Controllable and Explainable Personality Sliders for LLMs at Inference Time

Florian Hoppe[1,2], David Khachaturov[2], Robert Mullins[2], Mark Huasong Meng[1,3†]

[1]Technical University of Munich, [2]University of Cambridge, [3]University College Dublin

Motivation

Large Language Models (LLMs) often lack a durable, consistent persona, suffering from contextual drift in long conversations. Existing solutions like Supervised Fine-Tuning (SFT) or Direct Preference Optimization (DPO) are effective but monolithic. Training a separate model for every desired combination of personality traits leads to a combinatorial explosion.

Our Solution: Inference-time **Activation Steering** via a modular framework for Personality Alignment, structured around the Big Five (OCEAN) traits.

System Overview

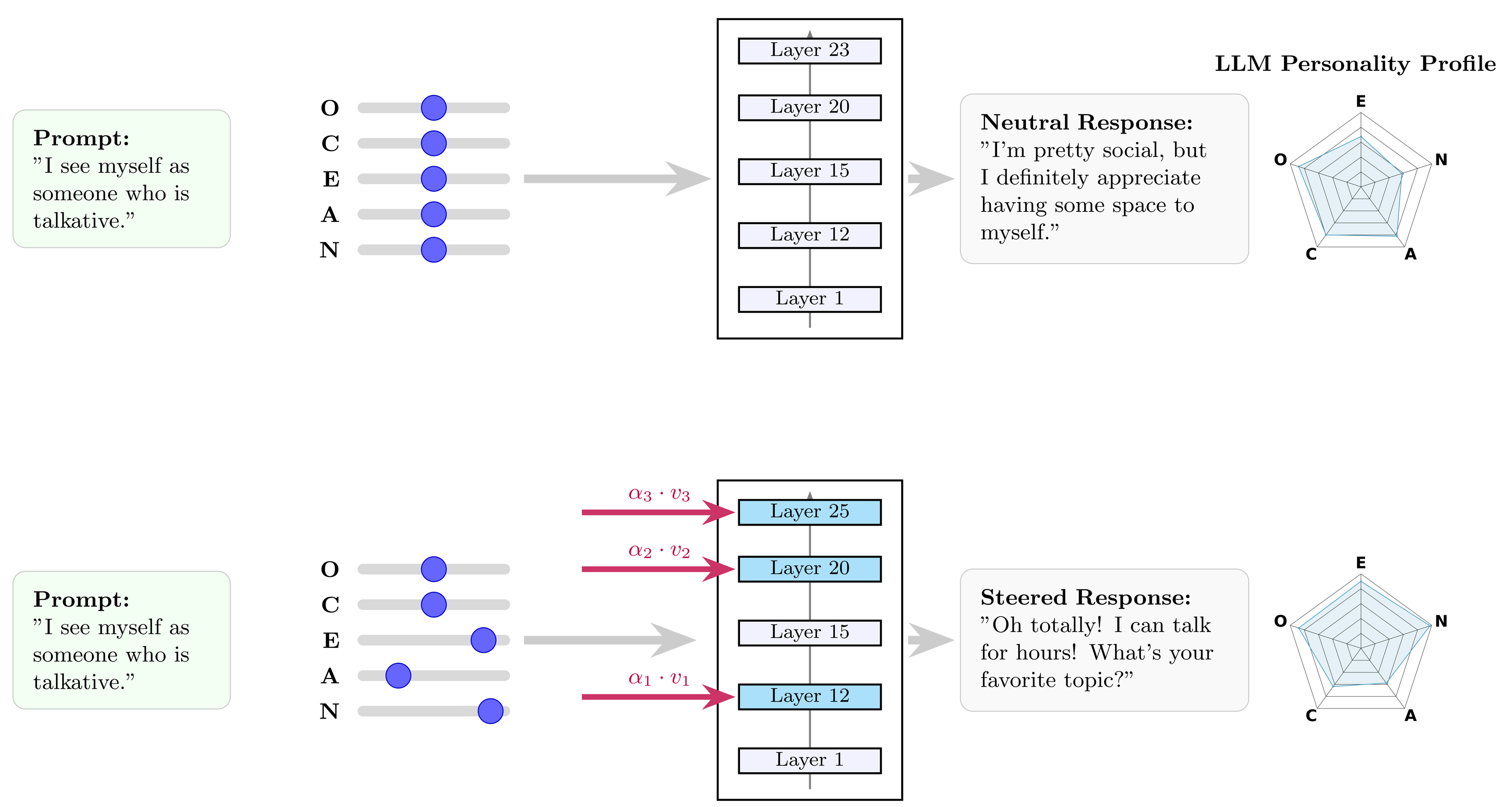


Figure: **OCEAN Sliders**: Users can dynamically adjust model personality via Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Key Idea: Sequential Adaptive Steering (SAS)

Naive multi-vector steering fails due to representation collapse: preceding interventions shift the activation manifold, causing subsequent probes to encounter unseen distributions. **Sequential Adaptive Steering (SAS)** effectively orthogonalizes steering vectors by training probes on activation distributions shifted by prior interventions, mitigating destructive interference.

Key Contributions

- **SAS Framework:** Modular, inference-time composition of personality traits without weight updates.
- **Interference Mitigation:** Novel training procedure that effectively orthogonalizes steering vectors.
- **Fisher-Ratio Selection:** Automated, data-driven identification of optimal intervention layers.
- **Cross-Model Validation:** Demonstrated Pareto dominance across Llama-3, Mistral, and Qwen architectures.

Methodology

We construct probes sequentially, ensuring that each new probe is robust to the shifts induced by prior probes. During training, we sample the steering intensity α_{train} uniformly to simulate distributional shifts. This exposure forces the probe to learn a direction invariant to these perturbations.

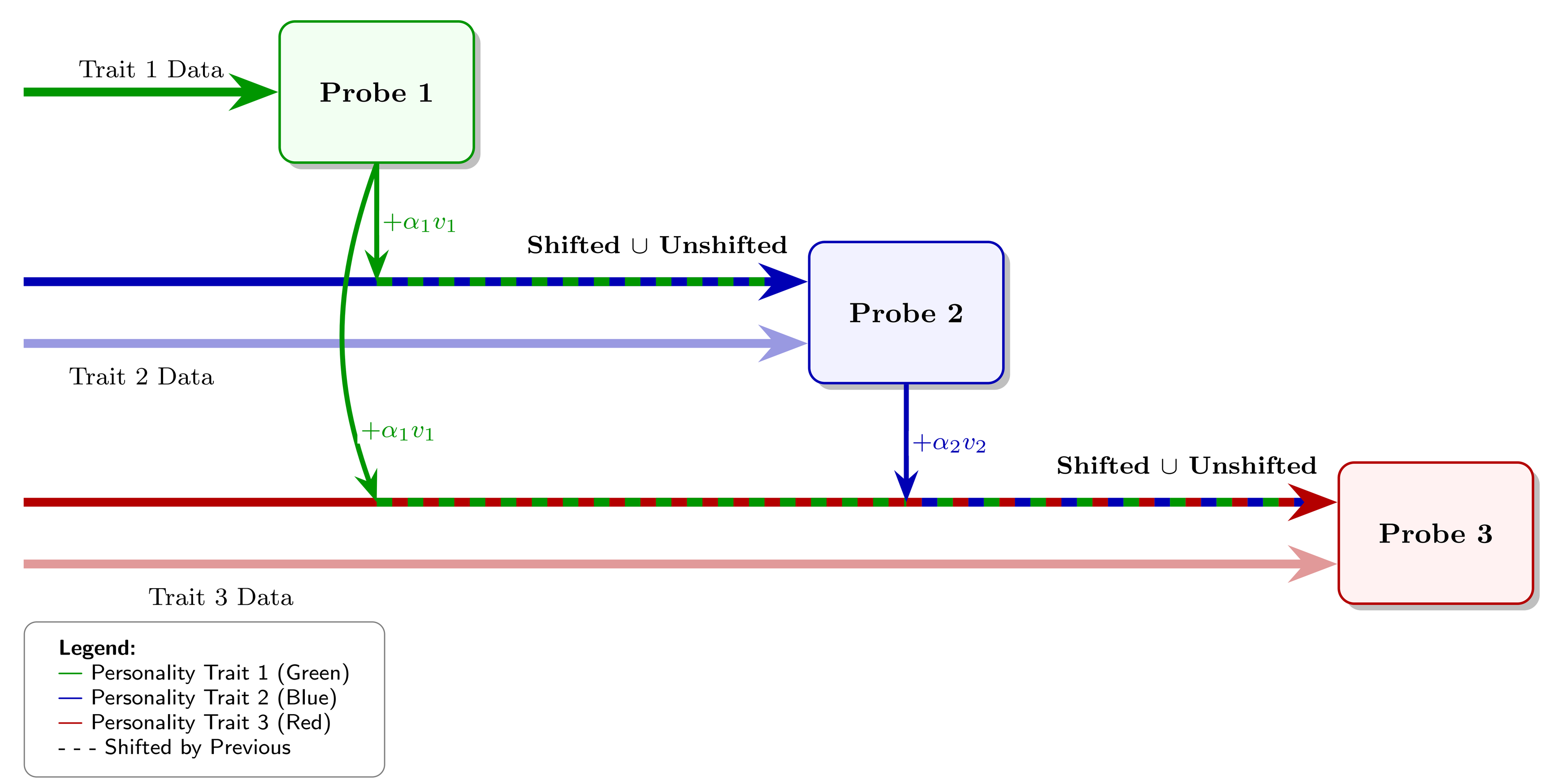


Figure: **SAS Visualization:** Each subsequent probe is trained on a combined dataset of unsteered activations and activations steered by predecessor probes.

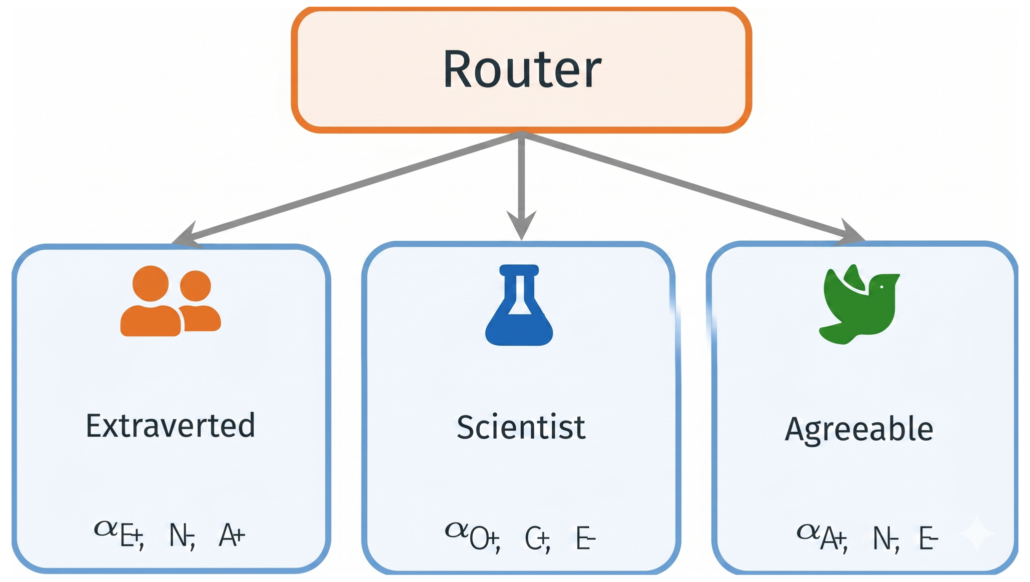
Automated Layer Selection & Calibration

Selecting the optimal intervention layer l^* is critical. We automate this using the **Fisher Ratio (FR)** to measure the separability of class distributions, ensuring we intervene where traits are most disentangled.

We also define safe operating bounds $[\alpha_{min}, \alpha_{max}]$ via grid search to guarantee stable text generation without perplexity degradation.

Future Direction: AI agents and Mixture of Personalities

Moving towards autonomous AI agents, we envision a **"Mixture of Personalities"** architecture. Instead of monolithic models, agents can dynamically route to specific, compositionally steered personality modules based on task context and user needs.



Empirical Results

Our framework achieves Pareto dominance over baselines in the trade-off between goal adherence and model perplexity across Llama-3, Mistral, and Qwen architectures.

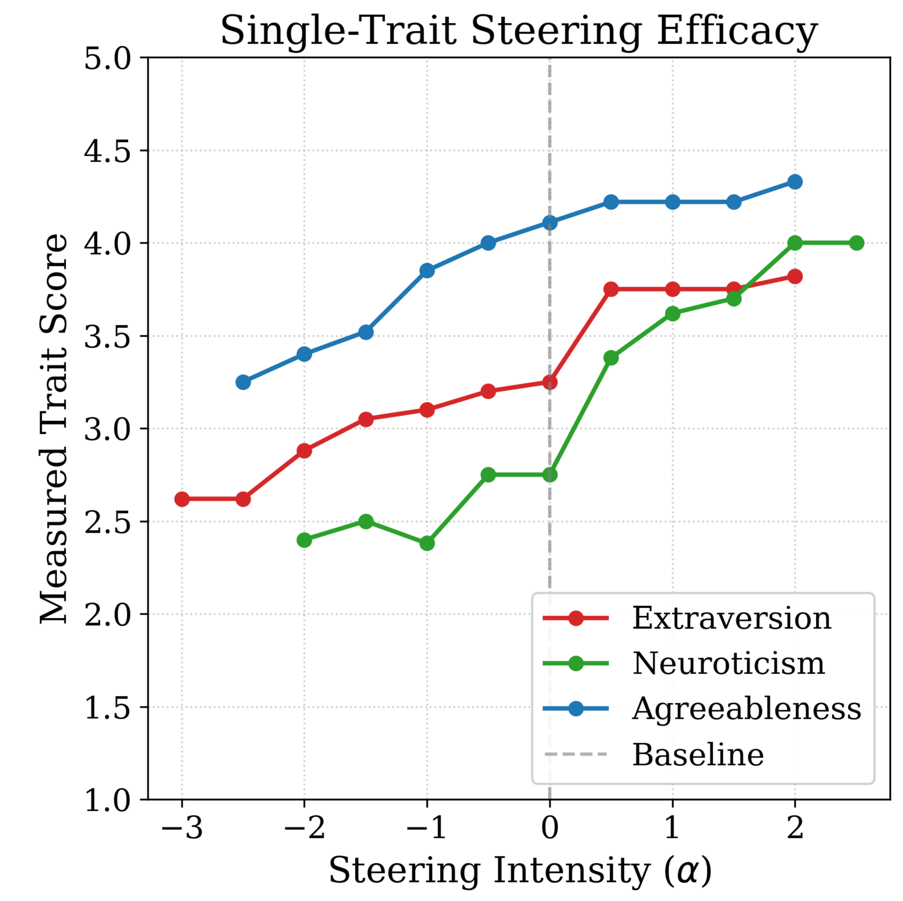


Figure: Goal adherence vs. perplexity trade-off during a single probe sweep.

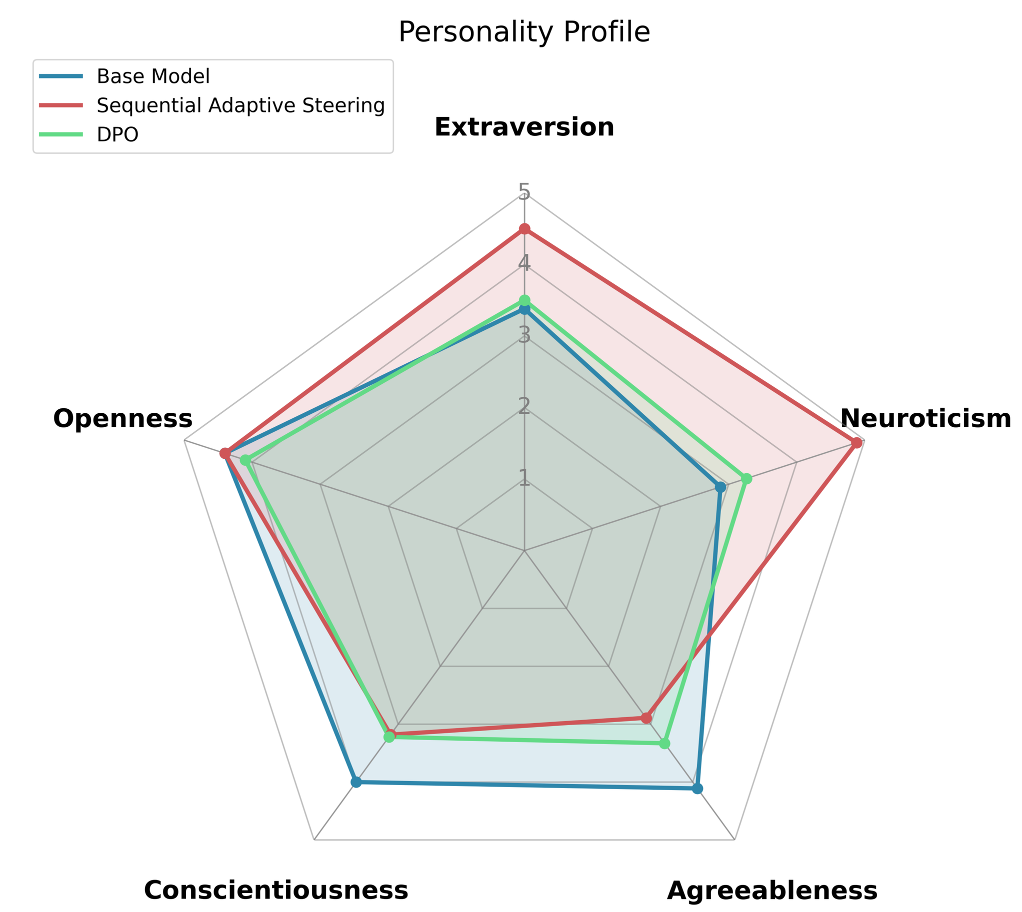


Figure: Radar plot demonstrating independent steering of multiple traits without interference.

Conclusion

Sequential Adaptive Steering (SAS) is a parameter-efficient framework that mitigates destructive interference in multi-vector steering. By accounting for geometric shifts from preceding interventions, our approach enables precise, simultaneous control over Big Five traits while maintaining model stability, offering a highly modular alternative to costly fine-tuning.

We gratefully acknowledge primary travel support from G-Research, alongside funding from the University of Cambridge, TUM, and the Alexander von Humboldt-Stiftung.