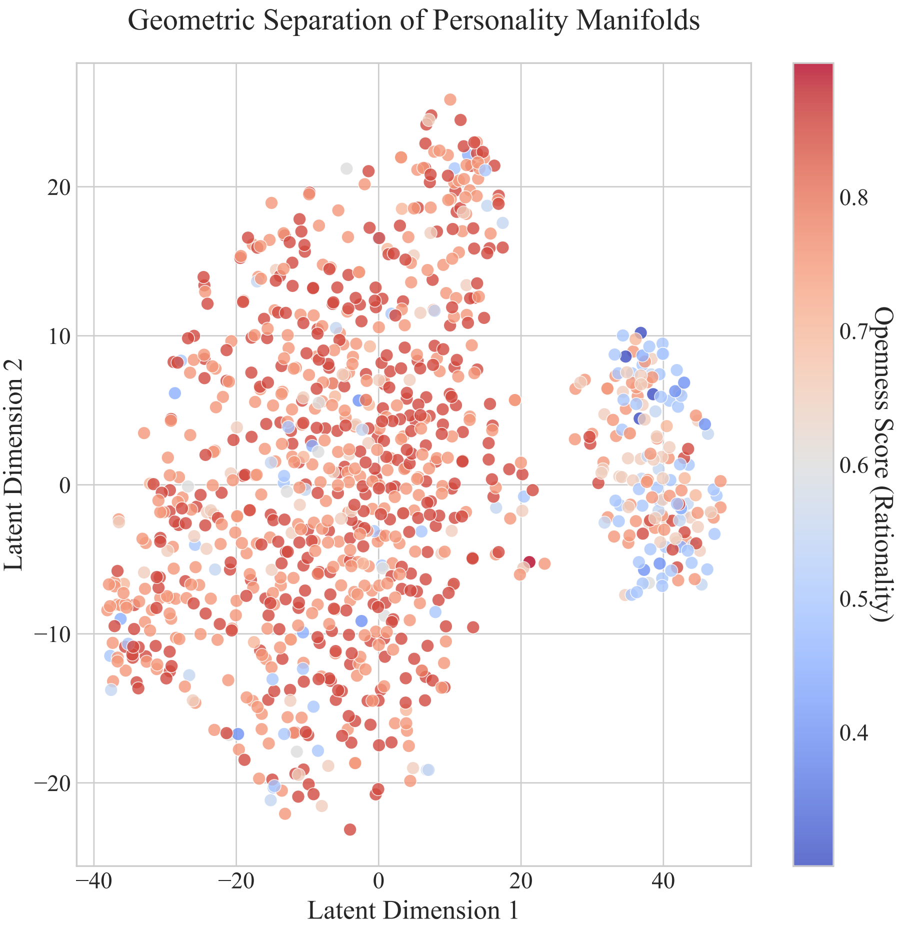

Figure 2: **Geometric Separation of Personality Manifolds.** T-SNE projection of 1,000 character embeddings. Points are colored by their "Openness" score. The clear gradient separation confirms that the Soul Encoder has successfully mapped discrete psychological traits onto a continuous geometric manifold.

3.4 Ablation Study: Steering Stability

Finally, we evaluate the effectiveness of our vector injection mechanism. We perform a grid search to find the optimal intervention layer and strength. Figure 3 illustrates the trade-off between "Villainy" (Steering Effectiveness) and "Sanity" (Model Coherence).

We observe that injecting vectors into the middle layers yields the most robust control. Early-layer injection fails to influence high-level semantics effectively, while late-layer injection tends to disrupt syntax generation.

4 Discussion

Our findings on the Qwen2.5-0.5B model provide compelling empirical support for the *Linear Representation Hypothesis* in the domain of computational psychology. Beyond the quantitative metrics, several key implications emerge regarding the nature of personality in Large Language Models.