

# The Geometry of Persona: Disentangling Personality from Reasoning in Large Language Models

Zhixiang Wang<sup>1</sup>

<sup>1</sup> Precision and Intelligence Medical Imaging Lab, Beijing Friendship Hospital, Capital Medical University

December 9, 2025

## Abstract

**Background:** The deployment of personalized Large Language Models (LLMs) is currently constrained by the *stability-plasticity dilemma*. Prevailing alignment methods, such as Supervised Fine-Tuning (SFT), rely on stochastic weight updates that often incur an "alignment tax"—degrading general reasoning capabilities.

**Methods:** We propose the *Soul Engine*, a framework based on the **Linear Representation Hypothesis**, which posits that personality traits exist as orthogonal linear subspaces. We introduce **SoulBench**, a dataset constructed via *dynamic contextual sampling* ( $C(N, k)$ ). Using a dual-head architecture on a frozen Qwen-2.5 base, we extract disentangled personality vectors without modifying the backbone weights.

**Results:** Our experiments demonstrate three breakthroughs. First, **High-Precision Profiling:** The model achieves a Mean Squared Error (MSE) of **0.011** against psychological ground truth. Second, **Geometric Orthogonality:** T-SNE visualization confirms that personality manifolds are distinct and continuous, allowing for "Zero-Shot Personality Injection" that maintains original model intelligence. Third, **Deterministic Steering:** We achieve robust control over behavior via vector arithmetic (e.g.,  $\vec{v}_{Neutral} + \alpha \cdot \vec{v}_{Villain}$ ), validated through extensive ablation studies.

**Conclusion:** This work challenges the necessity of fine-tuning for personalization. By transitioning from probabilistic prompting to deterministic latent intervention, we provide a mathematically rigorous foundation for safe, controllable AI personalization.

## 1 Introduction

The evolution of Large Language Models (LLMs) is shifting from the pursuit of general-purpose reasoning to the creation of specialized, coherent agents [1, 2]. Whether for immersive role-playing in open-world environments [3] or empathetic engagement in therapeutic settings, the utility of an AI agent increasingly depends on its ability to maintain a stable, distinct psychological profile. However, achieving this "personality alignment" without degrading the model's core intelligence remains one of the field's most persistent challenges.

**The Stability-Plasticity Dilemma.** Current paradigms for steering LLM behavior are trapped in a trade-off between stability and capability. The dominant approach, **Supervised Fine-Tuning (SFT)** and its parameter-efficient variants like LoRA [4], treats personality as a distribution of tokens to be learned via gradient descent. While effective for short-term style mimicry, this method is fundamentally destructive. By updating the model's weights to fit a narrow