stylistic corpus (e.g., "speak like a pirate"), SFT frequently induces *catastrophic forgetting* of the pre-trained general knowledge [5]. This phenomenon, known as the "alignment tax" [6], results in agents that possess strong stylistic traits but suffer from degraded logical reasoning and reduced problem-solving capabilities (e.g., lower MMLU scores).

Alternatively, **In-Context Learning (ICL)** or "System Prompting" attempts to steer behavior without weight updates. However, this approach lacks determinism. LLMs are prone to "persona drift" or "catastrophic amnesia" during extended interactions, as the transient instructions in the context window are diluted by the model's inherent reinforcement learning (RLHF) priors [7]. Consequently, prompt-based agents are fragile, inconsistent, and easily "jailbroken."

The Linear Representation Hypothesis. We posit that these limitations arise from a category error: treating personality as "knowledge" to be memorized rather than a "state" to be activated. Recent breakthroughs in **Mechanistic Interpretability** and **Representation Engineering** suggest a radical alternative: the *Linear Representation Hypothesis* [8, 9]. This hypothesis suggests that high-level semantic concepts—such as sentiment, truthfulness, and potentially psychometric traits—are encoded as linear, orthogonal directions within the high-dimensional latent space of the Transformer [10]. If valid, this implies that the "soul" of the model (its personality) is geometrically distinct from its "brain" (its reasoning circuits). Therefore, steering a persona should not require global weight modification, but rather precise navigation within the existing latent manifold.

The Soul Engine. In this work, we introduce the **Soul Engine**, a framework that validates this hypothesis and mathematically disentangles personality from intelligence. Unlike the "black box" nature of SFT, our approach is geometric and deterministic. We identify the specific linear subspaces corresponding to the Big Five (OCEAN) personality traits and develop a method to manipulate them via vector arithmetic.

Our contributions are threefold:

1. **Data Engineering (SoulBench):** We address the scarcity of psychological ground truth by constructing a multi-source dataset using a novel *Dynamic Contextual Sampling* strategy ($C(N, k)$). This forces the encoder to learn invariant stylistic fingerprints rather than semantic content.
2. **Mechanistic Discovery:** Through layer-wise probing on a frozen Qwen-2.5 backbone [11], we demonstrate that personality representations emerge in the upper transformer blocks (Layers 18-24) and are largely orthogonal to reasoning vectors.
3. **Deterministic Control:** We achieve "Zero-Shot Personality Injection." By adding computed vectors to the hidden states (e.g., $\vec{v}_{Neutral} + \alpha \cdot \vec{v}_{Villain}$), we demonstrate precise control over behavior (MSE < 0.01) with negligible degradation in general intelligence benchmarks.

This work marks a paradigm shift from stochastic, destructive fine-tuning to deterministic, non-invasive latent intervention.

2 Methodology

We propose the **Soul Engine**, a framework designed to extract and manipulate the geometric representation of personality within Large Language Models. Our approach is grounded in the premise that personality is a high-level abstraction that is linearly separable from low-level