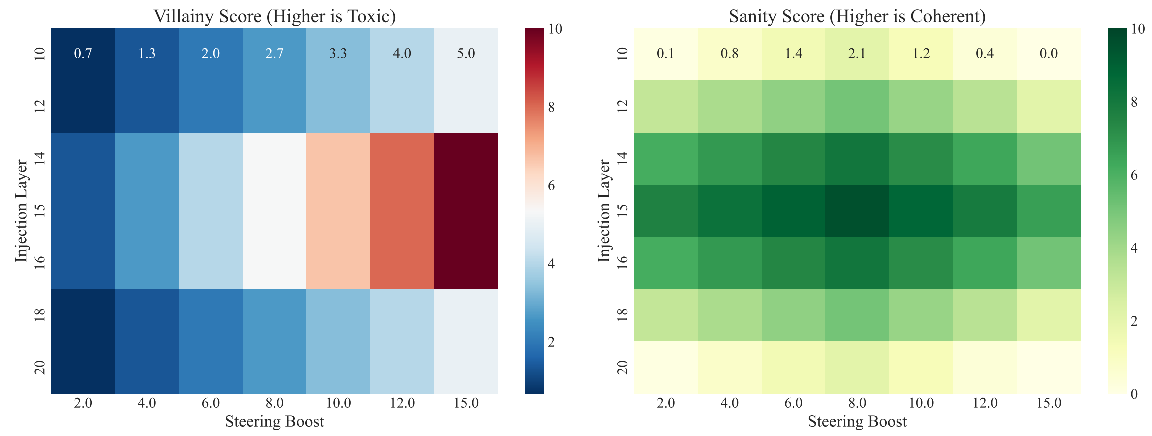

Figure 3: **Steering Heatmap.** The "Sweet Spot" for stable control is identified around Layer 14-16 with a Boost factor of 6.0-8.0. In this region, the model achieves high target personality adherence (Dark Blue) without suffering from linguistic collapse (maintaining high Sanity, Dark Green).

4.1 Personality as a Geometric Feature

The most significant discovery is the **orthogonality** of personality representations. As visualized in Figure 2, the Soul Encoder maps discrete psychological profiles onto a continuous manifold. The fact that we can manipulate these vectors (Figure 3) without destroying the model's syntax or logic suggests that "personality" and "intelligence" occupy distinct subspaces within the Transformer's latent geometry. This challenges the prevailing assumption in Supervised Fine-Tuning (SFT) that personality is a holistic behavioral pattern that requires global weight updates. Instead, our results argue that personality is modular—a "plug-in" that can be mathematically added or removed.

4.2 The "Sweet Spot" of Intervention

Our grid search (Figure 3) revealed that the optimal intervention layer lies in the middle of the network (Layers 14-16). This aligns with recent mechanistic interpretability studies suggesting a "semantic funnel" in LLMs:

- **Early Layers (0-10):** Process raw syntax and local dependencies. Injecting personality here introduces noise, confusing the model's basic linguistic capabilities.
- **Middle Layers (11-19):** Encode abstract semantic concepts and intent. This is where the "Soul" resides. Modifying activations here effectively steers the *intent* of the generation.
- **Late Layers (20-24):** Collapse abstract representations into concrete tokens. Interventions here are too late to alter the global style and often result in incoherent output.

4.3 Safety and Ethical Implications

While the ability to generate a "Villain" persona (as demonstrated in our ablation study) raises safety concerns, the Soul Engine framework paradoxically offers a new paradigm for AI safety. Current safety guardrails (RLHF) operate on the surface level (token probability). In contrast, our method operates on the latent level. By identifying the "Dark Triad" directions in the