semantic content. The framework consists of three components: (1) **SoulBench**, a dataset constructed via combinatorial sampling; (2) The **Scientific Soul Encoder**, a dual-head probe architecture; and (3) A **Deterministic Steering** mechanism based on vector arithmetic.

2.1 SoulBench: Mining Stylistic Invariance via Dynamic Sampling

A critical challenge in personality modeling is disentangling "style" (how something is said) from "content" (what is said). Static datasets often lead models to overfit to specific semantic phrases (e.g., associating "Joker" solely with the word "Batman").

To address this, we introduce a **Dynamic Contextual Sampling** strategy. Let $\mathcal{D}_c = \{s_1, s_2, ..., s_M\}$ be the corpus of sentences for a specific persona c. During training, we do not use fixed samples. Instead, for each iteration t, we construct an anchor A_t by randomly sampling a subset of k sentences:

$$A_t = \text{Concat}(s_{i_1}, s_{i_2}, ..., s_{i_k}), \quad \text{where } \{i_1, ..., i_k\} \sim \text{Uniform}(1, M) \tag{1}$$

In our experiments, we set chunk size $k = 3$. This combinatorial approach generates a virtual dataset of size $\binom{M}{k}$, which is effectively infinite. This forces the encoder to ignore the transient semantic content of individual sentences and converge on the **stylistic invariance**—the "common denominator" of the character's voice.

Ground truth labels $\mathbf{y}_{ocean} \in [0, 1]^5$ for each character are generated using a Teacher Model (**Doubao-Seed-1.6**) [12] prompted with the full character profile, ensuring psychological consistency.

2.2 The Scientific Soul Encoder

Our architectural design is governed by the principle of *Non-Invasive Probing*. We aim to extract personality representations without disrupting the pre-trained logical circuits of the base model. We denote the base LLM as $\mathcal{F}_\theta$.

Stratified Freezing Strategy. We partition the Transformer layers into two distinct regions: the *Syntactic Foundation* (θ_{frozen}) and the *Semantic Apex* (θ_{active}). **We operate on the hypothesis that abstract personality traits crystallize in the upper strata of the network, while lower layers handle syntax and basic semantics.**

For a model with L layers, we freeze the first K layers:

$$\theta_{frozen} = \{l_0, l_1, \ldots, l_{K-1}\}, \quad \theta_{active} = \{l_K, \ldots, l_{L-1}\} \tag{2}$$

In our primary experiments with **Qwen2.5-0.5B** ($L = 24$), we set $K = 20$, fine-tuning only the final 4 layers and the normalization heads. This ensures that the deep reasoning manifolds formed during pre-training remain intact.

Dual-Head Projectors. The latent embedding $e \in \mathbb{R}^d$ (where $d = 896$) is extracted from the final hidden state and bifurcated into:

- **Identity Head (P_{id}):** A 2-layer MLP mapping $e \to z_{id} \in \mathbb{R}^{256}$ for stylistic clustering.
- **Psychometric Head (P_{psy}):** A linear probe mapping $e \to \hat{\mathbf{y}} \in \mathbb{R}^5$ for OCEAN alignment.