This term is crucial for achieving "disentangled control," allowing us to adjust one personality trait without accidentally altering others.

2.4 Inference: Deterministic Steering via Vector Arithmetic

Unlike SFT, which permanently alters weights, Soul Engine enables dynamic, plug-and-play personality injection. We define the **Steering Vector** $\vec{v}_{steer}$ as the difference between a target persona and the neutral mean:

$$\vec{v}_{steer} = \mathbb{E}[e_{Target}] - \mathbb{E}[e_{Neutral}] \tag{7}$$

During inference, we intervene in the residual stream of layer $L-1$. The modified hidden state h' is computed as:

$$h' = h + \alpha \cdot \frac{\vec{v}_{steer}}{\|\vec{v}_{steer}\|} \tag{8}$$

Where α is the steering coefficient. Since $\vec{v}_{steer}$ is derived from the orthogonal subspace learned by our encoder, this intervention shifts the output distribution towards the target persona style without disrupting the logical coherence of the generated tokens.

3 Experiments

3.1 Experimental Setup

We conduct our analysis on **Qwen2.5-0.5B-Instruct**. Training was performed on a single NVIDIA A100 GPU (fp16).

- **Dataset:** SoulBench (Validation split: 1,000 samples).
- **Hyperparameters:** Batch Size $= 16$, Learning Rate $= 1e-4$.

3.2 Quantitative Results: Psychometric Precision

The Scientific Soul Encoder achieves rapid convergence. As shown in Table 1, the model achieves a Mean Squared Error (MSE) of **0.0113** on the held-out validation set. This implies that the learned "Psychometric Head" can predict the ground-truth OCEAN score with $\sim 99\%$ accuracy.

Table 1: Performance Metrics on SoulBench Validation Set

Metric	Value (Best)	Value (Final)
MSE (Psychometric)	**0.0113**	0.0118
Total Loss	-	1.3184

3.3 Qualitative Analysis: Geometry of Character

To verify the disentanglement of personality, we visualize the learned manifold using T-SNE on the embeddings from the Soul Encoder (Figure 2).

The visualization demonstrates that characters with similar profiles (e.g., High Openness vs. Low Openness) are naturally clustered together, confirming that the model has learned a continuous "spectrum of personality."