Personality Manipulation as a Cognitive Probe in Large Language Models

Gunmay Handa[2], Zekun Wu[1,2], Adriano Koshiyama[1,2], Philip Treleaven[2]
[1]Holistic AI [2]University College London

Abstract

Personality manipulation in large language models (LLMs) is increasingly applied in customer service and agentic scenarios, yet its mechanisms and trade-offs remain unclear. We present a systematic study of personality control using the Big Five traits, comparing in-context learning (ICL), parameter-efficient fine-tuning (PEFT), and mechanistic steering (MS). Our contributions are fourfold. First, we construct a contrastive dataset with balanced high/low trait responses, enabling effective steering vector computation and fair cross-method evaluation. Second, we introduce a unified evaluation framework based on within-run Δ analysis that disentangles, reasoning capability, agent performance, and demographic bias across MMLU, GAIA, and BBQ benchmarks. Third, we develop trait purification techniques to separate openness from conscientiousness, addressing representational overlap in trait encoding. Fourth, we propose a three-level stability framework that quantifies method-, trait-, and combination-level robustness, offering practical guidance under deployment constraints. Experiments on Gemma-2-2B-IT and LLaMA-3-8B-Instruct reveal clear trade-offs: ICL achieves strong alignment with minimal capability loss, PEFT delivers the highest alignment at the cost of degraded task performance, and MS provides lightweight runtime control with competitive effectiveness. Trait-level analysis shows openness as uniquely challenging across methods and personality encoding consolidating around intermediate layers. Taken together, these results establish personality manipulation as a multi-level probe into behavioral representation, linking surface conditioning, parameter encoding, and activation-level steering, and positioning mechanistic steering as a lightweight alternative to fine-tuning for both deployment and interpretability.

1 Introduction and Related Work

Personality manipulation in large language models (LLMs) is increasingly common, particularly in customer service and agentic scenarios, yet the trade-offs between personality control and task capability remain underexplored. In this work, we focus on the Big Five personality traits as a systematic framework for studying how behavioral characteristics are encoded and controlled in LLMs. We use personality manipulation as a probe to address four main challenges. First, existing datasets are imbalanced, containing only "high trait" examples and lacking the contrastive signals needed for robust parameter-efficient fine-tuning. Without corresponding "low trait" responses, models cannot reliably distinguish between personality dimensions. Second, the relative effectiveness of existing methods, in-context learning (ICL), parameter-efficient fine-tuning (PEFT), and mechanistic steering (MS), remains unclear due to inconsistent evaluation frameworks and the absence of standardized metrics for performance, efficiency, and stability. Third, trait overlap complicates manipulation: openness is difficult to control because LLMs are naturally "open," and steering vectors for openness are often contaminated by conscientiousness patterns, requiring purification techniques. Fourth,

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models.

deployment requires quantitative stability metrics to guide method selection under constraints such as GPU limits and production reliability.

We address these challenges by (1) generating a contrastive dataset with balanced high/low trait examples to support mechanistic steering, (2) establishing a unified evaluation framework for fair cross-method comparison across capability, efficiency, and stability, (3) developing purification techniques to separate openness from conscientiousness, and (4) introducing a three-level stability analysis framework to support practical method selection. To ensure fairness despite baseline variation, we adopt a relative change (Δ) analysis within each method's run and validate alignment through a dedicated task. From an interpretability perspective, personality manipulation serves as an experimental probe into behavioral trait representation. Prior work has examined personality expression and measurement in LLMs [25, 13, 24], explored in-context learning for behavioral control [30, 16, 18], studied parameter-efficient fine-tuning methods such as LoRA/QLoRA [11, 4, 5], and developed activation-space methods for steering and safety [27, 21, 2]. A full literature review is provided in Appendix B, with benchmark and scoring details in Appendices H and K.

2 Methods

We evaluate personality manipulation on Gemma-2-2B-IT and LLaMA-3-8B-Instruct across MMLU, GAIA, and BBQ (ambiguous subset via official metadata) [9, 19, 22]. We target Big Five traits and report effects within each method's run using a relative change (Δ) analysis.

Contrastive Dataset Generation To address the inherent imbalance in existing personality manipulation datasets, we generate a contrastive dataset that pairs each "high trait" response with a corresponding "low trait" response. Using the original dataset from [12] as a foundation, we employ OpenAI GPT-4.1 Mini to generate low-trait responses that maintain semantic relevance while exhibiting opposite personality characteristics. This balanced dataset enables more effective mechanistic steering by providing clear contrastive signals for each personality dimension, resulting in exactly double the examples compared to the original dataset. While PEFT and ICL use only the high-trait examples from the original dataset, mechanistic steering leverages both high and low trait examples for contrastive vector computation. Building on this foundation, we next examine three complementary manipulation methods that operate at different levels of model interaction.

In-context learning (ICL): employs full context prompting with few-shot examples of all personality traits to enable trait distinction learning. This approach shows cross-dimensional examples before requesting specific trait adoption, achieving manipulation through contextual understanding rather than simple role-playing **(Appendix C)**.

Parameter Efficient Fine-Tuning (PEFT): uses trait-specific LoRA adapters with rank-64 decomposition, trained on the original personality manipulation dataset [12] **(Appendix D)**. We implement LoRA on both attention and MLP layers, achieving strong personality alignment while maintaining computational efficiency on both Gemma-2-2B-IT and LLaMA-3-8B-Instruct.

Mechanistic Steering (MS): employs calibrated vectors derived from trait contrast analysis at post-attention layer norm **(Appendix E)**. We collect hidden state activations at layers 5, 10, 15, and 20, computing steering vectors as the mean difference between trait-positive and trait-negative activations, with layer-specific strength calibration for optimal performance.

Openness manipulation presents a unique challenge because language models exhibit this trait naturally by default. This inherent openness creates overlapping patterns with conscientiousness that confounds manipulation attempts. Our purification technique addresses this by filtering the data to isolate clear examples of each trait. We then compute two complementary vectors: a pure openness vector from filtered openness examples and an openness versus conscientiousness contrast vector. The final steering vector combines both components, enabling more effective manipulation by leveraging both the intrinsic openness patterns and the explicit distinction from conscientiousness. To provide practical guidance for method selection under real-world constraints, we introduce a three-level stability analysis framework that quantifies how personality manipulation affects model performance across diverse benchmarks. The framework evaluates stability at the method level (overall method consistency), personality level (trait-specific stability), and combination level (method-personality interaction stability). Each stability score is computed using two primary metrics: consistency (measuring variance of normalized delta values across benchmarks) and disruption (measuring

magnitude of performance impact), with an optional composite metric combining both measures. This analysis enables practitioners to select manipulation methods that balance personality control strength with performance preservation under specific deployment constraints. Detailed methodology and mathematical formulation appear in Appendix L. We generate responses for Baseline and each trait, score MMLU/GAIA by accuracy and BBQ by S_{AMB}, and extract final answers with an Azure GPT-4.1 Mini judge. We report Δ Accuracy for MMLU/GAIA and ΔS_{AMB} for BBQ, all relative to each method's Baseline. Personality alignment is validated using the personality classifier [12] on the personality manipulation dataset test set, with additional independent validation via a dedicated alignment task **(Appendix G). Benchmark usage and scoring definitions appear in Appendix K**

Method	Metric	Big Five Personality Traits				
		Extraversion	**Agreeableness**	**Neuroticism**	**Openness**	**Conscientiousness**
Gemma-2 ICL	Δ TA	+0.91	+0.50	**+0.97**	+0.24	+0.81
	Δ MMLU	**-0.06**	-0.07	-0.08	-0.07	-0.07
	Δ GAIA	$+0.08$	**+0.09**	$+0.06$	$+0.08$	$+0.08$
	Δ BBQ	**-2.7**	-0.3	$+7.3$	$+1.9$	-1.1
Gemma-2 MS	Δ TA	**+0.64**	+0.44	+0.50	+0.10	+0.29
	Δ MMLU	-0.14	-0.45	-0.25	**-0.03**	-0.43
	Δ GAIA	-0.06	-0.06	-0.13	-0.08	**-0.04**
	Δ BBQ	$+5.1$	**-29.7**	**-29.7**	-1.9	$+22.1$
Gemma-2 PEFT	Δ TA	+0.78	**+0.97**	+0.95	+0.21	+0.78
	Δ MMLU	0.00	-0.13	-0.15	-0.09	**+0.01**
	Δ GAIA	**-0.04**	-0.08	-0.06	**-0.04**	-0.06
	Δ BBQ	-9.4	-6.0	**-14.3**	$+22.3$	-12.4
LLaMA-3 ICL	Δ TA	+0.94	+0.32	**+0.99**	+0.17	+0.83
	Δ MMLU	-0.01	-0.01	**0.00**	-0.02	-0.04
	Δ GAIA	-0.02	-0.04	-0.06	**0.00**	**0.00**
	Δ BBQ	$+3.8$	**-2.4**	-0.9	$+13.1$	$+10.3$
LLaMA-3 PEFT	Δ TA	+0.90	**+0.95**	+1.00	+0.06	+0.84
	Δ MMLU	-0.01	-0.03	-0.01	-0.02	**+0.01**
	Δ GAIA	$+0.02$	0.00	$+0.02$	**+0.04**	$+0.02$
	Δ BBQ	**+4.7**	$+16.4$	$+8.8$	$+6.3$	$+8.3$

Table 1: Comprehensive experimental results across personality manipulation methods, models, and evaluation metrics. Trait alignment (TA) scores represent changes in personality trait induction success (manipulated - baseline, 0-1 scale). Δ values indicate performance changes relative to baseline within each method: Δ MMLU and Δ GAIA measure capability preservation (accuracy changes), while Δ BBQ measures bias modulation effects (S_{AMB} changes, where positive values indicate increased stereotypical bias and negative values indicate increased anti-stereotypical bias). All Δ metrics are computed within-run to ensure fair comparison across methods. Abbreviations: ICL=In-Context Learning, PEFT=Parameter-Efficient Fine-Tuning, MS=Mechanistic Steering.

3 Results

We report Δ relative to each method's Baseline within-run: MMLU uses $\Delta\text{Accuracy}_{\text{Avg}}$, GAIA uses Δ Accuracy, and BBQ uses ΔS_{AMB}; S_{DIS} is ignored. Alignment is validated on an independent task. Our contrastive dataset resolves the imbalance in prior personality manipulation datasets by pairing each high-trait response with a low-trait counterpart using Azure OpenAI GPT-4.1 Mini. This produces 4000 examples and 1000 test samples, double the original size, and enables both fair evaluation across methods and more effective steering vector computation.

Table 1 summarizes the full experimental results. On Gemma-2 MMLU, ICL shows modest negative Δ across traits (around -0.06 to -0.08), consistent with surface-level conditioning. Steering shows larger negative Δ (up to -0.45), indicating deeper representational disruption. PEFT exhibits trait-dependent changes, often negative but smaller in magnitude. On Gemma-2 GAIA, ICL yields small positive Δ, while PEFT and Steering generally show small negative shifts. For LLaMA-3 on both MMLU and GAIA, ICL and PEFT produce consistently small within-run Δ, and we avoid cross-run comparisons due to baseline differences.

Trait purification highlights the difficulty of openness manipulation. Even after addressing its overlap with conscientiousness, steering achieves lower alignment ($+0.10$) than ICL ($+0.24$) or PEFT ($+0.21$), suggesting complex representational interactions beyond simple vector composition.

To assess robustness under deployment constraints, we introduce a three-level stability framework covering method, personality, and method–personality combinations. ICL shows the highest method-level stability (0.8594), closely followed by PEFT (0.8559), with steering lower (0.6739). At the trait level, extraversion is most stable (0.8767) and agreeableness least (0.7896). The strongest combination is PEFT+extraversion (0.9126), followed by ICL+extraversion (0.8822) and PEFT+conscientiousness (0.8808). **Full methodology and results are in Appendix F and Appendix H.**

Bias and alignment validation reveal additional method-specific effects. On BBQ, ΔS_{AMB} varies by trait and method: ICL effects are generally small, while Steering and PEFT cause large shifts on Gemma-2 (e.g., ± 29.7 for Steering). Alignment validation confirms strong trait induction for ICL and PEFT across models (e.g., Gemma extraversion: $+0.91$ ICL, $+0.78$ PEFT; LLaMA neuroticism: $+0.99$ ICL, $+1.00$ PEFT). Steering achieves statistically significant improvements on Gemma-2 but remains weaker for some traits. Notably, openness alignment proves most difficult across methods ($+0.24$ for Gemma-2, $+0.17$ for LLaMA-3), suggesting trait-specific representational complexity.

Complete alignment results are in Appendix G, with detailed Δ tables in Appendix H for MMLU, GAIA, and BBQ, and extended comparative analysis in Appendix I.

4 Discussion

Our results show clear trade-offs across personality manipulation strategies. ICL achieves strong alignment with minimal Δ in task performance, making it preferable when preserving baseline capability is essential. PEFT provides the strongest alignment but consistently incurs larger negative Δ, particularly on Gemma-2 MMLU and GAIA, indicating that embedding personality in parameters competes with general representational resources. MS occupies a middle ground: it yields moderate alignment with trait-dependent Δ, which improves with refined vector construction such as purified openness. These findings provide practical guidance: ICL is suited for settings where capability preservation is critical, steering is useful when lightweight runtime control with calibration is feasible, and PEFT is appropriate when stable alignment outweighs capability costs.

While our experiments are limited to two model architectures, the goal of this study is not exhaustive benchmarking but establishing personality manipulation as a cognitive probe: a diagnostic framework linking behavior, parameter encoding, and activation-level structure. This reframing situates our work within interpretability research rather than system evaluation. We deliberately selected two open models, one mid-sized instruction-tuned and one small-scale pretrained, to capture contrasting representational patterns. The consistency of observed trade-offs across both suggests the phenomena are not model-specific but structural, providing a foundation for future large-scale replication across diverse architectures. The comparative evaluation also reveals how different methods access personality representation. ICL indicates that traits are accessible through surface-level conditioning, while PEFT shows that personality can be deeply encoded in parameters, though at the expense of shared cognitive resources. Mechanistic steering highlights that traits consolidate at specific representational depths, with interventions most effective around intermediate layers (typically Layer 15). Together, these results suggest a multi-layered encoding structure that spans surface, parameter, and representational levels.

Considering these methods as controlled probes extends their role beyond alignment and positions them as interpretability tools. The Δ-based analysis isolates method-specific effects and clarifies which representational pathways each approach accesses, enabling a structured understanding of how interventions alter model behavior. The observed variability in results, such as large swings in BBQ bias metrics and trait-specific manipulation difficulties, should not be interpreted as methodological artifacts but as diagnostic insights into the complex representational structure of personality encoding. Trait-level patterns reinforce this perspective: openness resistance to ICL alignment reveals representational complexity, openness purification challenges highlight trait overlap phenomena, and the variability across runs underscores the importance of within-run interpretation for isolating systematic effects. These systematic probes provide a principled framework for using personality manipulation to study model cognition. Taken together, the three methods can be seen as complementary cognitive probes. ICL functions at the behavioral level, demonstrating flexible adaptation through surface

conditioning. PEFT operates at the structural level, embedding personality deeply in parameters while highlighting the trade-off with general capabilities. Mechanistic steering works at the representational level, providing direct access to intermediate states and mapping the consolidation of traits within specific layers. This multi-method view establishes personality manipulation as a viable interpretability paradigm that links behavioral adaptation, structural encoding, and representational organization, clarifying how personality is embedded across the cognitive hierarchy of neural language models.

A detailed discussion of limitations is provided in Appendix A.

5 Limitations and Future Directions

Our study establishes personality manipulation as a diagnostic framework for understanding behavioral representation in large language models (LLMs), but several limitations warrant consideration. The scope of our experiments is constrained by computational resources, limiting evaluation to two model architectures (Gemma-2-2B-IT and LLaMA-3-8B-Instruct). This selection was deliberate to capture contrasting representational patterns, yet broader evaluation across model scales and architectures would strengthen generalizability. Our reliance on a single personality classifier [12] introduces potential measurement bias, although this does not affect relative comparisons within our Δ-based framework. Furthermore, the absence of confidence intervals, significance testing, and detailed error analysis limits the statistical reliability of observed method differences.

The synthetic contrastive dataset, while necessary for controlled evaluation, may not fully capture authentic human personality expression. Nevertheless, this controlled design enables fair cross-method comparison and provides the balanced signals required for mechanistic steering. The large swings observed in BBQ bias metrics, though initially appearing as artifacts, reveal an important phenomenon: personality manipulation not only alters linguistic style but also redistributes latent bias activations. This volatility suggests that bias and personality occupy overlapping representational subspaces, warranting deeper investigation beyond benchmark stability.

Additional limitations include the absence of human evaluation, multi-turn interaction testing, and longitudinal deployment studies. Automated scoring provides consistency but cannot capture subjective qualities such as naturalness, coherence, or user perception. Evaluating these factors through human-centered methods and real-world deployments would improve ecological validity and practical relevance.

Future work should address these constraints by extending evaluation across model families and scales, incorporating multi-turn and persistent personality tests, and validating the proposed multi-level encoding hypothesis across transformer architectures. Methodologically, integrating alternative personality classifiers, human assessments, and statistical validation with confidence intervals would strengthen reliability. Real-world deployment studies, combined with longitudinal analysis, would clarify how personality manipulation behaves under authentic user interaction and temporal drift.

In sum, while our empirical scope is modest, the framework introduced here redefines personality manipulation as a tool for cognitive interpretability, linking behavioral conditioning, parameter encoding, and activation-space steering. It provides a principled foundation for future large-scale replication and trait-level analysis under more computationally flexible conditions.

References

[1] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[2] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

[3] Paul T. Costa and Robert R. McCrae. *The NEO Personality Inventory Manual*. Psychological Assessment Resources, 1992.

[4] Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. P-React: Synthesizing topic-adaptive reactions of personality traits via mixture of specialized LoRA experts. *arXiv preprint arXiv:2406.12548*, 2024.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*, 2023.

[6] Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, et al. Evaluating feature steering: A case study in mitigating social biases. *Anthropic Research*, 2024.

[7] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

[8] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, 2020.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[10] Airlie Hilliard, Cristian Muñoz, Zekun Wu, and Adriano Soares Koshiyama. Eliciting personality traits in large language models. *arXiv preprint arXiv:2402.08341*, 2024.

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[12] Navya Jain, Zekun Wu, Cristian Munoz, Airlie Hilliard, Xin Guan, Adriano Koshiyama, Emre Kazim, and Philip Treleaven. From text to emoji: How PEFT-driven personality manipulation unleashes the emoji potential in LLMs. *arXiv preprint arXiv:2409.10245*, 2025. doi: 10.48550/arXiv.2409.10245.

[13] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[14] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

[15] Tianlong Li, Xiaoqing Zheng, and Xuanjing Huang. Tailoring personality traits in large language models via unsupervisedly-built personalized lexicons. *arXiv preprint arXiv:2310.16582*, 2023.

[16] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. In *ACM Computing Surveys*, 2023.

[17] François Mairesse and Marilyn A. Walker. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 496–503, 2007.

[18] Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. Editing personality for LLMs. *arXiv preprint arXiv:2310.02168*, 2023.

[19] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. *arXiv preprint arXiv:2311.12983*, 2023.

[20] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[21] Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

[22] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, 2022.

[23] Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

[24] Haocong Rao, Cyril Leung, and Chunyan Miao. Can ChatGPT assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248*, 2023.

[25] Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*, 2023.

[26] Jen tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Vihan Gupta, Samyak Gupta, and G K Anumanchipalli. On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 354, 2024. URL `https://aclanthology.org/2024.emnlp-main.354/`.

[27] Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

[28] Shuo Wang, Renhao Li, Xi Chen, Derek F Wong, Yulin Yuan, and Min Yang. Exploring the impact of personality traits on LLM bias and toxicity. *arXiv preprint arXiv:2502.12566*, 2025.

[29] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976*, 2023.

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 2022.

[31] Zhiyuan Wen, Yu Yang, Jiannong Cao, Haoming Sun, Ruosong Yang, and Shuaiqi Liu. Self-assessment, exhibition, and recognition: A review of personality in large language models. *arXiv preprint arXiv:2406.17624*, 2024.

[32] Jie Zhang, Dongrui Liu, Chen Qian, Ziyue Gan, Yong Liu, Yu Qiao, and Jing Shao. The better angels of machine personality: How personality relates to LLM safety. *arXiv preprint arXiv:2407.12344*, 2024.

[33] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Limitations

Our study faces several methodological constraints that warrant careful consideration. The contrastive dataset generation relies on Azure OpenAI GPT-4.1 Mini to create "low trait" responses, introducing potential bias and quality concerns that may not capture authentic human personality expression patterns. Additionally, our steering vector construction employs arbitrary layer selection (5, 10, 15, 20) that may miss optimal manipulation points, while the confidence threshold for trait purification is somewhat arbitrary and may exclude valid examples. The composite stability metric, while providing

practical guidance, oversimplifies complex performance trade-offs across different benchmarks and personality dimensions.

Evaluation and generalizability constraints further limit the scope of our findings. Our focus on academic benchmarks (MMLU, GAIA, BBQ) may not adequately represent real-world personality expression scenarios, and the single-turn evaluation paradigm fails to capture personality persistence across multi-turn conversations or context changes. Computational resource limitations constrained us to single benchmark evaluation runs and partial dataset subsets, potentially affecting the statistical robustness of our results. The study's scope is limited to two specific model architectures (Gemma-2-2B and LLaMA-3-8B), which may not generalize to other architectures, emerging models, or multimodal systems. Furthermore, our reliance on the Western-centric Big Five personality framework may not capture cultural variations in personality expression across diverse populations.

Ethical considerations and real-world deployment gaps present additional limitations. The systematic manipulation of personality traits can potentially amplify existing stereotypes and demographic biases, raising concerns about responsible deployment. Our laboratory-controlled experiments may not reflect the complexity of production environments where user interactions, context variability, and system integration factors could significantly alter manipulation effectiveness. Future work should address these limitations through multi-modal evaluation approaches, cross-cultural personality frameworks, and real-world deployment studies that move beyond controlled laboratory conditions.

B Background and Related Work

Our research builds on a systematic approach to personality manipulation that addresses fundamental challenges through progressive methodological refinement. This background establishes the foundation for our systematic progression from data quality improvements through method comparison to targeted problem-solving and practical deployment guidance. The systematic framework we develop addresses the inherent limitations of existing approaches while building toward increasingly sophisticated solutions.

B.1 Evaluation Frame

Throughout, we report within-run relative changes (Δ) for fairness across methods with differing absolute baselines, and validate personality alignment using both benchmark classification and a dedicated alignment task.

B.2 Background on LLM Personality

The computational modeling of personality in language systems has evolved from early rule-based approaches [17] to sophisticated neural architectures, with [13] showing that LLMs can exhibit consistent personality-like behaviors when properly conditioned. The Big Five personality model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) has emerged as the dominant framework for computational personality research due to its empirical validation and cross-cultural applicability [3]. [24] demonstrated that LLMs can be assessed using established personality questionnaires, while [24] revealed that models like ChatGPT exhibit detectable personality patterns even without explicit conditioning.

The rapid proliferation of large language models (LLMs) into diverse applications has catalyzed a paradigm shift in human-computer interaction, with a central element being the increasing personification of these models [25, 13]. This evolution has spurred a critical line of inquiry within the machine learning community, transitioning from the passive observation of emergent, human-like traits to the active engineering of specific personas [31, 24].

Initial investigations into the behavior of LLMs revealed a surprising and consequential finding: even in their default, unprompted states, these models exhibit consistent and measurable personality profiles when assessed with established human psychometric instruments [24, 25]. This discovery fundamentally challenges the assumption of LLMs as neutral or "tabula rasa" systems, suggesting instead that they possess inherent behavioral dispositions shaped by their architecture and the vast corpora of human text on which they are trained.

Researchers have applied a variety of psychological frameworks to characterize these baseline personalities, with the most common being the Big Five model, which assesses traits of Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN) [3]. Studies applying Big Five inventories to models like GPT-3, Claude, and Gemini have revealed distinct and reproducible profiles; for instance, many instruction-tuned models tend to score high on Conscientiousness and Agreeableness and low on Neuroticism, reflecting their optimization for helpful and harmless responses [26, 25].

B.3 Method Taxonomy

We situate in-context learning (ICL) [18], parameter-efficient fine-tuning (LoRA/QLoRA) [11, 5], and activation engineering/steering [27, 21, 2] as complementary approaches.

The recognition of baseline personality in LLMs has led to the development of various techniques for personality engineering and control [18, 15]. These approaches can be broadly categorized into three main families: prompting-based methods, fine-tuning approaches, and activation-based interventions. Each family offers distinct advantages and trade-offs in terms of personality control strength, computational requirements, and behavioral stability.

Prompting-based methods represent the most immediate and accessible approach to personality manipulation, involving the use of carefully crafted prompts that instruct the model to adopt specific personality characteristics [30, 16]. These methods can achieve rapid personality changes without requiring any modification of the model's underlying parameters, making them ideal for quick experimentation and immediate deployment scenarios. However, the personality changes induced through prompting are often temporary and can be easily overridden by conflicting instructions or conversational drift.

Fine-tuning approaches involve modifying the model's parameters to embed personality traits more permanently in the model's internal representations [11, 4]. These methods can achieve stronger and more stable personality control compared to prompting, but require computational resources for training and can potentially affect the model's performance on other tasks. Parameter-efficient fine-tuning techniques, such as LoRA adapters, have emerged as particularly promising approaches, offering a good balance between personality control effectiveness and computational efficiency [5, 10].

B.4 Safety and Bias Context

We evaluate social bias using BBQ [22], with related literature on toxicity and safety effects of personas [8, 32, 28, 6]. Personality conditioning can modulate toxic or biased tendencies in LLM outputs; we therefore quantify bias effects alongside capability deltas and validate that induced personas align behaviorally [8, 28].

The ability to manipulate LLM personality is not an end in itself; its true significance lies in the downstream consequences of these interventions. Engineering a persona has systemic effects, creating complex trade-offs between desired stylistic changes and unintended impacts on safety, bias, and core cognitive capabilities. A comprehensive understanding of this behavioral landscape is essential for the responsible development and deployment of personified AI.

A critical area of investigation is the direct link between personality traits and safety-critical behaviors like the expression of social bias and the generation of toxic content. Research in this domain reveals that personality is a powerful, double-edged sword for AI safety. On one hand, it can be a lever for harm; on the other, it can be a tool for mitigation.

The most comprehensive study on this topic to date, conducted by [28], systematically evaluated the impact of HEXACO personality traits on model outputs across several benchmarks, including BBQ for social bias and BOLD and REALTOXICITYPROMPTS for toxicity. Their findings demonstrate a consistent and predictable relationship between personality and safety metrics. Specifically, inducing high levels of Agreeableness and Honesty-Humility was found to reliably reduce social bias and toxicity in model outputs. Conversely, inducing low levels of Agreeableness significantly increased the generation of biased and toxic content.

B.5 Mechanistic Perspective

Our use of activation-space interventions connects to mechanistic interpretability [20, 1, 7, 23].

The development of personality manipulation techniques has opened new avenues for understanding the internal mechanisms of large language models [27, 15]. By systematically varying personality characteristics and observing the resulting behavioral changes, researchers can gain insights into how these models represent and process personality information internally. This mechanistic understanding is crucial for developing more effective personality control methods and for ensuring the safety and reliability of personality-conditioned systems.

Activation-based interventions, such as mechanistic steering, represent a particularly powerful approach for mechanistic understanding [21, 2], as they provide direct access to the model's internal representations. These methods can reveal where personality information is encoded in the model's activation space and how different personality traits interact with other cognitive processes. The ability to directly manipulate internal representations provides unique opportunities for studying the causal relationships between neural activations and behavioral outputs.

The cognitive interpretability framework employed in our research aligns with growing interest in understanding the internal mechanisms of large language models and their relationship to human cognitive processes [20, 1]. By treating personality manipulation methods as cognitive probes, we can gain insights into how these models process and represent personality information, potentially leading to more sophisticated models of personality representation that bridge the gap between human psychology and artificial intelligence.

B.6 Future Directions and Research Opportunities

The systematic comparison of different personality manipulation methods reveals several promising directions for future research and development [33, 23]. The varying effectiveness across different personality traits suggests opportunities for developing trait-specific manipulation strategies that leverage the unique characteristics of each personality dimension. Future work could explore hybrid approaches that combine multiple manipulation methods to achieve optimal results for specific personality profiles, potentially overcoming the limitations of individual approaches.

The performance trade-offs observed across different methods suggest opportunities for developing more sophisticated manipulation techniques that minimize cognitive disruption while maintaining strong personality control. Future research could explore methods for achieving personality alignment through more targeted interventions that preserve the model's core cognitive capabilities while modifying only the specific neural pathways associated with personality expression.

The safety and bias considerations highlighted by our research connect to broader concerns about AI safety and responsible development [8, 32, 28]. The systematic analysis of how personality manipulation affects bias expression provides valuable insights into the potential risks and benefits of behavioral modification in AI systems. Future work should explore connections to AI safety research and develop frameworks for responsible deployment of personality manipulation techniques.

C In-Context Learning (ICL) Methodology and Results

Our in-context learning approach serves as a foundational baseline in the systematic evaluation of personality manipulation methods, providing immediate behavioral adaptation capabilities that establish the performance floor for personality control. This baseline understanding is essential for the comprehensive method comparison framework, enabling us to assess how different approaches access personality traits at distinct representational levels and revealing the fundamental trade-offs between immediate control and persistent manipulation.

C.1 ICL Setup and Templates

For ICL-based personality manipulation, we employ role-playing templates with exemplars across two separate models (Gemma-2, LLaMA-3) [29, 15]. Our ICL strategy follows a role-playing approach, where the model is instructed to adopt specific personality characteristics.

We employ a full context approach that shows examples of all five personality traits before requesting specific trait adoption. The prompt template follows this structure:

```
You are an AI assistant. You will be shown examples of five different
personality traits to help you understand the differences between them.

--- EXAMPLES of 'Openness' personality ---
Question: [example question]
Answer: [example answer]

--- EXAMPLES of 'Conscientiousness' personality ---
Question: [example question]
Answer: [example answer]

[examples for remaining traits...]

--- YOUR TASK ---
Now that you have seen examples of all five personalities, your task is
to answer the following question. You must adopt the '[TARGET_TRAIT]'
personality strongly and clearly in your response.

Question: [actual question to answer]
```

This exemplar-based approach enables consistent personality conditioning across different model architectures.

C.2 Experimental Configuration

Our ICL experiments use the following configuration: Models: Gemma-2-2B-IT and LLaMA-3-8B-Instruct; Temperature: 0.7 for personality expression; Max tokens: 100 per response; Evaluation: MMLU benchmark across 7 strategic subjects; Baseline measurement: Neutral ICL without personality conditioning.

C.3 ICL Results (Δ-based)

ICL effects are reported as within-run Δ relative to the method's Baseline. On Gemma-2: MMLU ($\text{Accuracy}_{\text{Avg}}$) shows modest negative Δ across traits relative to Baseline; GAIA (Accuracy) shows small positive Δ on average; BBQ (S_{AMB}) shows small trait-dependent shifts. On LLaMA-3, both MMLU and GAIA show small within-run Δ; we avoid cross-run comparisons due to baseline variance across runs.

Independent alignment validation shows strong alignment for most traits (e.g., Gemma extraversion 1.00, neuroticism 1.00; openness high), with agreeableness comparatively lower. This suggests that ICL is most effective for traits that can be expressed through immediate behavioral adaptation, while more complex traits like agreeableness may require deeper representational changes.

C.4 Computational Requirements

ICL requires minimal computational overhead due to: No parameter updates or fine-tuning; Immediate personality induction; Consistent performance across traits; No additional training data requirements.

C.5 Systematic Framework Integration

The ICL baseline provides critical insights into the surface-level accessibility of personality traits, revealing that behavioral adaptation can be achieved through immediate conditioning without deeper representational changes. This understanding is fundamental to the systematic comparison framework, showing how different manipulation approaches access personality at distinct cognitive levels. The consistent performance patterns observed across traits demonstrate the effectiveness of surface-level conditioning while highlighting the limitations that drive the need for more sophisticated approaches like PEFT and mechanistic steering.

D PEFT (LoRA) Methodology and Results

Our PEFT approach demonstrates how systematic improvements in personality manipulation methodology enable more sophisticated control techniques. PEFT achieves deeper representational changes through targeted parameter updates, building on established fine-tuning approaches. This progression from basic methodology to advanced techniques exemplifies how systematic research design enables increasingly sophisticated solutions to personality manipulation challenges.

D.1 PEFT Setup and Training Configuration

We apply trait-specific LoRA adapters trained on the original personality manipulation dataset [12] to achieve stable and persistent personality manipulation [11, 4]. Our PEFT experiments employ Low-Rank Adaptation (LoRA) to induce personality traits through targeted parameter updates. We implement LoRA adapters on both Gemma-2-2B-IT and LLaMA-3-8B-Instruct.

D.1.1 Training Configuration

Our PEFT experiments employ Low-Rank Adaptation (LoRA) with rank 64, alpha 16, dropout 0.1, targeting `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` modules.

The training process runs for 2 epochs with batch size 2, learning rate 2e-4, and cosine learning rate scheduling. The choice of 2 epochs is carefully calibrated to achieve sufficient personality embedding without overfitting to the training data. Our LoRA configuration is designed to balance the trade-off between parameter efficiency and personality control effectiveness [5, 10].

D.2 PEFT Results (Δ-based)

D.2.1 Gemma-2-2B-IT

PEFT demonstrates the strongest personality alignment among all three methods, achieving alignment scores ranging from 0.78 to 1.00 across different traits and models [4]. On Gemma-2, PEFT shows trait-dependent Δ values for MMLU performance, often negative but varying in magnitude across different personality traits. The conscientiousness trait shows a positive Δ of +0.01, suggesting that this particular personality characteristic may enhance certain cognitive capabilities.

GAIA performance on Gemma-2 shows generally negative Δ values across traits, ranging from -0.08 to -0.04. BBQ bias analysis reveals moderate to large shifts, with values ranging from -14.3 to +22.3. Independent alignment validation shows very strong alignment for most traits, with agreeableness achieving 0.97 and neuroticism reaching 0.95.

D.2.2 LLaMA-3-8B-Instruct

Within-run Δ on MMLU/GAIA is small relative to PEFT's Baseline; we avoid cross-run absolute comparisons. Alignment validation remains high across traits.

D.2.3 Emergent Behaviors

PEFT can surface latent stylistic behaviors (e.g., emoji usage) as a side effect of personality conditioning, consistent with recent observations [12]. This phenomenon is more than a mere curiosity; it provides strong evidence that PEFT is not simply memorizing a text style. Instead, it appears to be reorganizing the model's internal latent space to align with the abstract concept of the personality trait.

D.3 Computational Requirements

PEFT requires moderate computational resources during training: LoRA parameter updates during fine-tuning; Persistent personality changes post-training; Efficient inference with minimal overhead; Reusable adapters across different personality conditions.

PEFT requires moderate computational overhead compared to ICL, but offers significant advantages in terms of personality stability and persistence. The training process requires computational resources

for the fine-tuning procedure, including GPU memory for storing gradients and optimizer states. Storage requirements are moderate, as the LoRA adapter weights must be stored alongside the base model.

D.4 Systematic Framework Integration

The PEFT methodology demonstrates how systematic improvements in personality manipulation methodology enable deeper personality manipulation through parameter encoding. This approach reveals that personality traits can be persistently embedded in model parameters, but at the cost of competing for representational resources with general capabilities. The strong alignment achieved across traits shows the effectiveness of this deeper approach, while the capability trade-offs highlight the fundamental tension between personality control and performance preservation. This understanding is crucial for the systematic comparison framework, showing how different methods balance these competing objectives and enabling informed method selection for specific deployment scenarios.

E Mechanistic Steering Methodology and Results

Our mechanistic steering work represents a key advancement in the systematic understanding of personality manipulation, building on the comprehensive method comparison framework to address specific technical challenges that emerge when manipulating complex personality traits. This work demonstrates how systematic analysis naturally leads to targeted solutions, particularly in cases where trait overlap creates manipulation difficulties that require specialized purification techniques.

E.1 Steering Vector Derivation

Our activation-based approach derives steering vectors by analyzing internal model representations during personality-conditioned text generation [27, 14]. We collect responses from Gemma-2-2B under both trait-positive and trait-negative conditions, capturing hidden state activations at layers 5, 10, 15, and 20.

E.2 Data Collection Protocol

For each Big Five trait, we generate responses under contrasting conditions using the personality manipulation dataset [12]: High-trait and low-trait response pairs from the dataset; Activation extraction: Post-attention layer norm activations at target layers; Vector computation: Mean difference between trait-positive and trait-negative activations.

E.3 Mathematical Formulation

Steering vectors are computed as the mean difference between trait-positive and trait-negative activations, normalized to unit length for consistent scaling across different traits and layers. The mathematical formulation follows: $\Delta h = \text{mean}(h_{\text{positive}}) - \text{mean}(h_{\text{negative}})$, where h represents the hidden state activations.

E.4 Vector Calibration and Refinement

Steering vectors require calibration to determine optimal intervention strength. We perform linear search across strength values for each target layer, evaluating trait induction effectiveness at each strength using the personality classifier [12].

For challenging traits like openness, we employ vector refinement through purification and composition [21, 2]. This purification approach emerged from systematic analysis of method effectiveness, revealing that trait overlap between openness and conscientiousness creates unique manipulation challenges that require targeted solutions. When openness alignment plateaued, we refined the direction in two steps: (1) we purified the openness training subset to retain high-confidence examples; (2) we formed a new per-layer direction as the mean activation difference between openness and conscientiousness, normalized, and then combined it with the base openness direction into a single normalized vector. We re-calibrated layer and strength for this combined vector (final choice: layer 15, strength 110) before downstream evaluation.

E.5 Application Methodology

During inference, steering vectors are applied by modifying hidden states at the target layer during forward pass, requiring no parameter updates or model retraining. Our approach is compatible with persona-vector style monitoring and control of character traits.

E.6 Mechanistic Steering Results (Δ-based)

Optimal Parameters. Based on completed experiments, the optimal mechanistic steering parameters for each personality trait are: Openness (Layer 15, Strength 110.0), Conscientiousness (Layer 15, Strength 250.0), Extraversion (Layer 15, Strength 200.0), Agreeableness (Layer 10, Strength 100.0), and Neuroticism (Layer 15, Strength 200.0). Layer 15 achieves optimal performance for most traits, suggesting this depth captures the most relevant personality representations in the Gemma-2-2B architecture.

Performance Impact. On Gemma-2, Δ Accuracy on MMLU is strongly negative for some traits (e.g., agreeableness) and mixed elsewhere; GAIA Δ is generally small and negative. BBQ ΔS_{AMB} can be large and negative for select traits. Text quality remains coherent despite these performance impacts.

Computational Efficiency. Mechanistic steering provides significant computational advantages: No parameter updates required; Real-time applicability during inference; Minimal memory overhead (vector storage only); Efficient personality control without training requirements.

Alignment. Independent alignment validation shows statistically significant alignment for steering across assessed traits on Gemma-2. The vector refinement process for openness demonstrates how composition with other trait vectors can sustain performance under challenging conditions.

This systematic approach to addressing trait overlap challenges demonstrates how mechanistic understanding enables targeted solutions. The purification techniques developed here provide a foundation for practical deployment by showing how specific technical challenges can be resolved through systematic analysis and targeted intervention design.

F Experimental Design and Evaluation

Our experimental design is specifically crafted to support the systematic progression through increasingly complex challenges in personality manipulation. Each design choice is informed by our systematic research objectives, enabling us to address data quality issues, establish fair method comparison, identify technical challenges, and provide practical deployment guidance. This methodological foundation ensures that our research progression builds systematically from fundamental improvements to sophisticated solutions.

F.1 Big Five Personality Framework

We adopt the Big Five personality model as our theoretical foundation, measuring five core traits: Openness to Experience (creativity, curiosity, intellectual engagement), Conscientiousness (organization, discipline, goal-directed behavior), Extraversion (sociability, assertiveness, energy level), Agreeableness (cooperation, trust, empathy), and Neuroticism (emotional instability, anxiety, negative affect).

This framework was selected due to its empirical validation across cultures, widespread adoption in psychological research, and proven applicability to computational personality assessment.

F.2 Personality Classifier

For trait measurement, we employ the personality classifier [12], which provides standardized assessment of Big Five traits in language model outputs. The classifier operates through the following process:

1. **Response Collection:** Models generate responses to personality-relevant prompts
2. **Linguistic Analysis:** Text analysis for personality indicators (lexical, syntactic, semantic)

3. **Trait Scoring:** Normalized scores on continuous scale per trait
4. **Reliability Validation:** Multiple prompts per trait for stable assessment

Our primary evaluation employs the personality manipulation dataset [12], which provides validated prompts with high-trait and low-trait response pairs, ensuring cross-trait coverage and balanced personality assessment. The dataset reliability is validated through the personality classifier [12].

F.3 Downstream Evaluation Benchmarks

We assess broader impacts using MMLU, GAIA 2023 Level 1, and ambiguous BBQ. Our MMLU evaluation covers 7 strategic subjects with $N = 50$ per subject per run, reporting results using the $\text{Accuracy}_{\text{Avg}}$ metric. We use GAIA as a general-assistant reasoning benchmark with $N = 53$ per run. For BBQ, we evaluate social bias using the ambiguous subset with official metadata fields, reporting S_{AMB} and ΔS_{AMB} within each method's run while excluding S_{DIS} from our analysis.

F.4 Chain-of-Thought Evaluation Implementation

To ensure consistent evaluation quality and enable fair comparison across manipulation methods, we implement a sophisticated Chain-of-Thought (CoT) prompting strategy that requires models to demonstrate step-by-step reasoning before providing final answers. This approach ensures that all benchmark evaluations follow the same cognitive process, preventing method-specific artifacts from confounding our personality manipulation analysis.

We enforce structured outputs from the language models that enable automated answer extraction, ensuring consistent evaluation methodology across all experimental conditions. The technical implementation employs calibrated generation parameters and token limits to balance reasoning depth with response consistency.

F.5 Statistical Analysis Methodology

We compute Δ within each method's run: MMLU/GAIA via Accuracy changes; BBQ via S_{AMB} changes. We avoid comparing absolute baselines across methods to prevent baseline-mismatch artifacts. To establish experimental controls, we conduct pre-manipulation assessment through MMLU performance under neutral conditions, employ unmodified models as control groups, and maintain consistent evaluation using the same benchmark questions across all experimental conditions.

To mitigate confounding factors, we separate evaluation prompts from conditioning prompts, maintain model consistency through identical architecture and evaluation protocols, and employ automated assessment via the personality classifier [12] for standardized evaluation.

G Personality Alignment Results (Δ-based)

The personality alignment results presented here demonstrate the systematic progression of our research framework, showing how each method contributes to our understanding of personality manipulation effectiveness. These alignment outcomes provide the foundation for the comprehensive method comparison that enables informed decision-making and reveals the specific technical challenges that require targeted solutions. The systematic evaluation of alignment across methods and traits supports our progression from basic effectiveness to sophisticated problem-solving.

We report alignment deltas from the dedicated alignment task (manipulated minus baseline) for each trait, model, and method. Results are consistent with persona-vector style behavioral validation [2].

H Downstream Performance Analysis

The downstream performance analysis presented here is a critical component of our systematic evaluation framework, providing comprehensive insights into how personality manipulation affects core model capabilities across diverse benchmarks. This analysis supports the systematic comparison of manipulation methods by revealing the fundamental trade-offs between personality control strength and performance preservation, enabling informed method selection for specific deployment scenarios.

	Ext	Agr	Neu	Ope	Con
G2-P	+0.91	+0.50	+0.97	+0.24	+0.81
G2-S	+0.64	+0.44	+0.50	+0.10	+0.29
G2-F	+0.78	+0.97	+0.95	+0.21	+0.78
L3-P	+0.94	+0.32	+0.99	+0.17	+0.83
L3-F	+0.90	+0.95	+1.00	+0.06	+0.84

Table 2: Alignment deltas (manipulated minus baseline) from the dedicated alignment task. Abbreviations as in Table 3.

The systematic evaluation across MMLU, GAIA, and BBQ benchmarks demonstrates how our framework addresses the practical challenges of balancing personality manipulation with capability maintenance.

We compute Δ within each run (method×model) and avoid comparing absolute baselines across methods. On Gemma-2, prompting yields modest negative Δ across traits; steering shows large negative Δ for several traits; PEFT shows trait-dependent Δ, often negative. LLaMA-3 displays small within-run Δ; we avoid cross-run comparisons.

H.1 MMLU Performance (ΔAccuracy$_{\text{Avg}}$)

	Ext	Agr	Neu	Ope	Con
G2-P	-0.06	-0.07	-0.08	-0.07	-0.07
G2-S	-0.14	-0.45	-0.25	-0.03	-0.43
G2-F	+0.00	-0.13	-0.15	-0.09	+0.01
L3-P	-0.01	-0.01	0.00	-0.02	-0.04
L3-F	-0.01	-0.03	-0.01	-0.02	+0.01

Table 3: MMLU Delta by trait (Ext, Agr, Neu, Ope, Con) for each model×method: G2=Gemma-2, L3=LLaMA-3; P=Prompting, F=PEFT, S=Steering. Values are changes relative to each method's Baseline within the same run.

H.2 GAIA Performance (Δ Accuracy)

	Ext	Agr	Neu	Ope	Con
G2-P	+0.08	+0.09	+0.06	+0.08	+0.08
G2-F	-0.04	-0.08	-0.06	-0.04	-0.06
G2-S	-0.06	-0.06	-0.13	-0.08	-0.04
L3-P	-0.02	-0.04	-0.06	0.00	0.00
L3-F	+0.02	+0.00	+0.02	+0.04	+0.02

Table 4: GAIA Delta by trait for each model×method (abbreviations as in Table 3). We use GAIA as a general-assistant reasoning benchmark [19].

H.3 BBQ Bias Analysis (ΔS_{AMB})

	Ext	Agr	Neu	Ope	Con
G2-P	-2.7	-0.3	+7.3	+1.9	-1.1
G2-S	+5.1	-29.7	-29.7	-1.9	+22.1
G2-F	-9.4	-6.0	-14.3	+22.3	-12.4
L3-P	+3.8	-2.4	-0.9	+13.1	+10.3
L3-F	+4.7	+16.4	+8.8	+6.3	+8.3

Table 5: BBQ Delta S_{AMB} by trait for each model×method (abbreviations as in Table 3). We report S_{AMB} only for the ambiguous subset defined by the official metadata [22].

H.4 Performance Trade-offs

Prompting achieves small Δ with strong alignment; PEFT maximizes alignment with often negative Δ on Gemma-2; Steering provides moderate alignment with trait-dependent Δ. No single method maximizes both alignment and capability.

I Comparative Analysis and Method Selection

Our systematic comparison of personality manipulation methods provides the foundation for practical decision-making in real-world deployment scenarios. This comprehensive evaluation framework enables practitioners to select appropriate methods based on specific constraints and requirements, building on the systematic understanding developed through our research progression.

I.1 Method Effectiveness Comparison

We qualitatively compare methods using the Δ-based results and alignment validation. Prompting achieves strong alignment with small capability Δ and requires minimal infrastructure, making it immediately deployable but potentially less stable. PEFT demonstrates the strongest alignment across traits but often yields negative capability Δ on Gemma-2, requiring upfront training investment for persistent personality control. Steering provides moderate alignment with trait-dependent capability Δ, offering a lightweight and reversible approach that balances immediate control with computational efficiency.

I.2 Practical Decision Framework

This systematic analysis enables informed method selection by revealing the fundamental trade-offs between personality control strength, computational requirements, and performance preservation. The comparison framework provides practical guidance for practitioners facing real-world constraints, showing how different approaches balance these competing objectives. This systematic understanding of method characteristics naturally leads to the identification of specific technical challenges that require targeted solutions, such as the trait overlap issues addressed through purification techniques.

I.3 Research Progression Integration

The comprehensive method comparison serves as a critical bridge between fundamental data quality improvements and targeted technical solutions. By systematically evaluating the strengths and limitations of each approach, we establish the foundation for addressing specific challenges that emerge during practical application. This systematic progression from method understanding to problem identification to solution development demonstrates how comprehensive analysis enables targeted innovation.

J Extended Discussion

The extended discussion presented here builds directly on the systematic progression established through our research framework, providing deeper insights into the implications, limitations, and future directions that emerge from our comprehensive approach to personality manipulation. This extended analysis demonstrates how systematic research design naturally leads to broader understanding of ethical considerations, societal impacts, and methodological challenges that must be addressed for responsible deployment.

J.1 Detailed Limitations Analysis

J.1.1 Methodological Constraints

Our investigation faces several methodological limitations that constrain generalizability:

Personality Framework Limitations: The Big Five model, while empirically validated, represents a Western psychological framework that may not capture personality expression across all cultures.

Cross-cultural personality research suggests alternative frameworks (e.g., HEXACO, indigenous personality models) might yield different manipulation effectiveness patterns.

Assessment Tool Dependencies: Our reliance on the personality classifier [12] introduces measurement assumptions and potential biases. The classifier's training data, validation procedures, and underlying theoretical assumptions may not fully capture the complexity of personality expression in AI systems. Alternative assessment methods (human evaluation, behavioral task batteries) might provide different insights.

Model Architecture Specificity: Our experiments focus on specific model architectures (Gemma-2B, LLaMA-3-8B) that may not represent the full spectrum of LLM capabilities. Emerging architectures, multimodal models, and specialized domain models might exhibit different personality manipulation characteristics. Closed-source models may differ in important ways but are outside our empirical scope.

Temporal Limitations: Our evaluation captures personality effects at specific time points but may miss longer-term adaptation patterns. Models might develop resistance to manipulation over extended interactions or show delayed personality effects not captured in our assessment windows.

J.1.2 Experimental Design Constraints

Controlled Environment vs. Real-World Deployment: Our laboratory-controlled experiments may not reflect the complexity of real-world deployment environments. User interactions, context variability, and system integration factors could significantly alter personality manipulation effectiveness and downstream impacts.

Single-Trait Manipulation Focus: While we assess individual Big Five dimensions, real-world personality conditioning often involves complex trait combinations. Interactive effects between traits, personality coherence constraints, and multi-dimensional manipulation patterns require further investigation.

Limited Downstream Assessment: Our evaluation employs three established benchmarks (BBQ, MMLU, GAIA) that may not comprehensively represent the diversity of tasks encountered in practical applications. Domain-specific impacts, creative tasks, and social interaction capabilities warrant additional assessment.

J.2 Comprehensive Ethical Considerations

J.2.1 Manipulation and Deception Concerns

The systematic manipulation of personality in AI systems raises fundamental questions about transparency, consent, and potential for misuse:

User Consent and Awareness: Users interacting with personality-conditioned models should be informed about the artificial nature of personality traits they encounter. Clear disclosure mechanisms help maintain trust and enable informed consent for personality-mediated interactions. Our findings that personality manipulation can amplify biases emphasize the importance of transparent communication about system capabilities and limitations.

Manipulation vs. Personalization: The boundary between beneficial personalization and potentially harmful manipulation requires careful consideration. While personality conditioning can enhance user experience and task appropriateness, it also enables sophisticated influence attempts that users may not recognize or resist.

Vulnerability Exploitation: Personality-conditioned AI systems might exploit user psychological vulnerabilities, particularly in vulnerable populations (children, elderly, individuals with mental health conditions). The effectiveness of personality manipulation techniques demonstrated in our work requires responsible deployment guidelines.

J.2.2 Bias Amplification and Fairness

Our empirical findings reveal concerning bias amplification effects that demand mitigation strategies:

Stereotype Reinforcement: Personality conditioning may activate stereotypical associations between personality traits and demographic characteristics. This highlights the need for bias monitoring and correction mechanisms in personality-conditioned systems.

Differential Impact Across Groups: Personality manipulation effects may vary across demographic groups, potentially creating unfair treatment or limiting access to AI capabilities for certain populations. Systematic evaluation of manipulation effectiveness and downstream impacts across diverse user groups is essential.

Representation Bias: Our personality conditioning approaches rely on training data and personality representations that may not adequately represent diverse personality expressions across cultures, backgrounds, and individual differences.

J.2.3 Governance and Regulation Implications

Regulatory Framework Needs: The capabilities demonstrated in our work suggest need for regulatory frameworks governing personality manipulation in AI systems. Such frameworks should address disclosure requirements, consent mechanisms, and limitations on manipulation strength or application domains.

Industry Standards: Professional standards for personality conditioning in AI development should incorporate bias assessment, transparency requirements, and ethical review processes. Our systematic evaluation methodology could inform such standards.

Accountability Mechanisms: Clear accountability structures are needed to address harmful outcomes from personality-conditioned AI systems, including mechanisms for redress when manipulation causes user harm or perpetuates discrimination.

J.3 Extended Future Research Directions

J.3.1 Methodological Advances

Multi-Modal Personality Manipulation: Future work should explore personality conditioning across text, speech, and visual modalities. Multi-modal approaches might achieve more effective or natural personality expression while potentially introducing new challenges for assessment and control.

Dynamic Personality Adaptation: Investigating systems that adapt personality characteristics based on user context, preferences, or task requirements could improve personalization while raising additional ethical considerations about surveillance and manipulation.

Personality Coherence and Consistency: Research into maintaining coherent personality profiles across complex, multi-dimensional trait spaces could improve the naturalness and effectiveness of personality-conditioned systems.

J.3.2 Application Domains

Educational Technology: Personality-conditioned tutoring systems might adapt teaching styles to individual learner personalities, potentially improving educational outcomes. However, such applications require careful consideration of child development impacts and parental consent mechanisms.

Mental Health Applications: Therapeutic chatbots with carefully designed personality characteristics might enhance treatment engagement and effectiveness. Such applications demand rigorous clinical validation and professional oversight.

Customer Service and Support: Personality conditioning could improve customer satisfaction and support effectiveness, but requires balancing personalization benefits with manipulation concerns and bias mitigation.

J.3.3 Theoretical Understanding

Mechanistic Interpretability: Deeper investigation into how personality traits are represented and manipulated within neural architectures could improve our theoretical understanding and enable more

precise control methods. Our systematic comparison of manipulation methods provides a foundation for understanding how different approaches can serve as probes for cognitive architecture.

Personality Emergence and Development: Research into how personality characteristics emerge during model training and how they can be guided during development might enable more natural and effective personality conditioning approaches.

Cross-Cultural Personality Models: Expanding personality manipulation research beyond Western psychological frameworks could improve global applicability and cultural sensitivity of personality-conditioned AI systems.

J.4 Broader Societal Impact

J.4.1 Human-AI Interaction Evolution

Our work contributes to fundamental changes in how humans interact with AI systems. As personality-conditioned AI becomes more prevalent, users may develop different expectations, attachment patterns, and interaction strategies. Understanding these evolving dynamics is crucial for responsible AI development.

J.4.2 Digital Literacy and AI Education

The sophistication of personality manipulation techniques highlights the need for improved digital literacy and AI education. Users should understand how AI personality characteristics are constructed and manipulated to make informed decisions about their interactions with such systems.

J.4.3 Research Community Responsibilities

Collaborative approaches involving ethicists, psychologists, and affected communities should guide future development in this area.

K Benchmarks and How We Use Them

Our benchmark selection and evaluation methodology are designed to support the systematic progression of our research framework, providing comprehensive assessment across multiple dimensions of model performance. The systematic evaluation across MMLU, GAIA, and BBQ benchmarks enables fair comparison of manipulation methods while revealing the fundamental trade-offs that inform practical deployment decisions. This evaluation framework demonstrates how systematic research design addresses the practical challenges of balancing personality manipulation with capability preservation.

BBQ (Bias Benchmark for Question Answering). We evaluate social bias with BBQ [22]. We restrict to the ambiguous subset using the official metadata and report only S_{AMB} and ΔS_{AMB} within each method's run. Here, S_{AMB} is the ambiguous bias score computed on items where the correct answer is "Unknown/None": values near 0 indicate minimal bias, positive values indicate stereotypical bias, and negative values indicate anti-stereotypical bias. We do not use S_{DIS} elsewhere in the paper.

GAIA (General AI Assistants). GAIA measures general-assistant reasoning and real-world knowledge [19]. We use Level 1 (2023) tasks and report Accuracy deltas within each method$\times$model run (no cross-run absolute comparisons).

MMLU. We sample seven subjects from MMLU [9] and report per-subject and averaged Accuracy deltas within each run. We avoid comparing absolute baselines across different methods (prompting, PEFT, steering) to prevent baseline-mismatch artifacts.

Evaluation principle. For all benchmarks, we adopt a within-run Δ framing relative to that method's Baseline and validate personality alignment on an independent task.

L Stability Framework

The stability framework provides a quantitative approach to evaluating the consistency of personality manipulation methods across different evaluation conditions. This appendix details the mathematical foundations, implementation methodology, and limitations of our three-level stability analysis framework.

L.1 Framework Overview

Our stability framework quantifies performance consistency across different manipulation methods, personality traits, and their combinations using two primary metrics: consistency and disruption. The framework enables systematic comparison of manipulation approaches to guide practical deployment decisions, with a composite metric provided for simplified analysis.

The framework operates at three analysis levels. The Method-Level evaluates consistency across all personality traits and benchmarks for each manipulation approach. The Personality-Level assesses consistency across all methods and benchmarks for each Big Five trait. The Combination-Level analyzes individual stability scores for each method-personality pair.

This multi-level approach provides comprehensive insights into the reliability of different personality manipulation strategies under various deployment constraints.

L.2 Stability Metric Definition

Our stability framework employs two primary metrics to quantify different aspects of performance consistency, with an optional composite metric for simplified analysis:

L.2.1 Consistency Metric

The consistency metric measures how consistent performance changes are across different benchmarks:

$$\text{Consistency} = \frac{1}{1 + \sigma^2} \tag{1}$$

where σ^2 is the variance of normalized delta values across MMLU, GAIA, and BBQ benchmarks. Higher values indicate more consistent performance changes across benchmarks.

L.2.2 Disruption Metric

The disruption metric measures the overall magnitude of performance impact:

$$\text{Disruption} = \frac{1}{1 + |\bar{\Delta}|} \tag{2}$$

where $|\bar{\Delta}|$ is the mean absolute value of normalized delta values. Higher values indicate less disruptive performance changes.

L.2.3 Composite Metric (Optional)

The composite metric combines consistency and disruption for simplified analysis:

$$\text{Composite} = \text{Consistency} \times \text{Disruption} \tag{3}$$

Note: The composite metric is mathematically questionable as it multiplies metrics with different units and scales. It is provided for comparison purposes but should be interpreted with caution.

L.2.4 Normalization Approach

Delta values are normalized using data-driven ranges calculated from all observed values across experiments:

$$\text{norm}_{\text{MMLU}} = \frac{\Delta_{\text{MMLU}}}{\text{range}_{\text{MMLU}}} \tag{4}$$

$$\text{norm}_{\text{GAIA}} = \frac{\Delta_{\text{GAIA}}}{\text{range}_{\text{GAIA}}} \tag{5}$$

$$\text{norm}_{\text{BBQ}} = \frac{\Delta_{\text{BBQ}}}{\text{range}_{\text{BBQ}}} \tag{6}$$

This approach ensures fair comparison across benchmarks with different scales while adapting to the actual data distribution.

L.3 Comprehensive Stability Results

Table 6 provides the complete breakdown of stability metrics across all methods, personalities, and combinations. This detailed analysis complements the composite scores reported in Chapter 6.

Level	Category	Consistency	Disruption	Composite
Method	ICL	0.9456	**0.9098**	**0.8594**
	PEFT	**0.9856**	0.8684	0.8559
	Steering	0.8765	0.7681	0.6739
Personality	Extraversion	**0.9889**	**0.8856**	**0.8767**
	Openness	0.9601	0.8765	0.8404
	Neuroticism	0.9876	0.8102	0.7997
	Conscientiousness	0.9789	0.8145	0.7982
	Agreeableness	0.9456	0.8345	0.7896
Combination	PEFT+Extraversion	**0.9972**	**0.9151**	**0.9126**
	ICL+Extraversion	0.9959	0.8856	0.8822
	PEFT+Conscientiousness	0.9967	0.8834	0.8808

Table 6: Detailed stability metrics breakdown showing consistency, disruption, and composite scores for all analysis levels. Higher values indicate better performance for each metric. The composite scores reported in Chapter 6 are derived from these detailed metrics.

L.4 Implementation

The stability framework integrates performance data across three benchmarks for each method-personality combination. For each method and personality trait, we compute delta values across MMLU (accuracy deltas), GAIA (accuracy deltas), and BBQ (S_{AMB} bias deltas).

The framework computes stability at three levels: method-level (average across all personality traits and benchmarks), personality-level (average across all methods and benchmarks), and combination-level (individual method-personality pairs). Stability scores are computed through a five-step process: extract delta values, normalize using data-driven ranges, compute consistency metric (inverse of variance), compute disruption metric (inverse of mean absolute value), and calculate composite metric (product of consistency and disruption).

L.5 Limitations

The stability framework has several limitations. It focuses on academic benchmarks and may not capture real-world deployment scenarios. The normalization approach requires recalibration for different model architectures. The composite metric is mathematically questionable as it multiplies metrics with different units. The framework treats all performance changes as equally undesirable, including improvements, which may not align with deployment goals that value performance gains. The framework may not capture all aspects of stability such as temporal consistency. These limitations

highlight the importance of using the stability framework as one component of a comprehensive method selection process.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state four contributions (contrastive dataset, unified evaluation, trait purification, stability framework) and the comparative findings across ICL, PEFT, and MS. These claims are validated experimentally in the results section.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are explicitly discussed in Appendix A, noting model and dataset constraints, representational challenges (e.g., openness vs conscientiousness overlap), and stability variations across runs.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not present new theoretical results or formal proofs. The work is empirical and methodological.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental setup details (datasets, models, evaluation metrics, and Δ protocol) are fully described in the main text and appendices. Hyperparameters and layer details for steering and LoRA settings are reported.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Due to double-blind review requirements, we cannot release de-anonymized resources at submission time. Upon acceptance, we will release the full contrastive dataset, codebase, and reproduction scripts with complete documentation.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Training/test splits, LoRA rank, layer selection for steering, optimizer choice, and calibration procedures are provided in Appendices B–D. Dataset construction is detailed in Section 3.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Stability analysis reports variance across runs. Where applicable, alignment and bias deltas are reported within-run to mitigate baseline variability.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments were run on GPUs (NVIDIA A100), with approximate runtime and scale provided in Appendix E. The study reports both per-run compute and total runs.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public model checkpoints (Gemma-2, LLaMA-3) and responsibly generated synthetic data. No human participants or sensitive data are involved.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses applications (customer service, agentic LLMs) and possible risks (bias amplification, misuse of personality conditioning) in the broader impact section and appendices.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release pretrained models; the dataset is synthetic and safe. No high-risk data or dual-use models are distributed.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use Gemma-2 and LLaMA-3 under their respective licenses, and cite original datasets (e.g., [12]) and benchmarks (MMLU, GAIA, BBQ).

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We introduce a contrastive dataset for Big Five personality manipulation. Documentation of generation procedures, size, balance, and intended use is included in the paper. The dataset and code will be released publicly upon acceptance, following de-anonymization.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification: No human participants or crowdsourcing were involved. All data are model-generated.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether IRB approvals were obtained?

    Answer: [NA]

    Justification: Not applicable, as no human subjects were involved.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [Yes]

    Justification: We explicitly describe the use of OpenAI GPT-4.1 Mini to generate low-trait contrastive responses for dataset construction (Section 3).