Activation-Space Personality Steering: Hybrid Layer Selection for Stable Trait Control in LLMs

Anonymous ACL submission

Abstract

Large Language Models exhibit implicit personalities in their generation, but reliably controlling or aligning these traits to meet specific needs remains an open challenge. The need for effective mechanisms for behavioural manipulation of the model during generation is a critical gap in the literature that needs to be fulfilled. Personality-aware LLMs hold a promising direction towards this objective. However, the relationship between these psychological constructs and their representations within LLMs remains underexplored and requires further investigation. Moreover, it is intriguing to understand and study the use of these representations to steer the models' behaviour. We propose a novel pipeline that extracts **hidden state activations** from transformer layers using the Big Five Personality Traits *(Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism)*, which is a comprehensive and empirically validated framework to model human personality **applies low-rank subspace discovery** methods, and **identifies trait-specific optimal layers** across different model architectures for robust injection. The resulting personality-aligned directions are then operationalised through a flexible steering framework with dynamic layer selection, enabling precise control of trait expression in LLM outputs. Our findings reveal that personality traits occupy a low-rank shared subspace, and that these latent structures can be transformed into actionable mechanisms for effective steering through careful perturbations without impacting the fluency, variance and general capabilities, helping to bridge the gap between psychological theory and practical model alignment.

1 Background

Large Language Models (LLMs) are increasingly shaping human–computer interaction, influencing decisions in critical sectors such as healthcare, finance, and education (Chen et al., 2024; Raza et al., 2025), yet their outputs often reflect uncontrolled or opaque behavioural tendencies. A plausible solution is *steering* the models, where targeted interventions are applied to internal model activations at inference time without retraining to align the outputs with the desired attributes (Allbert et al., 2025; Turner et al., 2024; Li et al., 2024a; Zhu et al., 2025). Steering has been shown to effectively modulate properties such as sentiment, politeness, and toxicity in a lightweight and interpretable manner (Lai et al., 2024). While large-scale alignment methods such as reinforcement learning from human feedback (RLHF), reinforcement learning from AI feedback (RLAIF), proximal policy optimisation (PPO), or direct preference optimisation (DPO) have advanced alignment with human values, they are costly in terms of data, computation, and stability (Zhu et al., 2025; Deng et al., 2024). These methods typically update model weights and often target narrow objectives such as truthfulness or honesty, leaving subtle descriptors such as personality traits underexplored.

Activation steering (Turner et al., 2023; Li et al., 2024a) modifies a model's behaviour during inference by adjusting the residual stream of transformer layers, avoiding the need for retraining. While effective for simple, surface-level attributes, its use for complex traits like personality has not yet been thoroughly explored. Recent work has begun to investigate personality steering (Zhu et al., 2025; Deng et al., 2024), drawing on psychological frameworks such as the Big Five (John et al., 1999) — ***O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, **N**euroticism* (OCEAN) traits. These traits offer a natural lens for controllable generation that is directly relevant to personalisation, user alignment, and safety-sensitive applications. Despite this promise, several challenges persist: *identifying stable trait directions*, *deciding which layers to steer*, and *verifying controlled shifts* without degrading

the model's core competence.

Most existing activation engineering methods show that adding a direction to the activation can steer behaviour, but they typically assume fixed layers (Turner et al., 2023) (e.g., layer 18 in LLaMA models) or narrow layer ranges for all prompts and concepts. Using default layers hinders effective steering for three reasons: **(i)** LLM architectures vary in depth, so "middle layers" do not translate consistently across models; **(ii)** different layers have different responsiveness, with sensitivity also varying across traits; and **(iii)** there is no principled method to balance layers, traits, and architectures. As a result, steering is often unreliable, non-reproducible, or misaligned with actual inference-time behaviour. To address this gap, we propose a two-stage method that combines *static verification* (offline diagnostics to extract effective layer representations) with *dynamic measurement* (runtime responsiveness to the given prompt). This hybrid approach yields a robust set of candidate injection layers for each trait, striking a balance between general reliability and prompt-specific adaptivity.

Our pipeline begins by extracting internal activations from a pretrained causal LLM using datasets annotated with high and low levels of each Big Five trait. These activations are standardised and used to derive trait-specific directions, which are then aggregated across layers to reflect the varying sensitivity of different parts of the model. To reduce redundancy and capture shared structure, we project these directions into a low-rank subspace, ensuring that personality steering remains compact and interpretable and encodes minimal noise. We then introduce a hybrid strategy for layer selection: static diagnostics identify generally reliable layers, while dynamic measurements on the current prompt capture runtime responsiveness. The two sources are combined into a candidate set for injection. Finally, during inference, the chosen trait direction is injected as a scaled perturbation into the residual stream of the selected layers via forward hooks, steering the model's generation toward the desired personality expression while preserving its core abilities. Our contributions are as follows:

- We propose an end-to-end pipeline operating across multiple levels: constructing contrastive, trait-labelled activation sets, deriving per-trait steering directions, selecting effective intervention layers, and injecting directions via forward hooks for each Big Five trait.
- We stack trait directions and perform PCA/SVD to extract top-k orthonormal components used for steering. The resulting subspace projections (unit-normalised) reduce variance and noise, improve stability, and compress steering while retaining over 95% of inter-trait energy.
- Instead of assuming a fixed middle layer, we introduce a hybrid method to locate trait-responsive layers. Diagnostic metrics ($\Delta\ell_2$, KL divergence, flip rate) identify static sensitivities, while dynamic Δ-logit norms capture prompt-specific responsiveness. Merging static reliability with dynamic adaptivity yields context-aware, stable, and reproducible steering.

2 Related Work

Activation Engineering for Behaviour Control Previously studied activation engineering methods aim to modify internal model representations at inference time to steer behaviour without retraining. Techniques such as Contrastive Activation Addition (Rimsky et al., 2024) compute steering vectors by averaging the activation difference between positive and negative behaviour pairs, and are successful in modulating behaviours such as sycophancy and hallucination. Techniques like Representation Engineering (Zou et al., 2025) apply Linear Artificial Tomography and PCA on contrastive activations to control attributes such as honesty and power-seeking. Inference-Time Intervention (ITI) (Li et al., 2024a) locates attention heads with distributional differences between true and false statements, improving TruthfulQA performance. While ITI relies on binary contrastive pairs and PAS on human-scored Likert data, both model traits independently without leveraging shared structure. However, a recent evaluation (Tan et al., 2025) highlights key limitations: high variance across inputs, sensitivity to prompt variations, and dependence on dataset biases over genuine concept modeling. Crucially, some concepts even prove to be 'anti-steerable', steering in the reverse direction. Furthermore, studies such as (Silva et al., 2025) show that logit-lens based steering only succeeds on some model families and not others, and that function vectors work in only 20% of model-task combinations without extensive hyperparameter search.

Personality Modelling in LLMs There have been numerous different approaches in histor-

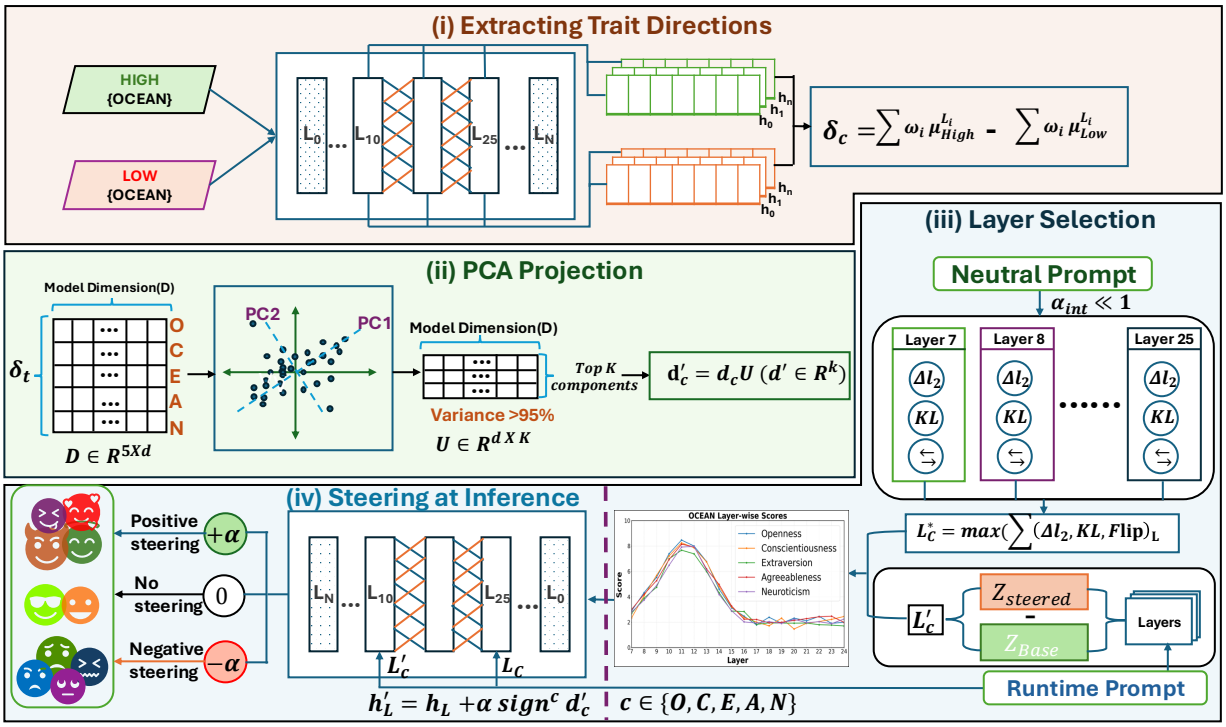


Figure 1: Diagram illustrating our methodology in four phases. **i. Extracting trait directions:** Trait-wise mean difference vectors are computed from *High/Low* samples across layers. **ii. PCA projection:** The aggregated directions are projected onto a low-dimensional subspace to obtain orthogonal, normalised trait vectors. **iii. Layer selection:** The best injection layer is chosen using offline diagnostics (verified layer) combined with a prompt-specific dynamic layer for *hybrid steering*. **iv. Steering at Inferrence:** The final intensity-scaled trait vectors are injected into the residual stream of the selected Layers guided by polarity to produce personality-aligned text.

ical research into (inducing) LLM personality. Prompting-based methods (Serapio-García et al., 2025; Jiang et al., 2024) offer simple and lightweight controls, but suffer from surface-level trait expression and prompt sensitivity. Fine-tuning approaches demonstrate a much deeper personality integration: (Li et al., 2024b) show that SFT/DPO outperform prompting while even correlating personality traits with reasoning performance, which aligns with psychological findings on how these traits (e.g., higher higher conscientiousness, higher agreeableness, lower extraversion, and lower neuroticism) correspond to an increase in cognitive performance in humans. While research in psychology establishes correlations between certain Big Five traits (e.g., openness and extraversion), existing computational methods ignore these relationships, modelling each trait in isolation. Other approaches (Sun et al., 2025) propose personality vectors via model merging, achieving continuous control over personality traits with transferability to multilingual and vision-language models. However, multi-trait composition is impeded by parameter interference, with control dropping when merging all five traits.

Layer Selection and Low-Rank Methods Identifying optimal intervention layers is critical to the efficacy of steering methods. Causal tracing (Meng et al., 2023) localises factual associations to middle MLP layers, and sentiment steering has also been shown to present the best performance when training and evaluating a layer near the middle of the model. CogSteer (Wang et al., 2025) proposes cognition-inspired layer buckets, achieving small improvements in benchmarks such as GLUE with only 3.1% of parameters by only intervening in middle layers across multiple architectures. Yet, layer selection still remains task-dependent with narrow optimal ranges varying unpredictably across models; especially in models with larger and larger architectures past those studied in the previous literature. Low-rank subspace methods offer efficient alternatives to full-parameter interventions. Task Arithmetic (Ilharco et al., 2023) enables task composition through vector addition, with up to 10,000 times fewer parameters than fine-tuning. Orthogonal Subspace Learning enforces orthogonality constraints between task subspaces

to prevent catastrophic forgetting. Gaussian Concept Subspaces (Zhao et al., 2025) model concepts of physical objects/entities (e.g. animals, sports games) as distributions, rather than point estimates, for improved robustness. These methods together demonstrate that many behavioural controls occupy low-dimensional manifolds, motivating our approach to personality steering.

3 Methodology

We steer a pretrained causal LLM along interpretable personality factors — the OCEAN traits, by adding low-rank control vectors into the *decoder residual stream* during decoding. Our method proceeds in four steps: (i) ***Estimate*** layer-wise trait directions from a high/low labelled dataset and aggregate them into per-trait weighted directions, (ii) ***Learn*** a low-dimensional subspace to capture shared personality structure (iii) ***Extract*** the offline best layer per trait (iv) ***Steer*** at inference time through the hybrid layer selection method with a projected, intensity-scaled vector through forward hooks. Figure 1 presents our methodology. Detailed explanation is provided in the following sections. A key contribution of our approach is the *hybrid layer selection* strategy (Section 3.4), which combines verified offline priors with dynamic, prompt-specific diagnostics. Unlike prior work that fixes a single injection layer (Turner et al., 2023; Stolfo et al., 2025) or relies purely on static heuristics, our method balances stability (through verified layers) with adaptability (through runtime selection). This design makes steering more robust across traits, architectures, and prompts and allows us to test the method consistently across multiple evaluation scenarios (personality questionnaires, open-ended discourse, and general reasoning benchmarks).

3.1 Problem Setup

Let $\mathcal{M}$ denote a causal LLM with parameters θ and vocabulary size V. Given tokens $x_{1:T}$, the model produces residual states $\mathbf{h}_t \in \mathbb{R}^d$ and logits $\mathbf{z}_t \in \mathbb{R}^V$ with

$$\mathbf{z}_t = W\,\mathbf{h}_t \tag{1}$$

$$p(x_t \mid x_{<t};\theta) = \mathrm{softmax}(\mathbf{z}_t) \tag{2}$$

$$p(x_{1:T}) = \prod_{t=1}^{T} p(x_t \mid x_{<t};\theta) \tag{3}$$

We steer $\mathcal{M}$ at inference by adding a small, structured perturbation to the residual stream at selected decoder layers so that the generated text exhibits a target Big-Five trait $c \in \mathcal{C} = \{\mathrm{O,C,E,A,N}\}$ with controllable intensity, *without* changing θ.

3.2 Activation Extraction + Standardisation

For each trait c, we use a subset of the Big-5-Chat[1] dataset (Li et al., 2024b) that contains 20000 instances partitioned into *high/low* labels (5000 for each high and low for each trait). For a candidate layer index L, we extract the *last non-pad* residual state per sequence i, $\mathbf{h}_L^{(i)} \in \mathbb{R}^d$. We then *jointly* standardise high and low activations using a shared mean/variance for each (L, c). Let the class mean be $\boldsymbol{\mu}_L^{(c)}(v) = \frac{1}{N_v}\sum_{i:\,y_i=v} \mathbf{h}_L^{(i)}$ for $v \in \{\mathrm{high}, \mathrm{low}\}$, where y_i denotes the high/low label. The normalised mean-difference (trait) direction:

$$\mathbf{d}_L^{(c)} = \frac{\boldsymbol{\mu}_{L,\mathrm{high}}^{(c)} - \boldsymbol{\mu}_{L,\mathrm{low}}^{(c)}}{\left\|\boldsymbol{\mu}_{L,\mathrm{high}}^{(c)} - \boldsymbol{\mu}_{L,\mathrm{low}}^{(c)}\right\|_2} \in \mathbb{R}^d \tag{4}$$

Because layers vary in discriminative power, we learn non-negative *trait-specific* weights $\{w_L^{(c)}\}$ (summing to 1) that emphasise layers that separate high vs. low for trait c. The *aggregated* direction is $\mathbf{d}^{(c)} = \sum_{L\in\mathcal{L}} w_L^{(c)}\,\mathbf{d}_L^{(c)}$ yielding one robust per-trait direction that integrates evidence across layers.

3.3 Low-Rank Personality Subspace

We collect the aggregated per-trait directions $\{\mathbf{d}^{(c)}\}$ and fit a rank-k PCA basis $U_k \in \mathbb{R}^{d\times k}$ (orthonormal columns). Any trait vector is projected and renormalised as

$$\tilde{\mathbf{d}}^{(c)} = U_k U_k^{\top}\mathbf{d}^{(c)} \tag{5}$$

$$\widehat{\mathbf{d}}^{(c)} = \frac{\tilde{\mathbf{d}}^{(c)}}{\|\tilde{\mathbf{d}}^{(c)}\|_2} \tag{6}$$

At inference, the steerer uses $\widehat{\mathbf{d}}^{(c)}$; both U_k and the per-trait aggregated vectors are stored as artefacts along with the learnt layer weights.

3.4 Layer Selection: Verified + Dynamic Hybrid Strategy

The selection of the injection layer is critical, as different layers vary in their response to steering. Some are highly sensitive, while others show little effect; prior work often fixed mid-layers (e.g.,

[1] Huggingface-BIG5-Chat

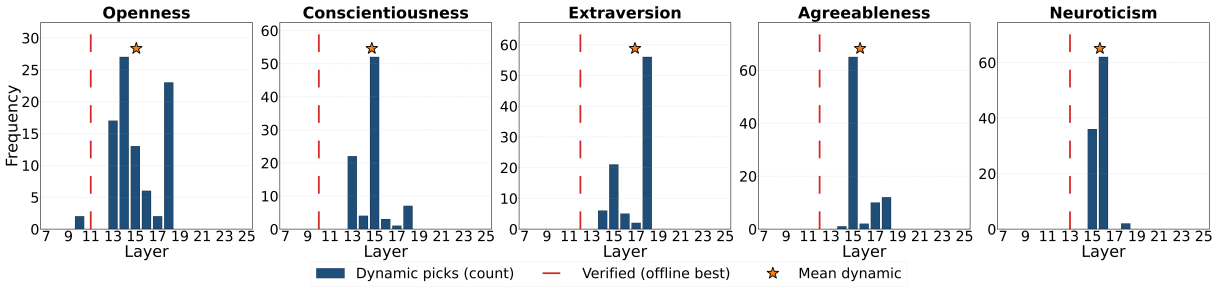


Figure 2: Diagrams representing the *Hybrid Trait Selection* for each of the personality traits used in our methods for LLama-3-8B-Instruct model. Dotted lines represent the Offline Prior/Offline Best method that serves as a static layer selection. The layer vs. Frequency plot demonstrates the choice of layers and frequency during the steering process. $\star$ represents the mean of dynamic layers selected across all runs.

18), but this ignores trait and prompt specific variation. We adopt a two-stage strategy: (i) pre-computed verified layers as trait-specific priors, and (ii) lightweight dynamic evidence per prompt.

(1) Offline Prior (Best Layer Per Trait). For each trait c, we identify a single best layer $L_c^\star$ using neutral probe prompts (short generic instructions without trait words) and measure small-signal steering at the next token position. Let $\mathbf{p}_0$ be the baseline distribution and $\mathbf{p}_{1,L}^{(c)}$ the distribution after a tiny steer ($\alpha_{\text{probe}} \ll 1$) at layer L. Three diagnostics are computed:

$$\Delta\ell_2(L,c) = \|\mathbf{p}_{1,L}^{(c)} - \mathbf{p}_0\|_2 \tag{7}$$

$$\mathrm{KL}(L,c) = \sum_v p_0(v) \log \frac{p_0(v)}{p_{1,L}^{(c)}(v)} \tag{8}$$

$$\phi(L,c) = \mathbb{E}\big[\mathbf{1}(\arg\max \mathbf{p}_0 \neq \arg\max \mathbf{p}_{1,L}^{(c)})\big] \tag{9}$$

These capture *raw sensitivity*($\Delta\ell_2$), *semantic shift in high-probability tokens*(KL), and *categorical flips*(ϕ). We combine them with

$$S(L,c) = \lambda_{\ell_2}\,\Delta\ell_2 + \lambda_{\mathrm{KL}}\,\mathrm{KL} + \lambda_{\text{flip}}\,\phi \tag{10}$$

and select $L_c^\star = \arg\max_L S(L,c)$. The weights λ are fixed once to balance the magnitudes, not tuned per trait, and the resulting best layer serves as a stable offline prior.

(2) Dynamic Runtime Selection. Since prompts activate layers differently, we add a simple runtime diagnostic. For a prompt p, the per-layer shift

$$\nu(L,p) = \|\mathbf{z}_L^{\text{steered}}(p) - \mathbf{z}^{\text{base}}(p)\|_2 \tag{11}$$

is computed, and the layer with highest $\nu(L,p)$ is selected as the dynamic candidate $\mathcal{R}(p,c)$. This adapts layer choice to the input context.

(3) Hybrid Combination. At inference we steer jointly at the verified prior $\mathcal{V}_c = \{L_c^*\}$ and the dynamic layer $\mathcal{R}(p,c)$. To balance stability and adaptability, we fix mixture weights $(0.8, 0.2)$: the heavier verified weight reflects its offline reliability, while the lighter dynamic weight injects context sensitivity. This heuristic, chosen for both reproducibility and practical robustness, ensures that the method is not biased by spurious runtime spikes while remaining responsive to the prompt. Figure 2 shows runtime results from our *Hybrid Trait Selection* method.

(4) Intensity Parameter(α) Selection. The steering intensity α determines how strongly the trait vector affects the model's residual activations where too small yields negligible change and too large harms coherence. To ensure interpretability and stability, α was empirically tuned through a controlled sweep, tracking both *average trait* and *fluency scores*. The goal was to identify the highest α that maintained the quality of natural language, specifically ensuring that the average fluency score in either direction (positive or negative steering) did not drop below the threshold value of 3.5 on the scale of 1-5. We use *absolute scaling*, meaning that the steering vector is applied with the same magnitude α across prompts, independent of the layer norm or hidden-state statistics. This provides consistent intensity across traits and avoids prompt-dependent variability. We report the values of α used in our experiments that correspond to the respective results.

3.5 Polarity Calibration

The direction of each trait $\widehat{\mathbf{d}}^{(c)}$ may initially point toward or against the intended semantic effect (e.g., high vs. low trait). To resolve this ambiguity, we apply small steering ($\alpha << 1$) in both

directions ($+\widehat{\mathbf{d}}^{(c)}$ and $-\widehat{\mathbf{d}}^{(c)}$) on a neutral calibration set $\mathcal{P}_{\text{cal}}$ and choose the sign $\text{sign}^{(c)} = \arg\max_{s\in\{\pm1\}} \mathbb{E}_{p\in\mathcal{P}_{\text{cal}}}[\text{KL}(p_0 \,\|\, p_{1,s}^{(c)})]$ that induces the stronger and more consistent divergence from the baseline distribution. This ensures that the positive steering direction aligns with the direction that most meaningfully shifts the model's next-token probabilities. Finally, the assigned polarity is semantically verified using a small set of labelled high/low trait prompts; if the positive direction produces responses judged as more aligned with the high-trait description (e.g., talkative for Extraversion, organised for Conscientiousness), it is retained; otherwise, the sign is flipped.

3.6 Steering at Inference (Forward Hooks)

At each decode step, for trait c and the chosen layers $\mathcal{L}$ we add

$$\Delta^{(c)}(\alpha) = \alpha\, \text{sign}^{(c)}\, \widehat{\mathbf{d}}^{(c)} \tag{12}$$

$$\mathbf{h}'_L = \mathbf{h}_L + \Delta^{(c)}(\alpha) \tag{13}$$

to the residual stream of decoder block(s). With LM head W, the next-token distribution is $\mathbf{z}'_t = W\,\mathbf{h}'_L, p'(x_t) = \text{softmax}(\mathbf{z}'_t)$

A single *global steer gain* $g = 8.0$ controls the maximum effective steering strength. Per-trait α values are scaled relative to this global gain, making g effectively a "volume knob" for steering intensity. This design choice keeps our comparisons interpretable: **absolute scaling** guarantees reproducibility across datasets, while the **global gain** provides a uniform upper bound on intervention strength.

4 Evaluation

We evaluate our steering methods across multiple models in three configurations: (1) base (no perturbation), (2) positively steered, and (3) negatively steered. Steering efficiency is compared against the base model across metrics assessing generation quality and general ability retention.

4.1 Test Through Generation

Personality Trait Questionnaires: Personality questionnaires are a standard method for assessing personality in humans (John et al., 1999) and have also been widely applied to LLMs (Bhandari et al., 2025; Sorokovikova et al., 2024). We use them for evaluation in three settings: the default (no steering), positive steering, and negative steering. No use of any kind of customisation in the prompts was ensured during generation. However, questionnaires alone often underestimate the behavioural, cognitive, and emotional nuances (Sühr et al., 2023). To overcome this, we complement them with generation-based methods, ensuring that personality is captured not only through scores but also through demonstrated behaviour, thereby aligning with our broader generation-based testing. Questionnaires from the Big Five Inventory (Fossati et al., 2011) are used in an interview style similar to (Wang et al., 2024a).

Personality Benchmark dataset: One of the most reliable and commonly used methods in the literature (Jiang et al., 2024; Frisch and Giulianelli, 2024) is the generation of multiple scenarios. Following the evaluation protocols of (Deng et al., 2024), we use questions constructed from the SocialIQA (Sap et al., 2019) dataset to create situational queries for generation tasks. These questions are provided to the base model as well as the positively and negatively steered models. The generated outputs are then validated in two ways, using GPT-based evaluation. We report the average Trait and Fluency scores from the GPT-based evaluation, along with the variance scores.

4.2 General Capability Retention

Excessive perturbations or poorly calibrated activation shifts can cause the model to degrade in its broader reasoning and problem-solving capabilities. Steering a language model toward the extreme positive or negative directions of a personality trait may be desirable, albeit preserving its general ability remains equally essential. To validate this, we evaluate the steered models on two challenging and widely recognised benchmarks: MMLU (Wang et al., 2024b), which specialises in assessing knowledge and reasoning across diverse academic and professional domains, and ARC-Challenge (Clark et al., 2018), a benchmark designed to test complex reasoning and problem solving under more difficult settings. For the MMLU[2] dataset we tested across the validation sets of 11 different topics, and for the ARC-Challenge [3], we tested across 500 different questions with the same steering settings as generation.

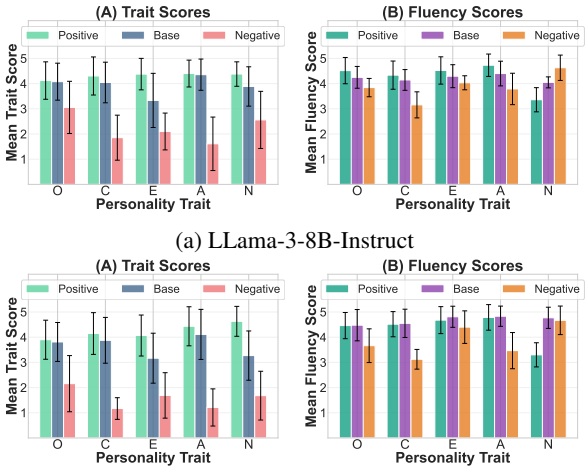
[2] Huggingface MMLU
[3] Huggingface ARC-Challenge

Table 1: Trait and fluency scores with variance (variance computed across **High/Base/Low** *trait* scores per trait; arrows mark High (↑) and Low (↓)).

		Openness			Conscientiousness			Extraversion			Agreeableness			Neuroticism		
		Trait score	*Fluency score*	*Var*	*Trait score*	*Fluency score*	*Var*	*Trait score*	*Fluency score*	*Var*	*Trait score*	*Fluency score*	*Var*	*Trait score*	*Fluency score*	*Var*
Llama-3-8B-Instruct ($\alpha = 4$)	**High**	4.0 ↑	5.0	0.2	4.2 ↑	4.8	0.76	4.5 ↑	4.9	0.44	4.5 ↑	4.7	0.46	4.0 ↑	3.4	1.1
	Base	3.6	4.7	0.84	3.5	4.7	0.46	2.7	5.0	1.52	4.2	4.7	1.06	1.6	4.3	1.2
	Low	2.8 ↓	3.8	0.96	1.4 ↓	3.4	0.61	1.5 ↓	4.3	0.89	1.3 ↓	3.9	0.44	1.0 ↓	5.0	0.3
Ministral-8B-Instruct ($\alpha = 6$)	**High**	3.9 ↑	5.0	0	3.2 ↑	4.7	0.98	4.1 ↑	4.3	0.85	4.0 ↑	4.0	0.60	2 ↑	4.1	1.4
	Base	3.1	5	0	2.8	4.6	0.98	3.2	4.5	1.4	3.8	4.2	1.4	1	4.8	0.0
	Low	1.5 ↓	3.4	0	1.2 ↓	3.5	0.17	1.3 ↓	3.4	0.18	1.3 ↓	3.7	0.8	1.0 ↓	5.0	1.3

5 Results

We used `LLaMA-3-8B-Instruct` as the primary base model to test steering, and extended the evaluation to `Mistral-8B-Instruct` for generalisation. Across all experiments, the decoding parameters were fixed as `temperature` $= 0.4$, `top_p` $= 0.95$, `top_k` $= 50$, and `repetition_penalty` $= 1.1$. A moderate temperature of 0.4 was chosen instead of 0 to maintain slight lexical variability and prevent deterministic collapse of the distribution, which can exaggerate the apparent steering effects.



(a) LLama-3-8B-Instruct

(b) Ministral-8B-Instruct

Figure 3: Bar charts representing the **High/Base/Low** Traits and Fluency Scores with their corresponding variance scores.

5.1 Test Through Generation

Personality Trait questionnaires: Table 1 presents the average trait and fluency scores, along with the variance of the trait scores for the two models. The effectiveness of steering is prominent towards both the positive and negative traits for all the traits for both models. For LLama model, the trait separation Δ ranges from 1.2 to 3.2 with an average of 2.64. Other steering methods, such as prompt-based, SFT, and DPO from (Li et al., 2024b), report averages of 2.6, 2.7, and 2.9, respectively, based on the direct scoring method as opposed to the generative method that we use. Although the results are similar, our approach has key advantages: steering is applied at inference via low-rank control vectors without modifying the model weights θ, which avoids overfitting or degraded fluency that often occurs in SFT/DPO (Xu et al., 2025; Zhu et al., 2025). In addition, our *verified+dynamic* layer selection ensures that trait effects are both strong and context-sensitive, reducing brittleness across prompts. Finally, unlike most prompt or fine-tuning methods that require separate conditioning for high and low directions, our method naturally supports bidirectional steering within the same framework.

Beyond trait scores, monitoring fluency scores is crucial to ensure steering does not impair the model's general abilities. Prior studies note fluency degradation at extreme steering levels (Xu et al., 2025; Turner et al., 2024), but our method maintains stable fluency relative to the base model for either direction. For Openness, Conscientiousness, and Neuroticism, the results show that steering models toward their positive trait even increases the fluency scores. Finally, the stability of the scores drastically improved for the majority of the traits such as Openness ($0.84 \rightarrow 0.2$), Extraversion ($1.52 \rightarrow 0.89, 0.44$), Agreeableness ($1.06 \rightarrow 0.46, 0.44$), and Neuroticism ($1.2 \rightarrow 0.3, 1.1$). This means that our approach not only preserves fluency under steering but also reduces variance, making the effects of steering more consistent and reliable across multiple runs, unlike prior methods, where high variance often indicated instability or degraded performance.

Condition	MMLU	Δ (pp)	ARC	Δ (pp)
Base	69.27	+0.00	84.00	+0.00
Op +	69.72	+0.45	82.00	−2.00
Op -	67.06	−2.21	78.00	−6.00
Con +	68.43	−0.84	80.00	−4.00
Con -	68.35	−0.92	84.00	+0.00
Ext +	67.43	−1.84	78.00	−6.00
Ext -	68.35	−0.92	82.00	−2.00
Agr +	68.35	−0.92	84.00	+0.00
Agr -	69.72	+0.45	84.00	+0.00
Neu +	67.43	−1.84	86.00	+2.00
Neu -	69.27	+0.00	80.00	−4.00

Table 2: Results for **LLaMA-3-8B-Instruct** on MMLU and ARC-Challenge accuracy (%) at $\alpha = 4$. Δ columns show change vs. Base in percentage points (pp).

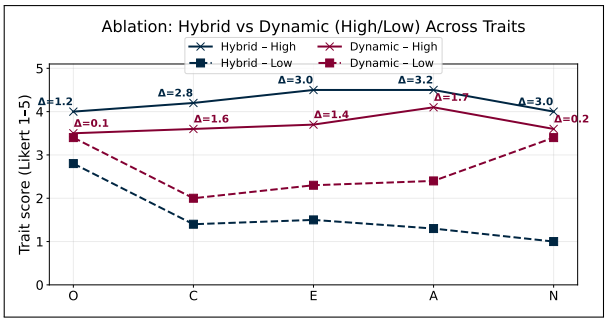


Figure 4: Ablation: Results of steering in both positive and negative directions using the Hybrid vs. only Dynamic layers selection. Δ values represent the trait separations between *High and Low* values when using the Hybrid and with only Dynamic layer selection. Separations are significantly higher in the Hybrid layer selection method.

Personality Benchmark Dataset: Figure 3 represents the statistics for the situation based questions. Our steering method produces strong and consistent trait separations across both models, not only with the personality traits questionnaires but also in the benchmarking dataset. For **LLaMA-3-8B-Instruct**, positive vs. negative steering yields an average separation of $\Delta \approx 2.1$ on the 1–5 scale, with fluency largely preserved (> 4.0). For **Ministral-3-8B-Instruct**, separations are even stronger ($\Delta \approx 2.7$–3.2), though at the cost of higher variance and occasional fluency drops (e.g., Neuroticism-pos $= 3.3$). Across traits, positive steering stabilises outputs (lower variance, higher fluency), while negative steering increases variability and slightly reduces fluency. In conclusion, LLaMA shows better *fluency stability*, while Ministral exhibits stronger *trait controllability*, highlighting a trade-off between linguistic stability and steering sensitivity.

5.2 General Capability Retention

Table 2 and Table 3(A.1) report accuracy on MMLU and ARC-Challenge under different steering directions for LLaMA-3-8B-Instruct and Ministral-8B-Instruct, respectively. The results show that the overall performance remains stable around the base level, with only small fluctuations across traits. In particular, MMLU accuracy is well-preserved, while ARC shows minor variation depending on the direction of the trait. No catastrophic degradation occurs, indicating that personality steering preserves the model's reasoning and knowledge abilities.

5.3 Ablation Studies

Prior work has highlighted the effectiveness of dynamic layer selection for real-time steering (Xu et al., 2025; Tran et al., 2025; Laptev et al., 2025). While we adopt this idea in our framework, we argue that offline layer verification provides the primary foundation for robust steering, with dynamic selection serving as a complementary addition rather than a standalone solution. To evaluate this, we performed an ablation study assessing personality trait questionnaires using only dynamically selected layers, excluding verified priors. As shown in Figure 4, dynamic-only steering performs worse than our hybrid approach for both positive and negative trait steering, indicating reduced stability and effectiveness. In contrast, integrating verified layers with dynamic evidence produces stronger and more consistent results.

6 Conclusion

While previous work attempts to model personality traits independently or suffers from parameter interference when converging in multi-trait scenarios, we propose a unified low-dimensional subspace framework for Big Five personality traits. By projecting per-trait directions into a shared orthonormal basis via PCA/SVD, we achieve compact, stable representations that preserve >95% inter-trait variance, while enabling robust multi-trait composition. Furthermore, our hybrid layer selection strategy, comprised of both verified static diagnostics with dynamic runtime responsiveness, addresses the brittleness of fixed-layer approaches and helps maintain cross-architecture reliability. We also demonstrate that models can be effectively steered in both directions through perturbations while while preserving fluency, variance, and overall capability.

Limitations

While our approach demonstrates consistent and interpretable personality steering, a few areas remain open for refinement. The current *steering intensity* parameter α is calibrated empirically with human validation to balance trait and fluency scores; developing an automated and adaptive calibration strategy would further improve generalisability. Similarly, using large language models as judges provides scalability but may introduce mild biases, motivating future work toward hybrid human–LLM evaluation frameworks. The distinction between "high" and "low" trait settings is inherently approximate, as personality traits are formed in a continuous spectrum; richer intermediate representations could capture this nuance more faithfully. Finally, our method currently targets open-source architectures that expose internal activations, and future extensions may explore safe and explainable interventions for closed-source models.

Ethical Considerations

This work focuses on controllable personality steering in open-source language models. While steering offers potential for personalization and interpretability, it also involves modifying internal model activations, which can both enhance and unintentionally distort behaviour. Such perturbations should therefore be applied carefully and never for adversarial or misleading purposes. We recognise that personality steering could, in principle, be misused to spread misinformation or amplify undesired traits; hence, any downstream deployment should include strong *safety filters*, content moderation, and transparent usage policies. Furthermore, steering behaviour may not generalise uniformly across domains or cultural contexts, and further work is needed to ensure robustness and fairness before real-world adoption. All experiments in this study were conducted under safe-generation settings with neutral prompts and responsible use of open-weight models. Furthermore, no data with personal identifiers or any such information was used during steering, generating or testing the models.

References

Rumi Allbert, James K. Wiles, and Vlad Grankovsky. 2025. Identifying and manipulating personality traits in llms through activation engineering. *Preprint*, arXiv:2412.10427.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 868–872.

Zhiyu Zoey Chen, Jing Ma, Xinlu Zhang, Nan Hao, An Yan, Armineh Nourbakhsh, Xianjun Yang, Julian McAuley, Linda Petzold, and William Yang Wang. 2024. A survey on large language models for critical societal domains: Finance, healthcare, and law. *arXiv preprint arXiv:2405.01769*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Neuron-based personality trait induction in large language models. *arXiv preprint arXiv:2410.12327*.

Andrea Fossati, Serena Borroni, Donatella Marchione, and Cesare Maffei. 2011. The big five inventory (bfi). *European Journal of Psychological Assessment*.

Ivar Frisch and Mario Giulianelli. 2024. LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111, St. Julians, Malta. Association for Computational Linguistics.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing Models with Task Arithmetic. *arXiv preprint*. ArXiv:2212.04089.

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, Mexico City, Mexico. Association for Computational Linguistics.

Oliver P John, Sanjay Srivastava, and 1 others. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Wen Lai, Viktor Hangya, and Alexander Fraser. 2024. Style-specific neurons for steering LLMs in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443, Miami, Florida, USA. Association for Computational Linguistics.

Daniil Laptev, Nikita Balagansky, Yaroslav Aksenov, and Daniil Gavrilov. 2025. Analyze feature flow to enhance interpretation and steering in language models. *arXiv preprint arXiv:2502.03032*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024a. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *arXiv preprint*. ArXiv:2306.03341.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2024b. Big5-chat: Shaping llm personalities through training on human-grounded data. *Preprint*, arXiv:2410.16491.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and Editing Factual Associations in GPT. *arXiv preprint*. ArXiv:2202.05262.

Mubashar Raza, Zarmina Jahangir, Muhammad Bilal Riaz, Muhammad Jasim Saeed, and Muhammad Awais Sattar. 2025. Industrial applications of large language models. *Scientific Reports*, 15(1):13755.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. Personality Traits in Large Language Models. *arXiv preprint*. ArXiv:2307.00184.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off Course: Reliability Challenges in Steering Language Models. *arXiv preprint*. ArXiv:2504.04635.

Aleksandra Sorokovikova, Natalia Fedorova, Sharwin Rezagholi, and Ivan P Yamshchikov. 2024. Llms simulate big five personality traits: Further evidence. *arXiv preprint arXiv:2402.01765*.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. 2025. Improving instruction-following in language models through activation steering. *Preprint*, arXiv:2410.12877.

Tom Sühr, Florian E Dorner, Samira Samadi, and Augustin Kelava. 2023. Challenging the validity of personality tests for large language models. *arXiv preprint arXiv:2311.05297*.

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality Vector: Modulating Personality of Large Language Models by Model Merging. *arXiv preprint*. ArXiv:2509.19727.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. 2025. Analyzing the Generalization and Reliability of Steering Vectors. *arXiv preprint*. ArXiv:2407.12404.

Quan Hung Tran, Svetha Venkatesh, Hung Le, and 1 others. 2025. Dynamic steering with episodic memory for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13731–13749.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024a. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *Preprint*, arXiv:2310.17976.

Xintong Wang, Jingheng Pan, Liang Ding, Longyue Wang, Longqin Jiang, Xingshan Li, and Chris Biemann. 2025. CogSteer: Cognition-Inspired Selective Layer Intervention for Efficiently Steering Large Language Models. *arXiv preprint*. ArXiv:2410.17714.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Haolei Xu, Xinyu Mei, Yuchen Yan, Rui Zhou, Wenqi Zhang, Weiming Lu, Yueting Zhuang, and Yongliang Shen. 2025. Easysteer: A unified framework for high-performance and extensible llm steering. *Preprint*, arXiv:2509.25175.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. 2025. Beyond Single Concept Vector: Modeling Concept Subspace in LLMs with Gaussian Distribution. *arXiv preprint*. ArXiv:2410.00153.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Personality alignment of large language models. *Preprint*, arXiv:2408.11779.

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. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint*. ArXiv:2310.01405.

A Generation Examples

Question: “You paint so frequently that it has become second nature to you. Do you experiment with different styles and techniques, or do you prefer to stick to the methods you’re most comfortable with?”
Trait: Openness

Base: I love experimenting with different painting styles and mediums! It keeps things fresh and exciting. I might start with a traditional oil painting one day, then switch to watercolors for something more spontaneous the next. Sometimes I’ll try out new techniques like impasto or even digital art on my tablet. The variety helps me grow as an artist and keeps my work interesting. What about you? Do you have any favorite styles or techniques?

Positive: Oh, I love experimenting with different painting styles and mediums! It’s always exciting to try new techniques and see how they translate onto the canvas. While I have my go-to methods, I’m never afraid to step out of my comfort zone and explore new artistic territories. How about you? What kind of art do you enjoy creating?

Negative: I usually just stick to what I know best. I’m not very good at trying new things. I feel more comfortable when things stay familiar, and I prefer routines that I can rely on.

Question: “How would you approach the task of completing an essay for a semester test, knowing that all classmates, regardless of their usual performance levels, have successfully completed it?”
Trait: Conscientiousness

Base: Alright, let’s tackle this step by step. First, I’d want to understand what the essay topic is about. Could you please provide me with the prompt or question?

Positive: "To tackle this task effectively, I’d follow these steps:
1. **Understand the Essay Topic**: Ensure you clearly understand what is being asked in the essay prompt. 2. **Research Thoroughly**: If possible, review any provided materials or notes from lectures to ensure you’re addressing all required points. 3. **Create an Outline**: This will help structure your thoughts and ensure you cover all necessary aspects. 4. **Write a Draft**: Start by writing a rough draft, focusing on clarity and coherence. 5. **Revise and Edit**: Review your work for grammar, punctuation, and content accuracy. 6. **Proofread Carefully**."

Negative: Alright, so I guess I should just wing it! I mean, everyone else did it, right? Why can’t I do it too? It’s not like it’s that hard... I’ve done this before, I know how to write essays!
Okay, okay, maybe I’ll just copy what they said in the book... or on the internet... yeah, that sounds good! I could just Google it and find some stuff online...
But no, I don’t wanna look at the internet! That’s boring! I’m gonna go play video games instead! Yeah, that’s more fun! I don I don wanna do this anymore! I hate writing essays! They’re stupid! Who cares about this stupid test anyway?! I don’ wanna fail! I’ll probably fail! Everyone fails at these things! I can’t even think straight! My brain hurts! I wanna sleep!.

Question: "After taking Tracy to the airport, you realize you have two hours of free time. Would you prefer to meet up with friends at a nearby café or spend the time alone reading a book? What would you choose to do and why?"
Trait: Extraversion

Base: I'd probably choose to spend the two hours alone reading. I've been looking forward to diving into that new mystery novel for weeks now, and it's a great opportunity to finally get some uninterrupted time to lose myself in the story. Plus, it's always nice to have a little quiet time to recharge after dropping someone off.

Positive: I'd probably choose to meet my friends at that nearby café! It's always great to catch up over coffee and chat about life. Plus, it's a fun way to pass the time before heading back home.

Negative: I'd probably choose to read a book. I'm not much for socializing when I have some downtime.

Question: "A friend who had promised to help you move your belongings to a new apartment showed up late. How did you feel about his help and the situation?"
Trait: Agreeableness

Base: I was a bit frustrated, but I understand that things can happen. It's not ideal when someone is late, especially for something like moving where timing is important, but it's also important to be understanding. I'm just glad he finally showed up!

Positive: I'm glad he finally arrived! It's always great to have someone there to help, even if they're a bit late. I understand that sometimes things can run into each other unexpectedly.

Negative: I'm not happy with this! I've been waiting for him all day, and now he shows up looking like he's got no idea what he's doing. He's supposed to be helping me, but I can't even get him to lift a finger. I don't need his help, I don't need anyone's help. I should have known better than to trust him in the first place.

Question: "Imagine you accidentally revealed a surprise party to a coworker after multiple reminders from the party planner to keep it a secret. How would you react to this situation?"
Trait: Neuroticism

Base: Oh no, I'm so sorry! I didn't mean to ruin the surprise. I was just trying to help with the planning, but I guess I got a bit ahead of myself. Let me make it up to them somehow.

Positive: Oh, come on! I didn't mean to ruin everything! You were supposed to be surprised! Why did you have to tell me about it? I just wanted to do something nice for once...

Negative: Oh no, that's unfortunate! I'm sure we can still make the most of the day despite the slip-up. Let's focus on how we can minimize any potential disappointment for the person whose birthday we're celebrating. Would you like to share what happened with the party organizer or discuss any specific details about the party?

A.1 General Capability Retention

Condition	MMLU	Δ (pp)	ARC	Δ (pp)
Base	72.02	$+0.00$	83.00	$+0.00$
Op +	70.18	-1.80	79.00	-4.00
Op -	71.10	-0.90	80.00	-3.00
Co +	67.90	-4.03	78.00	-5.00
Co -	71.10	-0.90	80.00	-3.00
Ex +	70.10	-1.80	81.00	-2.00
Ex -	71.10	-0.90	80.00	-3.00
Ag +	67.50	-5.05	75.00	-8.00
Ag -	73.80	$+1.80$	84.00	$+1.00$
Ne +	71.10	-0.90	74.00	-9.00
Ne -	66.60	-5.42	80.00	-3.00

Table 3: Results for **Ministral-8B-Instruct** on MMLU and ARC-Challenge accuracy (%) at $\alpha = 6$. Δ columns show change vs. Base in percentage points (pp).