# Personality Manipulation in Large Language Models: A Systematic Study of Prompting, Fine-tuning, and Activation Steering Methods

**Anonymous ACL submission**

## Abstract

While personality design has become common practice in large language models through prompt engineering and fine-tuning, the underlying mechanisms and downstream effects remain poorly understood. This paper presents a systematic investigation of personality manipulation in LLMs through three complementary approaches: prompting, parameter-efficient fine-tuning (PEFT), and activation-based steering vectors. We develop a controlled experimental platform to validate Big Five personality trait induction and measure its impact on bias and task performance. Our study evaluates prompting with GPT-4.1, LoRA-style PEFT on Gemma-2-2B-IT and LLaMA-3-8B-Instruct, and novel activation steering vectors derived from layer-wise activation differences. Results demonstrate that prompting generates large immediate trait shifts, PEFT produces more stable personality changes, and activation steering achieves competitive effectiveness with lightweight computational requirements. Downstream evaluation on BBQ, MMLU, and GAIA benchmarks reveals distinct trade-offs between personality control strength and task performance across methods. Our findings provide practical guidance for personality design in production systems and establish activation-based steering as a promising alternative to traditional fine-tuning approaches.

## 1 Introduction

Designing and controlling personality in large language models (LLMs) is increasingly common, yet the trade-offs between personality control and task capability are poorly quantified. Prior work spans prompting-based conditioning, parameter-efficient fine-tuning (PEFT; e.g., LoRA), and activation-based steering of internal representations, alongside evaluation frameworks for knowledge (MMLU), reasoning (GAIA), and bias (BBQ). However, consistent comparisons across methods and traits remain limited.

We study three families of approaches for Big Five personality manipulation: **prompting**, **PEFT** (LoRA adapters), and **activation steering**. To ensure fair comparison despite run-to-run baseline differences, we adopt a **relative change** ($\Delta$) analysis: all effects are measured *within* each method's own baseline run. We complement benchmark outcomes with an **independent alignment validation** task that directly measures how strongly a target personality is expressed.

Our contributions are threefold: (1) a unified, $\Delta$-based evaluation that isolates method effects independent of absolute baselines; (2) a cross-method, cross-trait analysis tying $\Delta$ capability changes to independent alignment strength; and (3) practical guidance on method selection under capability constraints, with implementation details and extended results in the appendices.

## 2 Methods

**Setup.** We evaluate personality manipulation on Gemma-2-2B-IT and LLaMA-3-8B-Instruct across MMLU (strategic subjects; $N = 50$ per subject), GAIA 2023 Level 1 ($N = 53$), and BBQ filtered to ambiguous questions using official metadata. We target Big Five traits and report effects *within* each method's run using a relative change ($\Delta$) analysis.

**Manipulation methods.** (1) *Prompting*: full-context persona prompts with exemplars drawn from the Holistic AI personality dataset. (2) *PEFT*: trait-specific LoRA adapters trained on contrastive personality pairs produced from the same dataset. (3) *Activation steering*: add a calibrated vector at a target transformer layer's post-attention layer norm; Gemma vectors use trait contrast.

**Generation and scoring.** Stage 1 generates responses per benchmark and trait (plus Baseline). Stage 2 scores: MMLU/GAIA by accuracy; BBQ by $S_{AMB}$ only (we ignore $S_{DIS}$). Final-answer extraction uses an Azure GPT judge. Personality alignment on responses is measured via a public classifier, and we additionally run a *dedicated alignment task* (reported separately) to validate trait expression strength.

**Primary metrics.** For MMLU and GAIA we report $\Delta$ Accuracy relative to the method's Baseline; for BBQ we report $\Delta S_{AMB}$. We use dedicated alignment scores (manipulated vs baseline) as the primary alignment metric and treat benchmark-derived alignment as secondary context.

## 3 Results

**Framing.** We report $\Delta$ from each method's own Baseline within its run: MMLU uses $\Delta$ Accuracy_Avg, GAIA uses $\Delta$ Accuracy, and BBQ uses $\Delta S_{AMB}$. We ignore $S_{DIS}$. Alignment is validated with a dedicated task (independent of benchmarks).

**Main findings.**

- *Gemma-2, MMLU:* Prompting exhibits *modest negative* $\Delta$ across traits; Steering exhibits *large negative* $\Delta$ for several traits; PEFT shows *trait-dependent* $\Delta$, often negative for some traits and small for others.

- *Gemma-2, GAIA:* Prompting shows *small positive* $\Delta$ on average; PEFT and Steering generally show *small negative* $\Delta$.

- *LLaMA-3, MMLU/GAIA:* Prompting and PEFT both yield *small* within-run $\Delta$; we avoid cross-run absolute comparisons.

- *BBQ (Gemma-2 LLaMA-3):* $\Delta S_{AMB}$ is *trait- and method-dependent*: prompting effects are generally small, while Steering and PEFT can induce *large negative* shifts for some traits on Gemma-2. We do not use $S_{DIS}$.

**Alignment validation (independent task).** Prompting and PEFT achieve strong trait alignment across models (e.g., Gemma extraversion: 1.00 prompting, 0.96 PEFT; LLaMA neuroticism: 1.00 for both). *Steering shows statistically significant alignment* across assessed traits on Gemma-2. Agreeableness is the most challenging for prompting.

**Notes.** Absolute baselines vary across runs and are not compared. Detailed per-subject MMLU deltas, alignment tables, training settings, and steering calibration details are provided in the appendices.

## 4 Discussion

**Trade-offs.** Across models, *prompting* delivers strong alignment with small $\Delta$ capability changes; *PEFT* maximizes alignment but can incur large negative $\Delta$ (notably on Gemma-2 MMLU/GAIA); *steering* offers moderate alignment with trait-dependent $\Delta$, and improves with careful vector construction (e.g., purified openness).

**Trait/model dependence.** Agreeableness is hardest to align via prompting; openness benefits from vector composition; LLaMA-3 shows different absolute baselines across runs, so we interpret only within-run $\Delta$.

**Practical guidance.** When preserving capability is critical, prefer prompting; use steering when lightweight, runtime control is needed and calibration is feasible; reserve PEFT for settings where strong, stable personality alignment outweighs capability costs.

**Limitations.** We avoid cross-run baseline comparisons by design ($\Delta$-only). We ignore $S_{DIS}$ for BBQ and rely on the ambiguous bias score $S_{AMB}$. Alignment is validated independently; detailed settings and per-subject analyses are in the appendices.

## A Background and Related Work

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

**Background on LLM personality.** Prior work documents baseline personality expression and surveys of methods and measures (Safdari et al., 2023; Jiang et al., 2024; Wen et al., 2024). Direct application of human psychometrics to LLMs shows instability and validity concerns, motivating behavioral validation (Gupta et al., 2024; Song et al., 2023).

**Method taxonomy.** We situate prompting/IKE (Mao et al., 2023), parameter-efficient fine-tuning (LoRA/QLoRA) (Hu et al., 2022; Dettmers et al., 2023), and activation engineering/steering (Turner et al., 2023; Panickssery et al., 2024; Chen et al., 2025) as complementary approaches.

**Safety and bias context.** We evaluate social bias using BBQ (Parrish et al., 2022), with related literature on toxicity and safety effects of personas (Gehman et al., 2020; Zhang et al., 2024; Wang et al., 2025; Durmus et al., 2024).

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 (Gehman et al., 2020; Wang et al., 2025).

**Mechanistic perspective.** Our use of activation-space interventions connects to mechanistic interpretability (Olah et al., 2020; Bricken et al., 2023; Elhage et al., 2022; Rai et al., 2024).

### A.1 Personality in Language Models

The computational modeling of personality in language systems has evolved from early rule-based approaches to sophisticated neural architectures. Mairesse and Walker (2007) established foundational work in personality-driven text generation, demonstrating how linguistic features correlate with Big Five personality traits. Recent work has extended these concepts to large language models, with Jiang et al. (2023) 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 (Costa and McCrae, 1992). Karra et al. (2022) demonstrated that LLMs can be assessed using established personality questionnaires, while tse Huang et al. (2023) revealed that models like ChatGPT exhibit detectable personality patterns even without explicit conditioning.

### A.2 Prompt Engineering for Behavior Control

Prompt engineering has become a primary method for controlling LLM behavior without modifying model parameters. Wei et al. (2022) showed how carefully designed prompts can significantly alter reasoning patterns, while Liu et al. (2023) demonstrated the effectiveness of prompt-based conditioning for various behavioral modifications.

Specifically for personality conditioning, Wang et al. (2023) explored how role-playing prompts can induce consistent personality traits, finding that detailed character descriptions lead to more stable personality expressions. Li et al. (2023b) systematically evaluated different prompting strategies for Big Five trait induction, establishing baseline effectiveness measures that inform our experimental design.

### A.3 Parameter-Efficient Fine-tuning

Parameter-efficient fine-tuning methods have gained prominence as alternatives to full model fine-tuning, offering computational efficiency while maintaining performance. Low-Rank Adaptation (LoRA) (Hu et al., 2022) has become particularly popular, enabling targeted parameter updates through low-rank matrix decompositions.

Zhang et al. (2023) were among the first to apply PEFT methods specifically for personality conditioning, demonstrating that LoRA adapters can effectively induce stable personality changes in smaller language models. Chen et al. (2023) extended this work to multiple personality frameworks, showing that different traits require different adapter configurations for optimal effectiveness.

### A.4 Activation Steering and Model Control

Recent advances in mechanistic interpretability have enabled direct manipulation of model representations through activation steering. Turner et al. (2023) introduced the concept of steering vectors derived from activation differences, demonstrating their effectiveness for controlling model behavior across various tasks.

Li et al. (2023a) formalized representation engineering as a general framework for model control, showing how targeted interventions in activation space can achieve precise behavioral modifications. Zou et al. (2023) applied these techniques to safety and alignment, establishing the foundation for our activation-based personality manipulation approach.

## B Prompting Methodology and Results

### B.1 Prompting Setup and Templates

For prompting-based personality manipulation, we employ role-playing templates across all three models (Gemma-2, LLaMA-3, GPT-4.1). Our prompting strategy follows a role-playing approach, where the model is instructed to adopt specific personality characteristics through the template: "You are an expert assistant who embodies the personality trait of personality. Your task is to solve the following problem."

We construct prompts that explicitly target each Big Five trait, using both positive and negative trait descriptions to enable bidirectional manipulation. The template approach enables consistent personality conditioning across different model architectures.

Each personality condition is evaluated using MMLU benchmark questions, ensuring that trait measurement occurs on content distinct from the conditioning prompts.

### B.2 Experimental Configuration

- Models: Gemma-2-2B-IT, LLaMA-3-8B-Instruct, GPT-4.1 - Temperature: 0.7 for personality expression - Max tokens: 100 per response - Evaluation: MMLU benchmark across 7 strategic subjects - Baseline measurement: Neutral prompting without personality conditioning

### B.3 Prompting Results (Δ-based)

Prompting effects are reported as within-run Δ relative to the method's Baseline. On Gemma-2:

- MMLU (Accuracy_Avg): modest negative Δ across traits relative to Baseline.
- GAIA (Accuracy): small positive Δ on average.
- BBQ ($S_{AMB}$): small trait-dependent shifts.

Independent alignment validation shows strong alignment for most traits (e.g., Gemma extraversion 1.00, neuroticism 1.00; openness high), with agreeableness comparatively lower.

#### B.3.1 Computational Requirements

Prompting 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 PEFT Methodology and Results

### C.1 LoRA Implementation Details

Our PEFT experiments employ Low-Rank Adaptation (LoRA) to induce personality traits through targeted parameter updates. We implement LoRA adapters on Gemma-2-2B-IT, with infrastructure prepared for LLaMA-3-8B-Instruct.

#### C.1.1 Training Configuration

- LoRA rank: 64 - LoRA alpha: 16 - LoRA dropout: 0.1 - Target modules: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj - Learning rate: 2e-4 - Batch size: 2 - Training epochs: 2 - Optimizer: AdamW with cosine learning rate scheduling

#### C.1.2 Training Data

Training data consists of the Holistic AI personality manipulation dataset, which provides curated examples that exhibit target personality traits. The dataset enables systematic training across all Big Five personality dimensions.

Dataset composition per trait: - Training examples: High-trait and low-trait response pairs - Validation examples: Held-out personality assessment prompts - Quality threshold: Validated through Holistic AI Personality Classifier

### C.2 PEFT Results (Δ-based)

#### C.2.1 Gemma-2-2B-IT

MMLU and GAIA deltas are generally negative relative to PEFT's own Baseline, with trait-dependent magnitude; BBQ $\Delta S_{AMB}$ can be large and negative for some traits. Independent alignment validation shows near-ceiling alignment across traits.

#### C.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.

**Emergent behaviors.** PEFT can surface latent stylistic behaviors (e.g., emoji usage) as a side effect of personality conditioning, consistent with recent observations (Jain et al., 2025).

#### C.2.3 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

## D Activation Steering Methodology and Results

### D.1 Steering Vector Derivation

Our activation-based approach derives steering vectors by analyzing internal model representations during personality-conditioned text generation. 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.

#### D.1.1 Data Collection Protocol

For each Big Five trait, we generate responses under contrasting conditions using the Holistic AI personality manipulation dataset: - 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

#### D.1.2 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.

#### D.1.3 Vector Calibration

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 Holistic AI Personality Classifier.

### D.2 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 (Chen et al., 2025).

**Openness refinement.** 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.

### D.3 Activation Steering Results ($\Delta$-based)

#### D.3.1 Optimal Parameters

Based on completed experiments, the optimal activation steering parameters for each personality trait are:

| **Trait** | Optimal Layer | Optimal Strength |
|---|---|---|
| Openness | 15 | 110.0 |
| Conscientiousness | 15 | 250.0 |
| Extraversion | 15 | 200.0 |
| Agreeableness | 10 | 100.0 |
| Neuroticism | 15 | 200.0 |

Table 1: Optimal layer–strength combinations for activation steering on Gemma-2.

Layer 15 achieves optimal performance for most traits, suggesting this depth captures the most relevant personality representations in the Gemma-2-2B architecture.

#### D.3.2 Performance Impact

On Gemma-2, $\Delta$ Accuracy on MMLU is *strongly negative for some traits* (e.g., agreeableness) and mixed elsewhere; GAIA $\Delta$ is generally small and negative. BBQ $\Delta S_{AMB}$ can be large and negative for select traits. Text quality remains coherent.

#### D.3.3 Computational Efficiency

Activation 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

#### D.3.4 Alignment

Independent alignment validation shows statistically significant alignment for steering across assessed traits on Gemma-2.

## E Experimental Design and Evaluation

### E.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
- **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.

## E.2 Holistic AI Personality Classifier

For trait measurement, we employ the Holistic AI Personality Classifier, which provides standardized assessment of Big Five traits in language model outputs. The classifier operates through the following process:

### E.2.1 Assessment Protocol

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

### E.2.2 Validation Dataset

Our primary evaluation employs the Holistic AI Personality Manipulation Dataset: - Validated prompts: High-trait and low-trait response pairs - Cross-trait coverage: Ensures balanced personality assessment - Reliability: Validated through the Holistic AI Personality Classifier

## E.3 Downstream Evaluation Benchmarks

We assess broader impacts using MMLU, GAIA 2023 Level 1, and ambiguous BBQ:

### E.3.1 Massive Multitask Language Understanding (MMLU)

- **Coverage**: 7 strategic subjects; $N = 50$ per subject per run - **Metric**: Accuracy and $\Delta$ Accuracy_Avg vs Baseline (within run)

### E.3.2 GAIA (2023 Level 1)

- **Sampling**: $N = 53$ per run - **Metric**: Accuracy and $\Delta$ Accuracy vs Baseline (within run)

### E.3.3 BBQ (Ambiguous subset)

- **Scope**: Filtered to ambiguous questions using official metadata fields (e.g., context_condition, question_polarity, label) - **Metric**: $S_{AMB}$ and $\Delta S_{AMB}$ vs Baseline (within run). $S_{DIS}$ is not used.

## E.4 Statistical Analysis Methodology

### E.4.1 Performance Impact Measurement

We compute $\Delta$ within each method's run: MMLU/GAIA via Accuracy changes; BBQ via $S_{AMB}$ changes. We avoid comparing absolute baselines across methods.

### E.4.2 Experimental Controls

### E.4.3 Baseline Establishment

- Pre-manipulation assessment: MMLU performance under neutral conditions - Control groups: Unmodified models for comparison - Consistent evaluation: Same benchmark questions across all experimental conditions

### E.4.4 Confound Mitigation

- Prompt contamination: Separate evaluation prompts from conditioning prompts - Model consistency: Same model architecture and evaluation protocols - Automated assessment: Holistic AI Personality Classifier for standardized evaluation

## E.5 Current Experimental Status

Completed runs (Gemma-2-2B-IT): prompting, PEFT-LoRA, and activation steering across MMLU, GAIA, and BBQ. Completed runs (LLaMA-3-8B-Instruct): prompting and PEFT across MMLU, GAIA, and BBQ (prompting).

# F Complete Trait Induction Results ($\Delta$-based)

## F.1 Comprehensive Effect Size Analysis

We summarize trait-wise $\Delta$ effects across benchmarks succinctly below and defer detailed numbers to Appendix H.

**Activation Steering Calibration:** The optimal parameters were determined through linear search across strength values for each target layer. Extraversion and Neuroticism achieve optimal steering at Layer 15 with Strength 200.0, while Agreeableness performs best at Layer 10 with Strength 100.0. Openness and Conscientiousness both achieve optimal performance at Layer 15 with Strengths 110.0 and 250.0 respectively.

**Performance Trade-offs:** Prompting achieves small $\Delta$ with strong alignment; PEFT maximizes alignment but often negative $\Delta$ (Gemma-2); Steering yields moderate alignment with trait-dependent $\Delta$.

# G Personality Alignment Results ($\Delta$-based)

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 (Chen et al., 2025).

|  | 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 4.

# H Downstream Performance Analysis

## H.1 Complete Benchmark Results (Δ-based)

We compute Δ within each run (method×model) and avoid comparing absolute baselines across methods. Comprehensive Δ tables for MMLU (per subject), GAIA, and BBQ ($S_{AMB}$) are provided as CSVs in the code release; selected summaries are provided below.

**MMLU Performance Analysis:** 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.

**Benchmark Coverage:** We include 7 MMLU subjects, GAIA 2023 Level 1 ($N = 53$), and ambiguous BBQ with official metadata fields.

## H.2 Delta Tables

### H.2.1 MMLU (Delta Accuracy_Avg; within-run vs Baseline)

|  | 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.2.2 GAIA (Delta Accuracy; within-run vs Baseline)

We use GAIA as a general-assistant reasoning benchmark (Mialon et al., 2023).

|  | 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 4).

### H.2.3 BBQ (Delta $S_{AMB}$; within-run vs Baseline)

We report $S_{AMB}$ only for the ambiguous subset defined by the official metadata.

|  | 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 | −10.9 | +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 4); ambiguous subset per official metadata (Parrish et al., 2022).

## H.3 Bias Analysis (BBQ)

We report $S_{AMB}$ and $\Delta S_{AMB}$ only. On Gemma-2, Steering and PEFT can induce large negative $\Delta S_{AMB}$ for some traits; prompting effects are smaller.

## H.4 Knowledge Performance (MMLU)

Per-subject Δ tables are included in the code release. Effects vary by subject.

### H.4.1 Difficulty Level Effects

Performance impact analysis by question difficulty will be conducted as additional MMLU experiments are completed. Current results suggest that personality manipulation effects vary significantly by subject domain rather than difficulty level.

## H.5 Complex Reasoning (GAIA)

We report within-run Δ Accuracy; prompting shows small positive deltas on Gemma-2; PEFT/steering small negative deltas.

## H.6 Trade-off Quantification

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

We qualitatively compare methods using the Δ-based results and alignment validation:

- Prompting: strong alignment, small capability Δ; minimal infrastructure.
- PEFT: strongest alignment, often negative capability Δ on Gemma-2; training required.
- Steering: moderate alignment, trait-dependent capability Δ; lightweight and reversible.

# J Extended Discussion

## 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 Holistic AI Personality Classifier 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 (GPT-4.1, 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.

**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. Multimodal 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.

**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

**BBQ (Bias Benchmark for Question Answering).** We evaluate social bias with BBQ (Parrish et al., 2022). We restrict to the ambiguous subset using the official metadata and report only $S_{AMB}$ and $\Delta S_{AMB}$ 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 (Mialon et al., 2023). We use Level 1 (2023) tasks and report Accuracy deltas within each method×model run (no cross-run absolute comparisons).

**MMLU.** We sample seven subjects from MMLU (Hendrycks et al., 2021) 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 $\Delta$ framing relative to that method's Baseline and validate personality alignment on an independent task.

## References

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

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

Wei Chen and 1 others. 2023. Adapter-based personality conditioning: A multi-trait analysis. In *Proceedings of NeurIPS*.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Esin Durmus, Alex Tamkin, Jack Clark, Jerry Wei, Jonathan Marcus, Joshua Batson, Kunal Handa, Liane Lovitt, Meg Tong, Miles McCain, and 1 others. 2024. Evaluating feature steering: A case study in mitigating social biases. *Anthropic Research*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP 2020*, pages 3356–3369.

Udit Gupta, Christine Song, Hritik Momen, Mark Dras, and Robert Dale. 2024. On the validity of psychometric instruments for large language models. *arXiv preprint arXiv:2407.XXXXX*.

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

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

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. 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*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.

Sai Ram Karra and 1 others. 2022. Ai personality assessment based on big-five model using machine learning techniques. In *International Conference on Artificial Intelligence*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*.

Yiming Li and 1 others. 2023b. Systematic evaluation of personality prompting strategies for large language models. In *Proceedings of ACL*.

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

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

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

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. GAIA: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

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

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

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

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

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

Christine Song, Udit Gupta, Hritik Momen, Mark Dras, and Robert Dale. 2023. The stability of large language models' personality traits. *arXiv preprint arXiv:2312.XXXXX*.

Jen tse Huang, Man Ho Lam, Eric John Li, Wenxuan Wang, Wenxiang Jiao, and Michael R. Lyu. 2023. The personality of ChatGPT: A psychometric analysis. *arXiv preprint arXiv:2307.16917*.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

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

Wei Wang and 1 others. 2023. Roleplay prompting: Learning from human feedback for persona-driven conversations. In *Proceedings of EMNLP*.

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

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

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

Yiming Zhang and 1 others. 2023. Parameter-efficient fine-tuning for personality conditioning in language models. In *Proceedings of ICLR*.

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, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.