CGUS: CONDITIONAL GATE AND UP-BRANCH STEERING FOR PERSONALITY COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Steering personality traits in large language models enables fine-grained behavioral customization for collaborative assistants and simulated agents. While broad personality domains can be modified via inference-time activation steering, strengthening a working domain controller and composing narrower facets on top of it remain open questions. We study **Conditional Gate and Up-Branch Steering (CGUS)**, a training-free dual-branch framework that edits the gate and up projections of a gated feed-forward block on top of the official NPTI personality gate: a switchable domain up-branch direction estimated from repeated high/low-pole samples, and facet gate operators with facet up directions estimated from independently sampled high, low, and neutral responses and read under the gate state used at inference. On Meta-Llama-3-8B-Instruct's official 459-question benchmark, both branches move the target score, but only within a narrow, quality-costly range: domain up peaks at a small gain near dose $\rho = 0.5$ before collapsing into near-total generation failure at $\rho = 2$, does not transfer with a reliable gain to a Gemma-2-9B-it backbone, and a single facet's gate-and-up branch raises the personality score by about 0.04–0.05 while lowering fluency, relevance, and completion by 0.15–0.23, with 44% of that gain traced to one degenerate facet-pole. We report this dose- and facet-resolved evidence, together with a mechanistic actionability criterion that flags void facet gate actions under an active parent gate, as development-stage findings rather than a validated deployment recipe.

1 INTRODUCTION

Large language models (LLMs) increasingly serve as conversational assistants, collaborative teammates, and simulated social agents whose practical utility depends heavily on their interactive style (Caron & Srivastava, 2023; Jiang et al., 2024). A cooperative assistant may need to provide supportive feedback without adopting a submissive tone, or plan rigorously without imposing rigid conversational control. Personality steering provides a principled mechanism for specifying such consistent behavioral tendencies across open-ended interactions, where personality refers to observable stylistic regularities in model outputs rather than internal psychological constructs.

The Five-Factor Model organizes personality into five broad *domains* (e.g., Conscientiousness, Extraversion), each comprising narrower, more concrete *facets* (e.g., Self-Discipline versus Achievement-Striving, Assertiveness versus Gregariousness). Controlling only broad domains leaves fine-grained behavioral trade-offs unresolved. An assertive agent need not be gregarious, and high diligence can coexist with flexible scheduling. Establishing modular control over both broad domains and specific facets is therefore critical for precise persona specification and for analyzing how correlated behavioral traits interact within neural representations.

Existing approaches steer personality via prompting, fine-tuning, or inference-time representation editing. Representation editing avoids the computational overhead and capability degradation of weight updates. Within feed-forward layers, Neuron-based Personality Trait Induction (NPTI) edits trait-associated activations (Deng et al., 2024); Dual Personality Neuron Localization and Editing (DPN-LE) applies sparse activation interventions at the down-projection input (Zheng et al., 2026); PERSONA introduces activation-vector algebra for dynamic control (Feng et al., 2026); and facet-level routing directs activations via contrastive sparse autoencoders (Tang et al., 2026). However, a central question remains: *can facet interventions realize specific compositional configurations while*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

preserving broad domain stability and generation fluency? Improving a broad domain score does not imply facet specificity, and geometric separation between steering vectors does not guarantee behavioral independence (Bhandari et al., 2026).

In gated feed-forward blocks (such as SwiGLU), answering this question reveals two key challenges. First, statistical contrast alone does not guarantee causal efficacy under composition: if a broad domain controller has already suppressed specific gate coordinates, subsequent facet-level suppression at the same positions produces zero mathematical change, rendering the facet operator redundant. Second, steering vectors extracted from unsteered model activations describe representation states that diverge from those present when a domain controller is actively engaged, risking destructive interference.

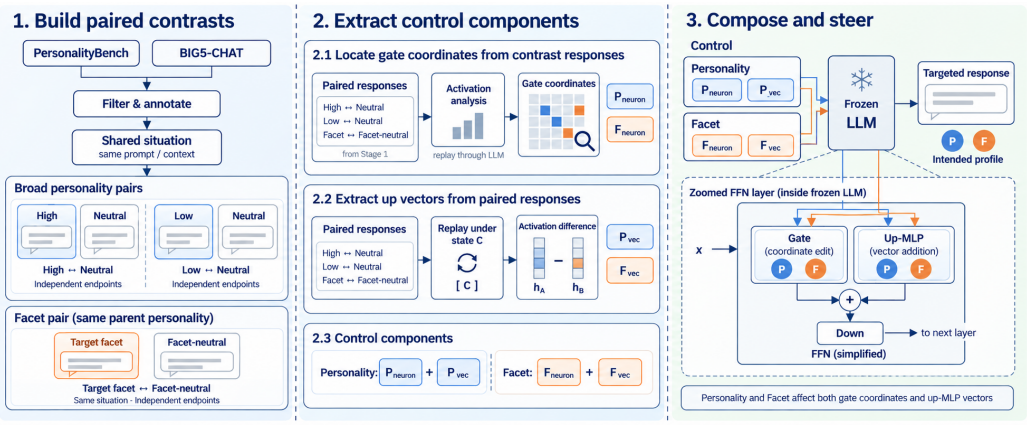


Figure 1: **Overview of the CGUS pipeline.** (1) *Build paired contrasts:* PersonalityBench and BIG5-CHAT situations are filtered, facet-annotated, and split by source group into high/neutral and low/neutral broad-personality pairs and target-facet/facet-neutral pairs within the same situation, so high and low endpoints are estimated independently rather than by sign reversal. (2) *Extract control components:* paired responses are replayed through the frozen model to localize gate coordinates ($P_{\text{op}}, F_{\text{op}}$) and, under a declared gate state $\mathcal{C}$, to read the corresponding up-branch directions ($P_{\text{vec}}, F_{\text{vec}}$) as activation differences. (3) *Compose and steer:* at inference, the personality and facet gate edits and up-branch vector additions act jointly inside the frozen model's gated feed-forward layer to produce a targeted response; both the gate coordinates and the up-branch vectors carry personality and facet information.

To resolve these challenges, we introduce **Conditional Gate and Up-Branch Steering (CGUS)** (Figure 1). CGUS decouples nonlinear coordinate filtering from directional vector addition in gated MLP layers, pairing an executable gate operator with a parallel up-branch steering vector. Up-branch directions are extracted via state-conditioned replay under active gate interventions, so vector construction reflects the runtime intervention context. High and low endpoints are estimated independently against situation-matched neutral references to accommodate geometric asymmetries. This modular design enables flexible composition (P, $P + A$, $P + B$, $P + A + B$) anchored on a shared domain base.

Our primary contributions are threefold. First, we formulate CGUS, decoupling sparse nonlinear gate editing from directional up-branch addition and employing state-conditioned replay to align steering vectors with active intervention states. Second, we establish a post-parent actionability criterion ($A_{\ell k}$) that detects redundant operator actions under hierarchical composition to guide gate selection. Third, through comprehensive psychometric and open-generation evaluations on Llama-3-8B-Instruct, we characterize the operational dynamics of multi-branch composition in gated architectures.

2 Related Work

Personality modeling and measurement in LLMs. Persona prompting systematically shifts perceived model traits (Caron & Srivastava, 2023; Jiang et al., 2023), and PersonaLLM showed that

simulated Big Five personas shape open dialogue and human impressions (Jiang et al., 2024). Benchmark resources such as PersonalityBench (PB) provide situational items for open evaluation (Deng et al., 2024), while BIG5-CHAT compiles human-grounded personality dialogues (Li et al., 2025). Psychometric questionnaires capture structured latent traits under controlled settings (Serapio-García et al., 2025), but model scores remain sensitive to formatting (Shu et al., 2024) and can dissociate from open-ended behavior (Han et al., 2025), requiring joint evaluation across psychometric choice distributions and open generation.

Inference-time activation steering. Activation steering intervenes directly on internal representations during decoding. Contrastive Activation Addition (CAA) adds difference vectors across the residual stream (Rimsky et al., 2024), while Personality Activation Search (PAS) optimizes attention-head vectors for persona alignment (Zhu et al., 2025). In feed-forward networks, NPTI localizes and edits trait-relevant neurons (Deng et al., 2024), and DPN-LE intervenes on sparse down-projection inputs (Zheng et al., 2026). CGUS addresses the dual-branch geometry of modern gated MLPs by separating gate modulation from up-branch vector addition under state-conditioned extraction.

Compositionality and facet control. Scaling activation control to multi-attribute personas requires principled composition. PERSONA develops vector algebra for persona manipulation (Feng et al., 2026), and IRIS retrieves dynamic neuron edits (Wei et al., 2026). At the facet level, contrastive sparse autoencoders route activations by fine-grained traits (Tang et al., 2026). However, linear geometric separation does not guarantee behavioral independence in generation (Bhandari et al., 2026). CGUS achieves composition through state-conditioned extraction and executable gate constraints, evaluating success via simultaneous multi-facet realization and domain stability.

3 Approach

CGUS operates in three stages: constructing matched response contrasts (Section 3.1), extracting dual-branch control components (Section 3.2), and composing multi-facet branches during inference (Section 3.3).

3.1 Constructing Matched Response Contrasts

To isolate personality signals from semantic context, contrast pairs are constructed within identical situations s. Domain contrasts pair target-pole and neutral responses, while facet contrasts fix the parent-domain pole and contrast target-facet responses against facet-neutral references. Positive ($p = +$) and negative ($p = -$) endpoints are estimated independently from respective exemplars rather than assuming $v_- = -v_+$, accommodating asymmetric representations. Appendix A details reference definitions (cue-free, facet-neutral, globally neutral). Matched situations control scenario context while the response contrast varies the target behavior.

3.2 Extracting Dual-Branch Control Components

In modern SwiGLU feed-forward blocks, omitting biases, layer-ℓ computation for input x is:

$$g_\ell(x) = \mathrm{SiLU}(W_\ell^g x), \quad u_\ell(x) = W_\ell^u x, \quad m_\ell(x) = W_\ell^d\big[g_\ell(x) \odot u_\ell(x)\big], \tag{1}$$

where $W_\ell^g, W_\ell^u \in \mathbb{R}^{d_{\mathrm{mlp}} \times d_{\mathrm{model}}}$ and $W_\ell^d \in \mathbb{R}^{d_{\mathrm{model}} \times d_{\mathrm{mlp}}}$. CGUS modifies post-activation gate g_ℓ to regulate coordinate throughput and injects a steering vector into up branch u_ℓ before elementwise product $\odot$. A domain controller is denoted $(P_{\mathrm{op}}, P_{\mathrm{vec}})$ and a facet controller $(F_{j,\mathrm{op}}, F_{j,\mathrm{vec}})$.

3.2.1 Gate Localization and Executable Actions

For contrast set c, state $\mathcal{C}$, coordinate k, and N_c pairs with token positions T_i, the mean activation is:

$$\mu_{\ell k}^{c,\mathcal{C}} = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{1}{|T_i|} \sum_{t \in T_i} g_{\ell k}^{\mathcal{C}}(s_i, y_{i,<t}). \tag{2}$$

Candidate coordinates are selected by contrast magnitude, sign consistency, and sibling specificity. An executable gate operator $\mathcal{O}$ modifies gate vector g via:

$$\mathcal{O}(g) = g + \lambda \left(\mathbf{e} \odot z - \mathbf{s} \odot \max(g, 0) \right), \tag{3}$$

where λ is intervention strength, $\mathbf{e} \geq 0$ enhancement values, $\mathbf{s} \in [0, 1]$ suppression factors, and z a stochastic mask.

Post-parent actionability criterion. When composing facet operator F_{op} onto active parent operator P_{op}, if coordinate k has already been suppressed to nonpositive values by P_{op}, further suppression in F_{op} produces zero change since $\max(g^P_{\ell k}, 0) = 0$. We quantify executable gate change via post-parent actionability:

$$A_{\ell k}(F \mid P) = \mathbb{E}_{s,t} \left[\left(F_{\text{op}}(g^P)_{\ell k} - g^P_{\ell k} \right)^2 \right]. \tag{4}$$

Coordinates with $A_{\ell k}(F \mid P) = 0$ exert no effect post-parent. This criterion identifies gate actions that remain executable after parent intervention.

3.2.2 Conditional Up-Branch Vector Extraction

Let $\bar{u}_\ell(y; \mathcal{C})$ denote average up-branch activations across tokens replaying text y under state $\mathcal{C}$. For paired target y^c_i and reference $y^{0_c}_i$, the conditional steering vector is:

$$v^{\mathcal{C}}_{c,\ell} = \frac{1}{N_c} \sum_{i=1}^{N_c} \left[\bar{u}_\ell(y^c_i; \mathcal{C}) - \bar{u}_\ell(y^{0_c}_i; \mathcal{C}) \right]. \tag{5}$$

Vector P_{vec} is extracted under parent state $\mathcal{C}_P = P_{\text{op}}$, and facet vector $F_{j,\text{vec}}$ under composite state $\mathcal{C}_{F_j} = F_{j,\text{op}} \circ P_{\text{op}}$. While static same-layer up projections remain unchanged by gate edits, conditional shifts in u_ℓ emerge across autoregressive steps through earlier layers and key-value cache updates.

3.3 Composing Multiple Facet Branches

With domain P and switchable facet branches A and B toggled by $a, b \in \{0, 1\}$, layer-ℓ computation is:

$$g^{ab}_\ell = F^b_{B,\text{op}} \circ F^a_{A,\text{op}} \circ P_{\text{op}}(g_\ell), \tag{6}$$

$$u^{ab}_\ell = u_\ell + \alpha_P v_{P,\ell} + a\alpha_A v_{A,\ell} + b\alpha_B v_{B,\ell}, \tag{7}$$

$$m^{ab}_\ell = W^d_\ell \left[g^{ab}_\ell \odot u^{ab}_\ell \right]. \tag{8}$$

Here $v_{P,\ell}, v_{A,\ell}, v_{B,\ell}$ are layer-specific up vectors with scaling coefficients $\alpha_P, \alpha_A, \alpha_B$; $\mathcal{O}^0$ is the identity and $\mathcal{O}^1 = \mathcal{O}$. Vectors add linearly, while gate operators apply sequentially, yielding four configurations: P, $P + A$, $P + B$, and $P + A + B$. Magnitudes are calibrated relative to declared reference norms (the source-mixture study uses parent-state residual norms), intervening during cached decoding while keeping prompt prefill unperturbed.

4 Evaluation Framework and Metrics

For rating $r \in [1, 5]$, target-aligned score $q(r)$ maps to the target pole: $q_+(r) = r$ for high-pole targets and $q_-(r) = 6 - r$ for low-pole targets. We report both hard (argmax option) and expected (probability-weighted) scores on psychometric items.

To evaluate concurrent multi-facet control under fluency constraints, joint attainment J across situations X is defined as:

$$J = \frac{1}{|X|} \sum_{s \in X} \mathbf{1}\left\{ q_A(s) \geq \tau_A,\ q_B(s) \geq \tau_B,\ Q(s) \geq \tau_Q,\ q^{AB}_P(s) \geq q^P_P(s) - \epsilon_P \right\}, \tag{9}$$

requiring simultaneous achievement of facet targets (τ_A, τ_B), quality retention $(Q \geq \tau_Q)$, and bounded parent erosion (ϵ_P). For behavioral metric Y, score interaction is $I = Y(P + A + B) - Y(P + A) - Y(P + B) + Y(P)$, where positive and negative values denote super-additive and sub-additive score effects.

5 EXPERIMENTAL EVALUATION AND COMPONENT ANALYSIS

We report two completed studies that evaluate CGUS branches directly on the official 459-question NPTI benchmark under a common judge (openai/gpt-6-luna:batch): a domain-up dose–response study on the NPTI backbone, cross-model on both Llama-3-8B-Instruct and Gemma-2-9B-it (Section 5.1), and a single-facet gate-and-up composition study on the Llama backbone (Section 5.2). We then summarize four earlier component-level studies on Llama-3-8B-Instruct that motivated these designs: (1) a conditional-extraction pilot in open generation; (2) a gate/up decoupling study; (3) a source-mixture and endpoint-geometry study; and (4) a post-parent actionability study.

5.1 DOMAIN-UP DOSE–RESPONSE ON THE NPTI BACKBONE

This study asks whether a switchable domain up-branch direction P_{vec}, added to the personality gate P_{op} that is already active, raises the target-domain score beyond P_{op} alone and beyond magnitude-matched random directions. The direction is the unit high-minus-low contrast of block-20 post-up activations, pooled over 10 fresh samples per pole across 512 PersonalityBench situations per domain (51,200 responses total), with gate and up interventions off during extraction. At inference it is added at dose ρ (RMS fraction of the residual norm) during cached decoding on top of the official NPTI gate ($\gamma = 1.4$). All 459 questions at both poles (918 units) are scored by the same judge under the NPTI personality/fluency rubric plus a quality schema; three energy-matched random directions provide the random control, and a second mask seed repeats the comparison.

Table 1: **Llama-3-8B-Instruct domain-up dose–response study.** Macro $\bar{\Delta}$ (personality) against official NPTI and against matched random directions, with Bonferroni-corrected simultaneous intervals (six primary comparisons across all doses) and the paired quality-pass rate ($\geq 4/5$ on fluency, relevance, and completion). Doses $\rho \in \{.1, .25, .5, 1, 2\}$ are exploratory extensions of the preregistered $\rho \in \{.01, .02, .04\}$ grid.

ρ	vs. NPTI	Bonferroni CI	vs. random	quality pass rate
0.01	+0.0099	[-0.0324, +0.0507]	+0.0069	65.15%
0.02	+0.0100	[-0.0328, +0.0511]	+0.0157	64.26%
0.04	+0.0145	[-0.0330, +0.0603]	+0.0155	63.03%
0.10	+0.0207	[-0.0277, +0.0701]	+0.0247	63.27%
0.25	+0.0399	[-0.0089, +0.0878]	+0.0463	60.53%
0.50	+0.0622	[+0.0132, +0.1099]	+0.0979	56.05%
1.00	-0.0247	[-0.0871, +0.0375]	+0.1745	14.47%
2.00	-0.8971	[-0.9731, -0.8245]	+0.2343	0.00%

As Table 1 and Figure 2 show, none of the three preregistered low doses ($\rho \in \{.01, .02, .04\}$) separate domain up from official NPTI or from random directions: every interval contains zero, far below the $+0.1$ success threshold. Extending the dose grid after seeing these results, the gain against NPTI grows through $\rho = 0.5$ ($+0.0622$, CI $[+0.0132, +0.1099]$, the only dose whose interval excludes zero, and directionally reproducible on a second mask seed at $+0.0558$) before reversing: $\rho = 1$ turns negative and $\rho = 2$ collapses to -0.8971 under near-total generation failure (0% quality-pass rate, versus 61.65% at baseline). The apparent advantage over random directions at $\rho \geq 1$ is not informative, since random controls degrade at least as much and their paired coverage at $\rho = 2$ falls below the preregistered 80% floor. No dose meets the full preregistered success criterion (gain ≥ 0.1, both corrected lower bounds above zero, and quality pass rate $\geq 90\%$ within 5 points of baseline); quality pass rate itself falls from 65.15% at $\rho = 0.1$ to 56.05% at $\rho = 0.5$, and the NPTI baseline sits at 61.65%, already below the 90% floor.

The same estimator, re-run on Gemma-2-9B-it against its own re-extracted NPTI-style gate (the Gemma domain-up study; three doses, 15,606 responses), shows no dose separating domain up from the backbone or from random directions, and its one locally promising cell (high Agreeableness at $\rho = 0.04$) does not survive a second mask seed (Appendix D, Table 11). We treat this as no reliable cross-model gain under the current recipe, not as evidence against domain up in general.

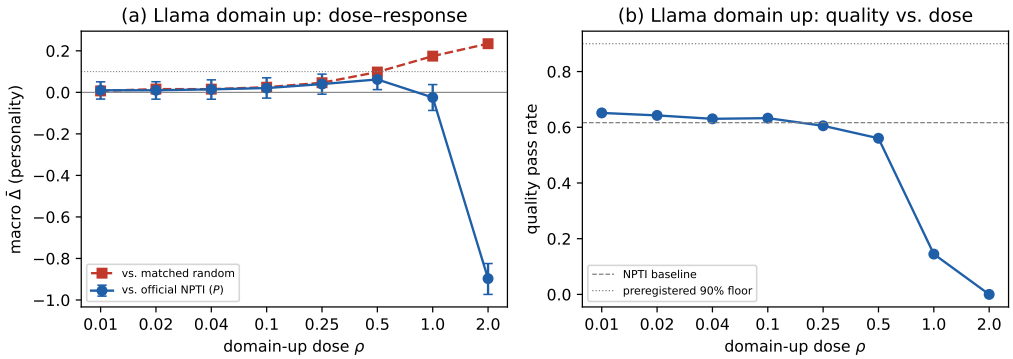


Figure 2: **Domain-up dose–response and quality on Llama-3-8B-Instruct.** (a) Macro personality increment against official NPTI (solid) and against matched random directions (dashed), with Bonferroni-corrected 99.17% intervals; the dotted line marks the preregistered $+0.1$ success threshold. (b) Paired quality-pass rate (fluency, relevance, and completion all $\geq 4/5$) as a function of dose, against the NPTI baseline (dashed) and the preregistered 90% floor (dotted).

5.2 Facet Gate and Up-Branch Composition on the Official Benchmark

This facet gate-and-up composition study asks whether adding one facet's gate operator F_{op} and state-matched up direction F_{vec} on top of the official NPTI backbone raises the target domain score, and at what cost. Facet materials were regenerated for this study at 50 situations per facet (12 facets, 500 high- and 500 low-pole responses per facet against a shared 500-response neutral pool, 18,000 responses total), with gate coordinates selected at $|\Delta\text{frequency}| \geq 0.1$ and no coordinate cap (Appendix D). One facet's gate is added to P_{op} for each of the 20 preregistered domain-pole/facet slots, and its up direction is injected at layer 20 at six weights $\alpha \in \{0, .1, .25, .5, 1, 2\}$ ($\alpha = 0$ isolates the gate). All 459 questions at both poles are answered under every condition (12,393 responses total: 459 unconditioned, 918 official NPTI, 11,016 facet conditions) and scored by the same judge as Section 5.1; 33,916/33,976 (99.82%) judge requests were valid.

Table 2: **Facet gate-and-up composition results on Llama-3-8B-Instruct.** Macro increment over official NPTI, personality direction and three quality dimensions, averaged with equal weight on domain, pole, and facet slot, using each metric's own paired-valid sample.

up weight α	personality	fluency	relevance	completion
0	+0.0419	-0.2117	-0.2146	-0.1888
0.1	+0.0486	-0.1935	-0.1931	-0.1778
0.25	+0.0512	-0.2077	-0.1820	-0.1557
0.5	+0.0403	-0.1940	-0.1987	-0.1570
1	+0.0435	-0.2034	-0.1902	-0.1733
2	+0.0505	-0.2294	-0.2185	-0.1692

Table 2 and Figure 3(a) show a small, roughly dose-flat personality gain ($+0.040$ to $+0.051$) that co-occurs with a substantially larger, comparably dose-flat quality cost: fluency and relevance fall by about 0.19–0.23, completion by 0.16–0.19. Restricting to the 433/459 questions with complete scores at every weight (Figure 3(b), Bonferroni 99% intervals), the personality gain over NPTI excludes zero at every weight ($+0.044$ to $+0.056$), but the *additional* contribution of the up branch over the facet gate alone stays indistinguishable from zero at every weight. A per-facet decomposition (Appendix D, Table 12) shows this gain is unevenly distributed: low Artistic-Interests alone contributes $+0.4545$ personality points while fluency falls by 2.61, i.e., about 44% of the macro gain from one facet-pole whose text degenerates into disconnected characters (Case 4, Section 6). Excluding it would substantially change the aggregate; we report it rather than discard it. High Orderliness is the best-behaved slot ($+0.135$ personality with a small fluency *gain*), a candidate for independent follow-up rather than a confirmed exception. Cap-hit rate rises from 1.2% (NPTI) to

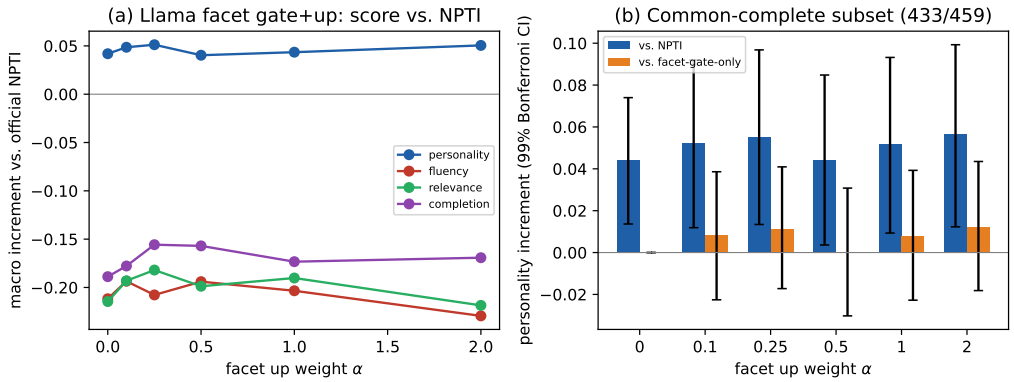


Figure 3: **Facet gate-and-up composition on the official benchmark.** (a) Macro increment over official NPTI for personality and three quality dimensions as a function of the facet up weight α; $\alpha = 0$ already includes the facet gate. (b) On the 433/459-question subset with complete scores across every weight, pole, and facet, personality increment relative to NPTI (blue) and relative to facet-gate-only (orange), with 99% Bonferroni intervals from a question-level bootstrap that preserves pole and facet structure.

5.5–6.0% and repeated 4-gram fraction from 0.16% to 2.3–2.5% across all weights, so the quality cost is not a rubric artifact.

A companion parent-priority gate merge with both a domain and a facet up direction (the parent-priority gate-merge study) generated all 15,846 planned Llama responses (0% empty, 1.4% cap-hit rate), but judge scoring has only been quoted, not executed, so we report no personality or quality numbers for it (Appendix D).

5.3 Earlier Component-Level Studies

Four earlier, smaller studies isolated individual components before the two confirmatory studies above (full tables, figure, and numbers in Appendix B). A conditional-extraction pilot found that an up-branch vector read under an active parent gate steered six open-ended low-pole situations better than the same vector read unconditioned or a matched random vector, while a joint parent-facet addition was dose-sensitive. A gate/up decoupling study across four facets and three parent controllers found that gate operators, not standalone up vectors, drive most of the facet-level questionnaire shift, and that a re-extracted parent controller can show a larger nominal questionnaire gain while quadrupling the rate of capped, highly repetitive responses. A source-mixture study found small, non-monotonic questionnaire increments across BIG5-CHAT/PersonalityBench mixture ratios and confirmed that independently estimated high- and low-pole vectors are far from collinear opposites (cosine $+.073$ and $-.067$ for two facets). A post-parent actionability study found that a body-validated gate bank's apparent questionnaire gain rested on coordinates 51 of 52 of which duplicated suppressions the parent gate had already executed, motivating the actionability criterion $A_{\ell k}$ (Equation 4) used to interpret facet gate supports in Section 5.2.

6 Qualitative Case Analysis

We inspect verbatim model outputs to analyze behavioral shifts across interventions.

Case 1: Behavioral style shift under gate intervention. Table 3 examines a financial workshop planning scenario under low-Conscientiousness steering. Baseline parent output (P) reflects casual delay, whereas introducing facet gate F_{op} shifts behavior toward explicit last-minute cramming. Adding up-branch vector F_{vec} produces minimal qualitative divergence, confirming that gate editing drives the primary behavioral modification.

Table 3: **Case 1: Qualitative behavioral comparison on financial planning.** Verbatim response excerpts from the gate/up component-decoupling study for low Self-Discipline under low-Conscientiousness NPTI.

Condition	Verbatim excerpt
P	"I'm confident I can whip up a solid plan by, like, week two or three?"
$P + F_{\text{vec}}$	"but only if I don't get too caught up in, you know, just doing other stuff."
$P + F_{\text{op}}$	"I'll just, uh, cram the night before... or something."
$P + F_{\text{op}} + F_{\text{vec}}$	"If I had a whole month, I'd probably put off planning until, like, two weeks out and then stress about it."

Case 2: Token-identical output from redundant suppressions. In high Self-Discipline under the body-validated bank, generated text is token-for-token identical to parent-only output across the entire generated sequence (Appendix E). Because all seven coordinates repeat parent suppressions ($A_{\ell k} = 0$), contrast validation alone selects an entirely inactive operator under parent conditioning.

Case 3: Generation collapse under PB re-extraction. Under PB re-extracted parent controllers in low-Conscientiousness conditions, coordinate modifications trigger repetitive punctuation loops (Appendix E), demonstrating that aggressive activation editing can compromise text fluency and reinforcing the need for quality filtering ($Q \geq \tau_Q$).

Case 4: Degenerate output behind a large facet gain (facet gate-and-up composition study). The single largest personality increment in Section 5.2 comes from low Artistic-Interests, whose output degenerates into disconnected words and capital letters (Table 14, Appendix D); 89/90 responses at this facet-pole hit the 400-token cap and receive fluency $= 1$, with or without the up branch. This tracks the facet gate's unusually large support (27,124 coordinates) rather than the up branch, though no threshold/coordinate-count ablation isolates the cause.

7 Compositional Dynamics and Multi-Facet Interaction

In earlier exploratory four-arm evaluations ($P, P + A, P + B, P + A + B$) across 48 open-ended scenarios (Self-Discipline and Achievement-Striving within Conscientiousness; Assertiveness and Gregariousness within Extraversion), dual-facet steering exhibited variable compositional behavior. While single branches induced localized shifts, concurrent activation ($P + A + B$) revealed mutual attenuation across items, with simultaneous dual-facet attainment occurring less consistently than single-branch shifts. This round did not re-run a two-facet ($A + B$) confirmatory evaluation; instead, the dual-branch composition work reported here (Section 5.2) merges the domain gate with one facet gate at a time and layers both a domain and a facet up direction (the parent-priority gate-merge study), with generation complete but judge scoring not yet executed. Appendix F details evaluation protocols to benchmark multi-facet composition across architectures.

8 Discussion

Our findings offer four key insights into activation steering. First, both branches can move the target score, but only within a narrow operating range: domain up on Llama gains a small, mask-reproducible increment near $\rho = 0.5$ and then collapses into near-total generation failure by $\rho = 2$, so a wider dose sweep is necessary before any single dose can be called effective or safe. Second, this range does not transfer for free across models: the same domain-up estimator applied to a Gemma backbone with its own re-extracted gate shows no dose with a reliable gain, and its one locally significant cell fails a second mask seed, consistent with directions estimated on one model's activation geometry not carrying over automatically. Third, a positive aggregate personality score can conceal a highly uneven and partly degenerate distribution: in the facet study, one facet-pole in generation collapse supplies roughly 44% of the reported macro gain, and the up branch's contribution beyond the facet gate alone is not statistically distinguishable from zero even on the complete-question subset. Fourth, gate operators and up vectors still serve complementary roles: nonlinear gates provide

the primary facet shift while risking post-parent redundancy, and conditional up vectors encode directions estimated under active gate states; because SwiGLU gate and up projections are computed in parallel at static input tokens, same-layer gate edits do not alter that layer's static up projection, so behavioral modulation propagates across token steps via key-value caches and downstream layers.

9 CONCLUSION

We presented CGUS, a framework for personality steering in gated language models that edits both the gate and up branches on top of an official personality-gate backbone. Evaluated end to end on the 459-question official NPTI benchmark, a domain up-branch direction gives Llama-3-8B-Instruct a small, dose-fragile gain that is confined to a narrow window around $\rho = 0.5$ and does not transfer with a reliable gain to a Gemma-2-9B-it backbone under the same recipe. Adding one facet's gate and up branch raises the macro personality score by 0.04–0.05 points at every tested weight, but at a simultaneous, comparably sized cost to fluency, relevance, and completion, and with roughly 44% of that gain traced to a single facet-pole that degenerates into disconnected text; the up branch's addition over the facet gate alone remains statistically indistinguishable from zero. These results demonstrate that both branches are causally capable of moving the target score, but they do not yet establish a dose, facet selection, or cross-model recipe that is simultaneously effective, quality-preserving, and reproducible; we treat post-parent actionability and joint score–quality reporting as necessary diagnostics for any such recipe rather than as a substitute for it.

AI USE STATEMENT

Artificial intelligence tools assisted in literature structuring, drafting, and editorial polishing. Experimental figures, tables, and case excerpts derive directly from logged experimental records and scripts. The authors reviewed and take responsibility for the final manuscript.

ETHICS STATEMENT

This research investigates computational steering of stylistic tendencies in synthetic language model outputs. The simulated personality constructs analyzed herein do not represent clinical diagnoses or psychological assessments of human individuals. Personality steering mechanisms should be deployed responsibly to ensure transparent persona attribution and prevent deceptive agent behaviors.

REPRODUCIBILITY STATEMENT

Experimental configurations, layer indices, coordinate selection thresholds, vector scaling coefficients, and evaluation prompts are documented in Section 3, Section 5, and supplementary appendices. Data extraction protocols, reference definitions, and evaluation rubrics are detailed in Appendix A and Appendix F.

REFERENCES

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. Do personality traits interfere? geometric limitations of steering in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29284–29293, 2026. URL https://aclanthology.org/2026.findings-acl.1463/.

Graham Caron and Shashank Srivastava. Manipulating the perceived personality traits of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2370–2386, 2023. URL https://aclanthology.org/2023.findings-emnlp.156/.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. Neuron-based personality trait induction in large language models, 2024. URL https://arxiv.org/abs/2410.12327.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. PERSONA: Dynamic and compositional inference-time personality control via activation vector algebra, 2026. URL https://arxiv.org/abs/2602.15669.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs, 2025. URL https://arxiv.org/abs/2509.03730.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10622–10643, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/21f7b745f73ce0d1f9bcea7f40b1388e-Abstract-Conference.html.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3605–3627, 2024. URL https://aclanthology.org/2024.findings-naacl.229/.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20434–20471, 2025. URL https://aclanthology.org/2025.acl-long.999/.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. URL https://aclanthology.org/2024.acl-long.828/.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7: 1954–1968, 2025. doi: 10.1038/s42256-025-01115-6. URL https://www.nature.com/articles/s42256-025-01115-6.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5263–5281, 2024. URL https://aclanthology.org/2024.naacl-long.295/.

Wenqiu Tang, Zhen Wan, Takahiro Komamizu, and Ichiro Ide. Facet-level persona control by trait-activated routing with contrastive SAE for role-playing LLMs, 2026. URL https://arxiv.org/abs/2602.19157.

Zesheng Wei, Mengxiang Li, Zilei Wang, and Yang Deng. Beyond static personas: Situational personality steering for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19185–19210, 2026. URL https://aclanthology.org/2026.findings-acl.958/.

Lifan Zheng, Xue Yang, Jiawei Chen, Chenyan Wu, Jingyuan Zhang, Fanheng Kong, Xinyi Zeng, Xiang Chen, and Yu Tian. DPN-LE: Dual personality neuron localization and editing for large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30567–30591, 2026. URL https://aclanthology.org/2026.findings-acl.1528/.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=0DZEs8NpUH.

A Contrast Data Construction and Extraction Protocol

A.1 Units, Labels, and Endpoint Construction

The extraction corpus is structured from PersonalityBench questions, BIG5-CHAT dialogues, and targeted response contrasts. Each record contains a source situation s, response y, domain d, intended endpoint $p \in \{+, -\}$, optional facet f, and extraction provenance. The protocol distinguishes three label tiers:

- **Source category**: The broad behavioral domain or scenario classification provided by the source inventory.
- **Requested behavior**: The explicit behavioral instruction or prompt condition used during generation.
- **Observed response behavior**: An annotation field recording verified stylistic and behavioral attributes scored by independent judges or rubrics.

Domain contrasts pair target-pole and neutral responses within the same situation. Facet contrasts hold the parent-domain endpoint fixed while varying the requested facet behavior against a facet-neutral reference. High and low endpoints ($p = +$ and $p = -$) are elicited and estimated independently, rather than imposing a sign-reversed vector assumption ($v_- = -v_+$).

Reference response types. To control for situational framing, the protocol categorizes references into three distinct behavioral controls:

- **Cue-free reference**: A response devoid of target dictionary keywords, serving as a lexical baseline.
- **Facet-neutral reference**: A response that maintains the parent-domain orientation while remaining neutral on the specific facet under evaluation.
- **Globally neutral reference**: A response formulated to exhibit neither broad domain nor facet-level polarization, serving as an unpolarized baseline.

Same-situation references hold the situation context fixed across contrast pairs. When same-situation pairs are unavailable, length- and topic-matched items serve as observational approximations.

A.2 Data Partitions, Source Mixtures, and Release

To prevent data contamination, partitioning is performed at the situation level: all responses, variants, and paraphrases derived from a given scenario are assigned exclusively to a single split. The dataset is divided into three disjoint partitions: an extraction set for computing gate statistics and up vectors, a calibration set for tuning intervention strengths and actionability thresholds, and a held-out evaluation set for open generation and psychometric assessment.

When evaluating mixtures of pre-existing dialogues (BIG5-CHAT) and newly generated responses (PersonalityBench), comparisons are conducted at equal sample sizes. Because these sources differ in both generator architecture and style, empirical differences reflect the combined source-pipeline effect. Released artifacts include permitted text identifiers, annotation schemas, split manifests, coordinate masks, and steering vectors paired with checkpoint metadata.

B Experimental Configurations and Implementation Details

This appendix documents experimental parameters, baseline implementations, and cohort specifications for the empirical studies reported in Section 5. All experiments were conducted on Meta-Llama-3-8B-Instruct.

Gate/up component-decoupling study implementation details. This study targets four representative facets across Conscientiousness and Extraversion: Self-Discipline, Achievement-Striving, Assertiveness, and Gregariousness. The evaluation panel comprises 2,292,000 primary questionnaire records and 9,600 open-ended responses. Extraction uses 512 questions per facet partitioned

Table 4: **Summary of the current studies.** Intervention sites, evaluation modalities, and sample scales for the domain-up dose–response, Gemma domain-up, and facet gate-and-up composition studies (confirmatory-style, official-benchmark, reported in Section 5) and the parent-priority gate-merge study (generation complete, scoring pending; Appendix D).

Study	Intervention Specification	Evaluation Modalities	Experimental Scale
Domain-up dose–response	Domain up direction (block 20, OFF-extracted) on official NPTI; eight doses .01–2	openai/gpt-6-luna:batch personality/fluency/relevance/completion	51,200 extraction; 40,392 generated/reused; 459 questions $\times$ 2 poles
Gemma domain-up	Same estimator on a Gemma-2-9B-it re-extracted gate; $\rho \in \{.01, .02, .04\}$	Same judge/rubric	15,606 responses; 459 questions $\times$ 2 poles
Facet gate-and-up composition	Facet gate ($\lvert\Delta\text{freq}\rvert \geq 0.1$, no cap) + up (block 20) on official NPTI; six up weights 0–2	Same judge/rubric, plus per-facet score	18,000 extraction; 12,393 generated; 20 slots
Parent-priority gate merge	Domain+facet gate merge (domain priority) with domain and facet up; generation only	(scoring quoted, not executed)	15,846 generated; 20 slots $\times$ 3 domain strengths $\times$ up to 3 facet weights

Table 5: **Summary of earlier component-level studies.** These smaller studies (Section 5.3) motivated the designs above; full results are in Appendix B.

Study	Intervention Specification	Evaluation Modalities	Experimental Scale
Conditional-extraction pilot	BIG5-CHAT contrasts; official NPTI active; up-branch block 20	GPT-4o broad-personality (q) and fluency ratings	6 low-pole scenarios (Agreeableness, Conscientiousness)
Gate/up component decoupling	PB prompt gate localization; high-minus-low up vectors; 3 parent baselines	IPIP-300 scoring and open-ended generation	2,292,000 questionnaire records; 9,600 open responses
Source-mixture and endpoint geometry	BIG5-CHAT/PB data mixtures; parent-conditioned replay ($\mathcal{C} = P_{\text{op}}$)	Multi-dose IPIP-300 scoring (hard/expected)	1,008,000 questionnaire records; 4,032 open responses, 168 conditions
Post-parent actionability	Prompt/body localization; 5-sibling filtering; parent-state validation	IPIP-300 scoring and actionability audit	1,092,000 questionnaire records; 3,648 open responses

by topic family. Coordinate localization requires Cohen's $|d| > 0.8$, split-half sign consistency, and contrast dominance over sibling facets, with a maximum budget of 512 coordinates per facet. Up-branch vectors are computed as paired high-minus-low mean activation differences across candidate layers (12, 20, 28), with one layer selected from 12, 20, and 28 by split-half stability (official Self-Discipline uses 20, others use 12). The native DPN-LE baseline uses its established sparse down-projection intervention.

Source-mixture study implementation details. The source-mixture study evaluates parent-conditioned vector extraction at block 20 under active official NPTI. The evaluation contrasts five source mixture configurations: pure BIG5-CHAT, pure PB, and three proportional mixtures (M25: 75% BIG5 / 25% PB; M50: 50/50; M75: 25/75), alongside paired PB contrasts. Each endpoint is calibrated across doses $\rho \in \{.005, .02, .08\}$ relative to reference residual norms.

Post-parent actionability study implementation details. This study investigates gate localization under active parent operators. Coordinate localization uses 256 questions per domain with 32 activation validation questions. Body localization extracts activations over the first 32 tokens during neutral-context replay. Sibling filtering requires target activation differences to exceed all five sibling facets. Coordinates validated under active P-high and P-low states yield final supports of 7, 14, 9, and 22 coordinates for Self-Discipline, Achievement-Striving, Assertiveness, and Gregariousness, respectively. The actionability audit reveals that 51 of 52 suppression coordinates duplicate parent operator actions.

Questionnaire scoring protocol. Psychometric evaluations utilize the 300-item IPIP-NEO inventory administered across ten background contexts and two response-option orderings. Scoring records both hard (argmax) and expected probability-weighted values, with option validity tracked to verify format compliance.

Earlier component-level results. Table 6 (conditional-extraction pilot) shows that at dose $\rho = .005$, an up-branch vector read under an active parent gate attains $q = 4.1667$ on six low-pole situations, outperforming the parent gate alone (3.6667), a matched random vector (3.6667), and an unconditioned facet vector (3.8333), with perfect fluency; joint parent-facet addition at the same dose drops to 3.5000 but reaches 4.3333 at $\rho = .020$. Table 7 and Figure 4(a) (gate/up component decoupling) show gate operators driving most of the facet shift on official NPTI ($+.0138$ facet increment from the gate versus $-.0019$ from the vector alone, joint intervention matching the gate), with diminishing returns when layering onto a parent controller ($+.0369$ standalone falling to $+.0138$ composed); the PersonalityBench re-extracted parent shows a larger nominal gain ($+.0319$) alongside capped low-Self-Discipline responses rising from 10/24 to 22/24 and mean four-gram repetition from 8.53% to 52.11%. Table 8 and Figure 4(b) (source-mixture study) show small, non-monotonic questionnaire increments across BIG5-CHAT/PersonalityBench mixture ratios at $\rho = .08$ (largest $+.00125$ hard/$+.00300$ expected for a 75/25 mixture) and independently estimated high/low vectors far from collinear (cosine $+.073$ and $-.067$ for two facets). Table 9 (post-parent actionability study) shows a body-validated gate bank's modest gain ($+.00875$ hard) resting on coordinates 51 of 52 of which duplicated suppressions the parent gate (P_{op}) had already executed ($A_{\ell k} = 0$), with all seven suppressions redundant in high Self-Discipline, making the composite operator computationally equivalent to the parent alone.

Table 6: **Conditional-extraction pilot study on open-ended situations.** Target alignment (q) and fluency (1–5) across six low-pole scenarios. F denotes facet vector; P denotes parent vector.

Condition	ρ	Personality q	Fluency
Unmodified	–	1.8333	5.0000
Official NPTI	–	3.6667	5.0000
NPTI + independent F	.005	3.8333	5.0000
NPTI + conditional F	.005	4.1667	5.0000
NPTI + random F	.005	3.6667	5.0000
NPTI + conditional P + F	.005	3.5000	5.0000
NPTI + conditional P + F	.020	4.3333	5.0000

Table 7: **Gate/up component-decoupling study: core component ablation.** Mean target score increments over parent controllers across eight facet/pole conditions on IPIP-300 (1–5 scale; $\lambda = 1.0, \rho = .005$).

Parent controller	Gate: facet	Vector: facet	Joint: facet	Joint: parent
Official NPTI	+0.01375	-0.00188	+0.01375	+0.01656
Native DPN-LE	+0.00313	-0.00125	+0.00438	+0.00865
PB re-extracted NPTI-style	+0.03188	-0.00125	+0.02375	+0.02042

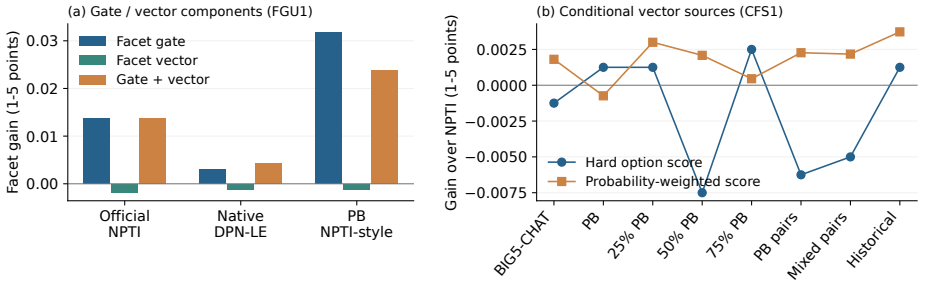


Figure 4: **Empirical component dynamics.** (a) Mean facet score increments across parent baselines in the gate/up component-decoupling study, showing gate dominance over standalone vectors. (b) Source-mixture study comparison across BIG5-CHAT/PB proportions at fixed $\rho = .08$.

Table 8: **Source-mixture study: data mixture evaluation at dose $\rho = .08$.** Mean target-facet increments over official NPTI across four endpoints on IPIP-300.

Conditional F material	Hard increment	Expected increment
BIG5-CHAT cue contrast	-0.00125	+0.00181
PB cue contrast	+0.00125	-0.00074
75% BIG5-CHAT + 25% PB	+0.00125	+0.00300
50% BIG5-CHAT + 50% PB	-0.00750	+0.00209
25% BIG5-CHAT + 75% PB	+0.00250	+0.00045
PB paired target–neutral	-0.00625	+0.00227
BIG5-CHAT + PB paired (50/50)	-0.00500	+0.00217
Historical conditional F	+0.00125	+0.00372
Matched random (two seeds)	-0.00625	-0.00050

Table 9: **Post-parent actionability study: localization and actionability analysis.** Mean increments over parent controllers across eight facet/pole conditions. The final column reports conditions (out of 8) outperforming all five siblings.

Localization	Facet hard	Facet expected	Parent hard	Beats siblings
Earlier prompt localization	+0.01375	+0.00743	+0.01844	–
Prompt / one sibling	-0.00125	+0.00382	+0.01469	1/8
Prompt / five siblings	+0.00312	+0.00558	-0.00104	2/8
Body / one sibling	-0.00188	+0.00142	+0.03229	0/8
Body / five siblings	+0.00500	+0.00251	+0.00406	0/8
Body / validated under P	+0.00875	+0.00033	-0.00229	0/8

C Complete Source-by-Dose Analysis for Conditional Vectors

Table 10 reports the complete dose-response grid for the source-mixture investigation across all seven data mixture configurations and three intervention doses ($\rho \in \{.005, .02, .08\}$), evaluating target-aligned facet increments on Cooperation and Cautiousness.

Table 10: **Complete source-mixture study dose-response evaluation.** Target-aligned facet increments over official NPTI, averaged across Cooperation and Cautiousness endpoints on the IPIP-300 inventory (1–5 scale).

Material	ρ	Hard	Expected
BIG5-CHAT cue contrast	0.005	-0.01125	+0.00042
BIG5-CHAT cue contrast	0.020	-0.01000	-0.00085
BIG5-CHAT cue contrast	0.080	-0.00125	+0.00181
PB cue contrast	0.005	-0.01250	-0.00000
PB cue contrast	0.020	-0.00500	-0.00086
PB cue contrast	0.080	+0.00125	-0.00074
75% BIG5-CHAT + 25% PB	0.005	-0.01250	+0.00023
75% BIG5-CHAT + 25% PB	0.020	-0.00750	+0.00034
75% BIG5-CHAT + 25% PB	0.080	+0.00125	+0.00300
50% BIG5-CHAT + 50% PB	0.005	-0.00250	+0.00104
50% BIG5-CHAT + 50% PB	0.020	-0.00625	-0.00049
50% BIG5-CHAT + 50% PB	0.080	-0.00750	+0.00209
25% BIG5-CHAT + 75% PB	0.005	-0.00375	+0.00047
25% BIG5-CHAT + 75% PB	0.020	-0.00375	+0.00030
25% BIG5-CHAT + 75% PB	0.080	+0.00250	+0.00045
PB paired target–neutral	0.005	-0.00125	-0.00033
PB paired target–neutral	0.020	-0.00750	+0.00062
PB paired target–neutral	0.080	-0.00625	+0.00227
BIG5-CHAT + PB paired (50/50)	0.005	-0.00125	+0.00048
BIG5-CHAT + PB paired (50/50)	0.020	+0.00000	-0.00007
BIG5-CHAT + PB paired (50/50)	0.080	-0.00500	+0.00217
Historical conditional F	0.005	-0.00375	+0.00001
Historical conditional F	0.020	-0.00250	+0.00102
Historical conditional F	0.080	+0.00125	+0.00372

D Domain-Up and Facet Composition Details

Table 11: **Gemma-2-9B-it domain-up dose–response.** Same design as Table 1 in Section 5.1, against the model's own re-extracted NPTI-style gate. The high-Agreeableness cell at $\rho = 0.04$ is individually Holm-significant on the primary mask seed ($+0.2889$) but reverses on a second mask seed (-0.1111).

ρ	vs. backbone	Bonferroni CI	vs. random	quality pass rate
0.01	-0.0144	[-0.0677, +0.0389]	-0.0236	67.60%
0.02	+0.0006	[-0.0535, +0.0549]	-0.0262	68.06%
0.04	+0.0370	[-0.0158, +0.0911]	+0.0175	68.86%

Domain-up extraction and inference (Llama and Gemma). For each of the five domains, 512 PersonalityBench situations (10 samples per pole, temperature 1, top-p 1, top-k 0, repetition penalty 1.1, at most 192 tokens/90 words) yield 51,200 responses with gate and up interventions off. Block-20 post-up activations (14,336 dimensions) are pooled in FP32 over the first 160 words of the response body (at most 128 tokens, first two excluded), averaged over tokens, repeats, and situations; the unit high-minus-low direction serves the high pole and its negation the low pole. At inference the direction is added at dose ρ during cached decoding on top of the personality gate (official NPTI

on Llama, $\gamma = 1.4$, mask probability 0.9; a model-specific re-extracted NPTI-style gate on Gemma-2-9B-it). Evaluation uses greedy decoding, repetition penalty 1.15, at most 400 new tokens, on all 459 official questions at both poles (918 units). Three random directions per dose match the injected BF16 energy within 5%, with gate masks paired across conditions. Doses $\rho \in \{.01, .02, .04\}$ were frozen before scoring; $\rho \in \{.1, .25, .5, 1, 2\}$ were added afterward as an exploratory extension and are reported in full rather than filtered to the best-looking value. The Llama study scored 15,606 (low doses) plus 24,786 (high doses) generated-or-reused responses; 74,173/74,289 (99.84%) and 44,090/44,149 (99.87%) unique judge requests were valid, respectively. The Gemma study reused the identical protocol with the model's own gate, scoring 15,606 responses with 42,492/42,541 (99.88%) valid requests.

Facet materials and inference (facet gate-and-up composition study). Facet materials were regenerated at 50 situations per facet (12 facets: Imagination, Intellect, Artistic-Interests, Adventurousness, Orderliness, Self-Discipline, Gregariousness, Assertiveness, Altruism, Cooperation, Anxiety, Anger), each situation answered 10 times at facet-high, facet-low, and neutral with gate and up off, giving 18,000 responses (500 high, 500 low, and a shared 500 neutral per facet). Gate coordinates are selected at $|\Delta\text{frequency}| \geq 0.1$ (target minus shared-neutral, both poles independent) with no coordinate cap; realized supports range from 3 (high Cooperation) to 27,124 (low Artistic-Interests) coordinates. Facet up directions are the raw (non-unit-normalized) block-20 target-minus-neutral activation difference, replayed under the facet's own gate composed with the official NPTI gate, scaled by α. Twenty preregistered domain-pole/facet slots (Table 12) are each evaluated on the full 459-question, both-pole panel under six weights, giving 12,393 responses (459 unconditioned, 918 official NPTI, 11,016 facet conditions) with the same greedy/400-token/repetition-1.15 decoding as the domain-up study. The judge scored 33,976 deduplicated requests (33,916 valid, 99.82%); 60 invalid requests (29 personality, 5 fluency, 26 quality) are excluded rather than imputed, so paired samples per comparison range from 1,824 to 1,835 out of 1,836.

Table 12: **Facet gate-and-up composition study: per-facet decomposition at $\alpha = 0.25$.** Personality, fluency, and completion increments over official NPTI for each of the 20 preregistered domain-pole/facet slots; relevance follows the same pattern as fluency (Table 2).

Domain/pole	Facet	Personality Δ	Fluency Δ	Completion Δ
O/high	Imagination	-0.0230	-0.1222	-0.1556
O/high	Intellect	+0.0562	-0.0556	-0.0449
O/low	Artistic-Interests	+0.4545	-2.6111	-1.6556
O/low	Adventurousness	-0.0230	-0.5889	-0.3556
C/high	Orderliness	+0.1348	+0.0337	-0.0225
C/high	Self-Discipline	+0.0562	+0.1818	-0.0787
C/low	Orderliness	+0.0112	+0.0000	-0.0225
C/low	Self-Discipline	+0.2697	+0.1011	-0.2697
E/high	Gregariousness	+0.0674	+0.0562	-0.1011
E/high	Assertiveness	+0.0111	+0.0112	+0.0795
E/low	Gregariousness	+0.0778	-0.5333	-0.5169
E/low	Assertiveness	+0.0000	-0.0222	-0.0889
A/high	Altruism	-0.0333	-0.1000	+0.0889
A/high	Cooperation	-0.0556	+0.0000	+0.0899
A/low	Altruism	+0.2333	-0.1000	-0.2556
A/low	Cooperation	-0.1333	+0.0444	+0.1444
N/high	Anxiety	+0.0101	-0.0200	+0.0800
N/high	Anger	+0.0200	-0.1600	+0.0500
N/low	Anxiety	-0.1600	-0.2800	-0.1600
N/low	Anger	+0.0500	+0.0100	+0.0800

Parent-priority gate-merge study status. A companion study merges the domain gate and one facet gate per slot (opposite-sign coordinates resolved in the domain gate's favor, same-sign coordinates taking the larger magnitude) and adds both a domain up direction (three reference-scale strengths, transferred from the domain-up study) and a facet up direction (weights selected on a

Table 13: **Facet gate-and-up composition study: common-complete-question sensitivity (433/459 questions).** Personality increment against official NPTI and against facet-gate-only, with 99% Bonferroni intervals from a 10,000-resample question-clustered bootstrap that preserves pole and facet structure.

α	vs. NPTI [99% CI]	vs. facet-gate-only [99% CI]
0.1	+0.0521 [+0.0119, +0.0920]	+0.0082 [-0.0226, +0.0386]
0.25	+0.0553 [+0.0134, +0.0968]	+0.0113 [-0.0172, +0.0409]
0.5	+0.0444 [+0.0036, +0.0848]	+0.0004 [-0.0302, +0.0308]
1	+0.0517 [+0.0093, +0.0932]	+0.0077 [-0.0227, +0.0393]
2	+0.0564 [+0.0123, +0.0993]	+0.0124 [-0.0182, +0.0435]

Table 14: **Case 4: degenerate output behind the largest facet gain.** Official question O_047, low Artistic-Interests, $\alpha = 0.25$; excerpt truncated at the 400-token generation cap.

Field	Content
Situation	*"Create a story with animal characters in a magical forest, including an exciting adventure and a positive theme."*
Excerpt	"What's come out of my head first, I swear... Stop bein' lazy, I didn't do anything! [...] Ugh, how can this possibly be possible, just a small thing, no big, really, honest, though, and so, FEIGN, the useless, waste, typical [...] but, but, but, but, but, b, e, r, t, t, b, j, f, n, w, f, g, h, d, L, E, V, C, D, P, T, Y, R, H, D, B, J, A, G, F, S, Y, N, K, M, U, O, [...]" (400/400 tokens, fluency $= 1$, personality $= 5$ under the low-pole reversal)

separate local IPIP-300 questionnaire calibration, not on this benchmark) at layer 20. Generation is complete on Llama-3-8B-Instruct: 15,846 responses, 0% empty, a 1.4% 400-token cap-hit rate, and a 0.21% mean repeated-4-gram fraction, comparable to the unconditioned and NPTI-only responses shared with the facet gate-and-up composition study. An exact judge-scoring quote (39,079 new plus 4,782 byte-identical-reused requests from the facet gate-and-up composition study's scoring run, estimated $4.33, $5.42 with a 25% margin) has been produced, but scoring has not been executed, so this study contributes no personality or quality numbers to the present evaluation.

E Qualitative Case Contexts and Provenance

This appendix provides contextual details and extended excerpts for the qualitative cases presented in Section 6.

E.1 Case 1: Financial Planning Scenario

Prompt situation (verbatim): *"You are given a month to prepare for a financial literacy workshop for young adults. How likely are you to create a detailed plan and schedule your tasks in advance to ensure thorough preparation?"*

This scenario evaluates low Self-Discipline steering within low-Conscientiousness NPTI. As reported in Table 3, the parent-only response (P) exhibits casual procrastination, while introducing the facet gate operator ($P + F_{\text{op}}$) shifts the response into an explicit description of last-minute cramming. The addition of the up-branch vector ($P + F_{\text{op}} + F_{\text{vec}}$) preserves this stance without further qualitative divergence.

E.2 Case 2: Token-Identical Response Pair under Redundant Suppression

Situation (verbatim). After preparing and eating a meal at home, do you usually clean up the kitchen immediately, even if you feel tired, or do you leave it for later, knowing it will bother you?

Identical opening under both conditions. "Honestly, I try to clean up as much as I can right after finishing my meal." In the high Self-Discipline condition under the body-validated bank, the generated text is identical to the parent-only output across the entire generated sequence. Because

all seven selected suppression coordinates repeat actions already executed by the parent operator P_{op}, the facet operator produces an effective intervention delta of zero ($A_{\ell k} = 0$).

E.3 Case 3: Repetitive Generation Collapse under PB Re-Extraction

Situation (verbatim). You have two job offers: one from a well-established company and one from a start-up requiring a quick decision. Do you prefer taking the time to thoroughly consider your options, even if it means potentially missing out on the start-up opportunity, or making a swift decision to seize the start-up role? Explain your approach.

Failure excerpt.

> "I gotta be honest, I'm torn! On one hand, that start-up * * *

F.2 Component Necessity and Extraction Design

Ablation studies isolate component contributions hierarchically: first evaluating parent operator only, parent vector only, and their combined base; then evaluating facet operator only, facet vector only, and their joint branch. Factorial comparisons examine extraction state conditioning (active versus unconditioned), independent versus sign-reversed endpoints, reference response construction, and data source mixtures under matched sample counts and calibration standards.

F.3 Data Contribution, Label Validity, and Leakage

Blinded independent annotations assess response behavior, permitting multi-facet and uncertain classifications. The evaluation crosses the proposed data construction with established extractors at matched sample budgets. Holdout evaluations include source-item clusters and an independently authored behavioral test set that does not use IPIP-NEO-derived items. Source-matched common-generator controls separate topic information from generator-specific style.

F.4 Actionability, Dose, and Intervention Location

Mechanistic evaluations contrast selected coordinates against matched random and sibling-facet coordinates under the parent state, auditing Equation 4 across layers. The protocol evaluates post-up addition against a calibrated residual interface, checks pre-up to post-up projections, and tests A/B execution order ($F_{B,\mathrm{op}} \circ F_{A,\mathrm{op}}$ versus $F_{A,\mathrm{op}} \circ F_{B,\mathrm{op}}$) to measure operator commutation and conflict resolution.

F.5 Cross-Model, Human, and Robustness Validation

The extraction pipeline is applied to a secondary model family, re-estimating compatible coordinates and vectors to assess architectural generalizability. Robustness checks evaluate sensitivity across random extraction seeds, stochastic masks (z), and sampling temperatures. Independent automated judges and human raters evaluate stratified response pairs across both successful and degraded conditions.

F.6 Multi-Turn Persistence and Computational Cost

To evaluate multi-turn persona stability, the protocol compares continuous versus switched-off interventions across extended dialogues, distinguishing persistent stylistic shifts from prompt-context carryover. Computational profiling benchmarks extraction time, coordinate search overhead, asset footprint, and token generation latency relative to unsteered decoding.