions

There are several frontier exploration directions, aiming to develop the forms of AI's personalities expressing**.**

The first one is the c**ognitive architecture innovation for current AI.** Future research will focus on decoupling personality from commonsense/context, avoiding inappropriate behaviors like an "optimistic AI still smiling in a tragic scene," and achieving more intelligent context-aware personality expression.

Additionally, **AI will do dynamic trait adjustment itself in the future.** Explore the possibility of adjusting AI personality modes in real-time and smoothly based on external environments (e.g., circadian rhythms) or user states, making interactions more adaptive and natural.

What's more, multimodal personality extension for AI could be expected. When AI possesses a physical embodiment (e.g., robots, virtual avatars), its personality expression extends from pure text to multimodal channels like voice, facial expressions, and gestures [15]. An "optimistic" AI would not only use positive text but also have a livelier vocal tone and richer facial expressions. This multimodal fusion of personality expression can significantly enhance the realism and immersion of interaction, representing an important development direction for personality customization technology in the field of embodied intelligence.

Safety and ethical framework are also being focused on nowadays. One of them is the **Personality Firewall,** establishing mechanisms to block dangerous combinations of personality traits (e.g., "high aggressiveness" with "high persuasiveness"). In addition, **transparency labeling** is being requested and conducted currently. Following proposals from bodies like the EU, clearly label the personality mode currently in use (e.g., "ENFJ") during AI interactions to ensure user awareness.

The scientific of Evaluation for AI has high request. Firstly, physiological signal feedback is suggested to be a useful technology for more accurate evaluation in the future. Objectively assess the realism brought by AI affinity by monitoring users' physiological indicators, such as EEG alpha wave fluctuations, during interaction with AI. Secondly, **multimodal Fusion Evaluation could also be used to complete this task.** Combine voice tremor detection, facial micro-expression blood flow analysis, etc., to construct a three-dimensional, comprehensive evaluation system for personality performance.

7. Conclusion

This survey has systematically examined the burgeoning field of personality customization for large language models (LLMs), a research area of shifting AI's positions from pure functional tools to personalized interactive partners. While AI system is endowed with a specific personality, those interactions between AI and human will be more engaged and authentic. Simultaneously, AI can be better applied in diverse domains such as professional services, entertainment, and social companionship.

To comprehensively summarize the technology of endowing AI with characters, this study traced the technical foundation of it first, beginning with the foundational Transformer architecture and its self-attention mechanism, which provides the essential substrate for contextual understanding and trait stability. The review detailed how pre-training constructs a vast "personality material library," while fine-tuning techniques, particularly Parameter-Efficient Fine-Tuning (PEFT) like QLoRA, enable efficient and accessible personality sculpting. Beyond these foundations, this essay categorized and analyzed core personality-shaping techniques into three main pathways: data-driven methods for stable persona implantation, parameter control methods like persona vectors for real-time trait adjustment, and memory-augmented architectures for ensuring long-term personality consistency. Furthermore, the survey introduced a multidimensional evaluation framework combining psychometric scales and long-term consistency verification, and explored concrete applications alongside persistent challenges.

The implications of these developments are profound. They collectively enable the creation of AI systems capable of serving not merely as tools but as consistent, context-aware interaction partners. This enhances user experience, trust, and effectiveness in applications ranging from empathetic healthcare assistants to immersive game characters. However, this survey also highlights several limitations of the current research landscape. These include a scarcity of non-English, large-scale personality-annotated corpora, immature mechanisms for resolving internal conflicts between multiple reinforced traits, and the substantial computational overhead associated with maintaining long-term personality memories for large user bases.

These limitations delineate clear directions for future work. The core challenge evolves from the technical question of "how to build" a personality to the more nuanced issues of "what kind of personality to build" and ensuring it is safe, ethical, and dynamically appropriate to context. Addressing this will require sustained interdisciplinary collaboration. Future research must focus on developing more sophisticated cognitive architectures, enabling dynamic and multimodal personality expression, establishing robust ethical safeguards like personality firewalls, and creating more scientific, objective evaluation systems. All in all, enforcing those aspects above collectively will ultimately guide the responsible and effective development of personalized AI.

References

[1] Zhao, Y., Huang, Z., Seligman, M. et al. Risk and prosocial behavioural cues elicit human-like response patterns from AI chatbots. Sci Rep 14, 7095 (2024). https://doi.org/10.1038/s41598-024-55949-y.

[2] Ding, J., Ma, S., Dong, L., Zhang, X., Huang, S., Wang, W., & Wei, F. (2023). LongNet: Scaling Transformers to 1,000,000,000 Tokens. arXiv. https://doi.org/10.48550/arXiv.2307.02486.

[3] Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … Kaplan, J. (2022). Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv. https://doi.org/10.48550/arxiv.2204.05862.

[4] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. arXiv. https://doi.org/10.48550/arXiv.2305.14314.

[5] Answer.AI. (2023). *FSDP-QLoRA: Scaling QLoRA to 70B Models on Consumer GPUs*. GitHub repository. https://github.com/answer-ai/fsdp-qlora.

[6] Tu, Q., Chen, C., Li, J., Li, Y., Zhou, Y., & Huang, X. (2023). CharacterChat: Learning towards Conversational AI with Personalized Social Support. arXiv. https://doi.org/10.48550/arXiv.2308.10278.

[7] Chen, R., Arditi, A., et al. (2025). Persona Vectors: Monitoring and Controlling Character Traits in Language Models. arXiv. https://doi.org/10.48550/arXiv.2501.21509.

[8] Shrivastava, V., Awadallah, A. H., Balachandran, V., & Zhou, Y. (2025). Sample More to Think Less: Group Filtered Policy Optimization for Concise Reasoning. arXiv. https:// doi.org/ 10. 48550/ arXiv. 2508. 09726.

[9] Zhang, Y., Li, Q., Wang, C., & Yu, D. (2025). Memo: Building Production-Ready AI Agents with Scalable Long-Term Memory. arXiv. https://doi.org/10.48550/arXiv.2504.19413.

[10] Zheng, et al. (2024). Language Model Linguistic Personality Assessment (LMLPA): A Big Five Personality Quantification System Based on Language Models. arXiv. https:// doi.org/ 10. 48550/ arXiv. 2410.17632.

[11] Zheng, et al. (2023). MemoryBank: Enhancing Large Language Models with Long-Term Memory. arXiv. https://doi.org/10.48550/arXiv.2305.10250.

[12] Dai, M., Yuan, C., & Nie, X. (2024). Managing the Personality of NPCs with Your Interactions: A Game Design System Based on Large Language Models. Lecture Notes in Computer Science, 13888, 245-256. https://doi.org/10.1007/978-3-031-60692-2_17.

[13] Wohn, D. Y., & Wash, R. (2012). A virtual "Room" with a cue: Detecting personality through spatial customization in a city simulation game. Computers in Human Behavior, 29(1), 155-159. https: // doi. org/ 10.1016/j.chb.2012.07.029.

[14] Pretty, E. J., Fayek, H. M., & Zambetta, F. (2023). A Case for Personalized Non-Player Character Companion Design. International Journal of Human–Computer Interaction, 40(12) 3051-3070. https: //doi. org/ 10.1080/10447318.2023.2181125.

[15] Sonlu, S., Güdükbay, U., & Durupinar, F. (2021). A Conversational Agent Framework with Multi-modal Personality Expression. ACM Transactions on Graphics (TOG), *40*(2), Article 15. https:// doi. org/ 10.1145/3439795.