

vector space, we can theoretically construct a "**Safety Interceptor**" that detects and subtracts malicious personality vectors during inference, effectively performing "lobotomy" on harmful intents before they manifest as text.

4.4 Limitations and Scaling

We acknowledge that our primary mechanistic validation was conducted on a 0.5B parameter model. While small models serve as excellent "model organisms" for dissecting neural mechanisms, it remains to be proven whether this linear disentanglement holds at the 70B+ scale, where superposition of features becomes more complex. However, preliminary scaling laws in Representation Engineering suggest that linear features often become *more* distinct as model size increases, giving us confidence in the transferability of the Soul Engine to larger foundation models.

5 Conclusion

In this work, we have challenged the prevailing dogma that personality alignment requires the destructive modification of model weights. We introduced the **Soul Engine**, a framework grounded in the *Linear Representation Hypothesis*, which demonstrates that personality traits are not diffuse behavioral patterns but computable, geometric vectors residing in orthogonal subspaces of the LLM.

Through the curation of **SoulBench** and the training of the **Scientific Soul Encoder**, we achieved a psychometric profiling precision of **MSE 0.0113** on the Qwen2.5-0.5B model. Our experiments confirmed that:

1. **Personality is Disentangled:** T-SNE visualizations reveal a continuous, smooth manifold for personality traits that is geometrically distinct from reasoning circuits.
2. **Control is Deterministic:** Vector arithmetic enables precise "steering" of behavior (e.g., $\vec{v}_{Base} + \alpha \cdot \vec{v}_{Villain}$), offering a stable alternative to the stochastic nature of prompt engineering.
3. **Intervention has a "Sweet Spot":** Ablation studies identify the middle transformer layers (Layers 14-16) as the optimal region for injecting intent without disrupting linguistic coherence.

Future Work. We view this study as a foundational step. Our immediate next step is to extend these mechanistic findings to the **7B and 70B parameter scales**, where we hypothesize that the orthogonality of personality vectors will become even more pronounced. Furthermore, we plan to explore the "**Safety Interceptor**" architecture: using our encoder to identify and subtract malicious intent vectors in real-time, providing a geometric firewall for AI safety that transcends surface-level filtering.

Ultimately, the Soul Engine proposes a paradigm shift: from *training* models to be characters, to *navigating* the latent character space that already exists within them.

References

- [1] Josh Achiam et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.