TRAIT: SELECTIVE AND ROBUST ACTIVATION-SPACE INTERVENTION FOR PERSONALITY TRAIT CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly used to simulate human behavior, enabling social simulation experiments. In such experiments, researchers typically control the personality of these models through prompting, fine-tuning, or activation steering. However, existing methods still fall short in the selectivity and robustness of control, either perturbing unrelated dimensions or failing to sustain the preset level across contexts, thereby undermining experimental validity. To this end, we propose TRAIT, an automated framework for selective and robust personality control. On the extraction side, Direction Extraction from Deliberation Spans (DeliSpan) takes items from the psychometric scale of the target construct as seeds, has the target model synthesize its own decision scenarios, reason through them, and make decisions, and extracts direction vectors solely from the reasoning segments in which the model reveals its own preferences while weighing trade-offs. On the injection side, Dynamic Injection with Context-Adaptive Dosing (CtxDose) trains a dose network, supervised by the deviation between post-injection scale scores and preset scores, to predict the injection strength based on context. Evaluated on the IPIP-300 inventory within the psychometric framework of Serapio-García et al. (2025), TRAIT sets a new representation engineering state of the art in selectivity and robustness. These suggest personality can be controlled precisely and stably, providing a foundation for variable control in social simulation experiments.

The source code will be released.

1 INTRODUCTION

Large language models are increasingly used to simulate human behavior, enabling social simulation experiments (Argyle et al., 2023; Aher et al., 2023; Park et al., 2023). Their validity requires selective and robust control of independent variables. **Selectivity** means that steering one trait leaves the other traits unchanged beyond what measurement noise and their correlations with the target trait can explain, corresponding to confounding in experimental validity theory (Campbell & Stanley, 1963; Shadish et al., 2002). **Robustness** means that a preset adjustment level takes effect consistently across behavioral contexts (Shadish et al., 2002). A recent systematic review finds that generative social simulations have extremely low validity and can scarcely yield reliable experimental results for social science research (Larooij & Törnberg, 2026). One important reason is that personality control lacks selectivity and robustness, although personality is one of the most commonly used independent variables in simulated social science experiments (Argyle et al., 2023; Park et al., 2023). We therefore urgently need an automated method for selective and robust control of model personality, and we present TRAIT to fill this gap.

Existing work controls model personality through three approaches. Prompting is the least costly, with graded descriptions producing ordered personality changes (Serapio-García et al., 2025). However, unclear boundaries between personality dimensions in natural language make specifications difficult to confine to a single construct. Fine-tuning encodes personality in model parameters (Shao et al., 2023; Wang et al., 2024), but fixes personality levels after training, requiring retraining for each target change and limiting its applicability to large-scale multi-agent social simulations. Representation engineering encodes traits as directions in activation space that permit intervention (Turner et al., 2023; Chen et al., 2025a). Its potential to combine fine-tuning effects with flexible adjustment makes representation engineering the most promising of the three approaches.

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.

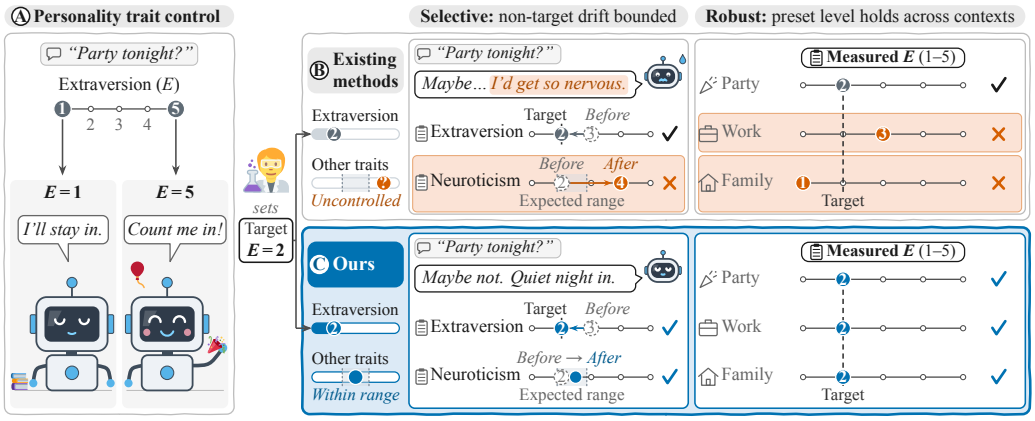

Figure 1: Social simulation requires controlling the personality of LLMs, but existing methods cannot control it both selectively and robustly, whereas TRAIT achieves both.

Current representation engineering methods still struggle to meet the selectivity and robustness requirements of experimental validity. Selectivity depends on whether extraction data yield vectors uncontaminated by other constructs. One line of research assumes that model self-reports can directly serve as high-quality extraction data (Turner et al., 2023; Rimsky et al., 2024). Yet research has shown that models' self-reports correlate only weakly with their actual behavior, so vectors extracted from self-report data may not transfer to behavior (Liu et al., 2026a). Another line uses frontier LLMs to generate or select extraction data (Chen et al., 2025a; Geng et al., 2026). Such data cannot reflect the natural behavioral distribution of the target model, creating a mismatch between the states from which vectors are extracted and those in which they are injected at inference. These limitations reveal ***the first key question (Q1): how can we automatically synthesize extraction data that respect psychological construct boundaries and match the distribution of the target model, yielding vectors that actually change its reasoning and behavior?***

Robustness depends on maintaining the preset expression level of the target psychological construct across contexts after vector injection. Early work injects vectors at a fixed strength for each output, causing the same preset level to produce different effects across contexts (Turner et al., 2023; Chen et al., 2025a). Recent methods dynamically adjust injection strength using the projection of the input onto the direction. However, the projection is not perfectly linearly related to the expression level of the target psychological construct, so the resulting strength cannot match the preset adjustment level (Zhao et al., 2025; Chen et al., 2025a). These limitations raise ***the second key question (Q2): how can an injection scheme determine the injection strength from context, keeping the expression level of the target psychological construct stable at the preset level across contexts?***

We propose TRAIT, an automated personality control method, to address these two questions. For the first, we propose Direction Extraction from Deliberation Spans (DeliSpan), which uses a psychometric scale of the target psychological construct as seed data, asks the target model to synthesize decision scenarios, reason within them, and make decisions, and extracts directions only from its deliberation spans. **Scale seeds provide psychological grounding**, self-generation matches the data to the distribution of the target model, and actual decisions ensure that the data capture genuine reasoning and action rather than self-reports inconsistent with behavior. For the second question, we propose Dynamic Injection with Context-Adaptive Dosing (CtxDose), in which **the dose network is supervised by deviations between post-injection scale scores and preset scores**, and adjusts injection strength using context and the preset level to ensure robustness.

To address these two questions, this paper makes three contributions:

1. **The TRAIT framework.** We present TRAIT, an automated framework that achieves selective and robust control over designated personality traits of LLMs in activation space. Given a target personality, TRAIT keeps the model's behavior faithfully aligned with the preset level across diverse contexts, while confining drift along non-target dimensions to the bounds that measurement noise and construct correlations can account for (§3).

2. **DeliSpan and CtxDose.** On the extraction side (Q1), DeliSpan takes scale items as seeds, invites the target model to synthesize its own decision scenarios and reason its way to a choice, and distills vectors solely from the reasoning segments in which the model reveals its own preferences while weighing trade-offs (§3.3). On the injection side (Q2), CtxDose trains a dose network supervised by the deviation between post-injection scale scores and preset scores, and predicts the injection strength based on context (§3.4).
3. **Findings.** Under the LLM psychometric evaluation framework of Serapio-García et al. (Serapio-García et al., 2025), TRAIT sets a new state of the art among representation engineering methods in both selectivity and robustness (Tables 1 and 2). Ablations confirm that restricting extraction to these segments markedly sharpens selectivity, and that context-adaptive dosing markedly narrows the target deviation relative to the static and projection dosing schemes (Figures 3 and 4).

2 Related Work

Personality measurement in LLMs. Personality is one of the most frequently manipulated independent variables in generative social simulation (Argyle et al., 2023; Park et al., 2023). Personality here does not imply that LLMs possess the same character or emotional experiences as humans. Synthetic personality refers to stable response patterns arising from the training data distribution of an LLM under specific types of prompts (Serapio-García et al., 2025). Serapio-García et al. evaluated 18 LLMs using IPIP-NEO and BFI (Serapio-García et al., 2025). They found that some models exhibit personality measurable with psychometric instruments.

Control via prompting and fine-tuning. Personality can also be adjusted. Prompting methods have progressed from personality descriptions and degree modifiers to quantitative control. For adjacent constructs such as cognitive bias, CoBRA aligns behavioral tendencies with quantitative targets through closed-loop behavioral measurement and iterative adjustment (Liu et al., 2026b). Yet prompting still lacks selectivity. Correlations and unclear boundaries among personality dimensions in natural language make interventions difficult to confine to the target psychological construct (Serapio-García et al., 2025). Fine-tuning encodes personality in model parameters through post-training on behavioral examples embodying the target personality (Shao et al., 2023; Wang et al., 2024). Building agent populations with heterogeneous personality distributions requires frequent retraining at high training and deployment costs.

Activation-space intervention. Representation engineering builds on linear representations of high-level concepts (Park et al., 2024). Behavioral traits form directions in activation space, and injection along these vectors can causally change trait expression, with continuously adjustable strength and no model weight updates (Turner et al., 2023; Zou et al., 2023). Persona Vectors has demonstrated the feasibility of monitoring and intervening on personality-related behavioral traits through these directions (Chen et al., 2025a). The paradigm comprises extraction and injection. Extraction determines what a direction encodes; impure directions perturb unrelated dimensions and undermine selectivity. Injection determines when and how strongly to intervene, affecting whether expression remains stable at the preset level across contexts.

Direction extraction and injection. Advances in extraction methods primarily reflect changes in their data. Early methods used human-authored or frontier-model-generated contrasting statements expressing opposite tendencies on the same construct and differenced activations at the starting position of the response (Turner et al., 2023; Wu et al., 2025). Subsequent work recognized the importance of psychological grounding, using psychometric scales or social science experiments as seeds for data synthesis (Liu et al., 2026b). Building on this work, we use psychometric scales as seeds for data synthesis by the target model and locate deliberation spans. Injection methods have gradually shifted from fixed strength (Turner et al., 2023; Rimsky et al., 2024; Chen et al., 2025a) to dynamic adjustment, as in AdaSteer, which regresses doses from the projection of the input onto the direction (Zhao et al., 2025). We instead train a dose network on deviations between post-injection scale scores and preset scores, an external psychometric anchor rather than the projection.

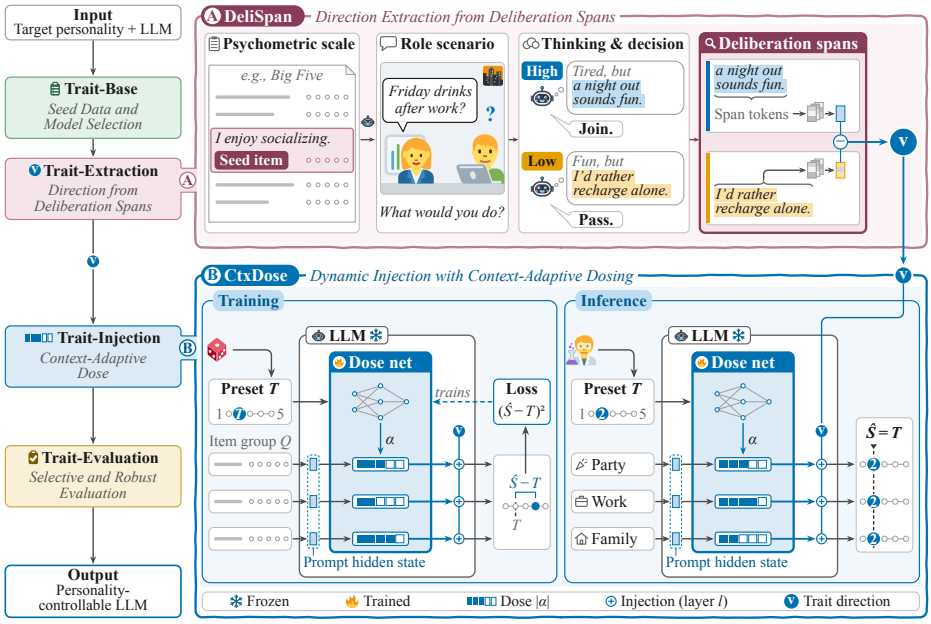

Figure 2: The TRAIT framework. (A) DeliSpan synthesizes scenarios from scale items and extracts the direction only from deliberation spans. (B) CtxDose sets the injection dose from the prompt hidden state and the preset score, trained against post-injection scale scores.

3 Method

3.1 Method Overview

We introduce TRAIT, a framework that controls specific personality traits of large language models selectively and robustly, without modifying the model's weights. Let M denote the target model. Given a situational input x and a prescribed target trait score $T \in [1, 5]$, our goal is to intervene on the model's activations as it generates y, so that its expressed behavior aligns with T while drift in non-target traits is tightly suppressed. As illustrated in Figure 2, TRAIT has four tightly integrated stages. Trait-Base (§3.2) anchors construct seeds in standardized psychometric inventories and establishes both the model to be controlled and a strictly isolated data foundation. Trait-Extraction (§3.3) uses DeliSpan, which elicits the model's intrinsic preferences within self-synthesized decision scenarios and distills construct-pure feature directions from them. Trait-Injection (§3.4) uses CtxDose, in which a lightweight dose network predicts injection strength dynamically from the scenario's hidden state and the target score and applies a bounded intervention. Trait-Evaluation (§3.5) uses an independent test item bank to assess the selectivity of control and its robustness across scenarios.

3.2 Trait-Base: Seed Data and Model Selection

To keep personality control construct-pure, we use the individual items of standardized psychometric inventories directly as the seed set $\mathcal{I}$. They carry well-defined construct boundaries and explicit forward- and reverse-keying rules, which prevents the cross-trait interference that free-form natural-language descriptions tend to introduce.

To select a base model M with a measurable personality, we adopt the psychometric evaluation protocol established by Serapio-García et al. (Serapio-García et al., 2025). A qualifying model must pass cross-inventory tests of reliability and discriminant validity without any intervention. Appendix C.2 reports the screening criteria and results.

3.3 TRAIT-EXTRACTION: DIRECTIONS FROM DELIBERATION SPANS

We propose DeliSpan, an automated pipeline for representation extraction.

Starting from the inventory items in the seed set, we prompt the target model M to synthesize first-person decision scenarios that involve behavioral conflict. Under prompts representing the high and low poles of the relevant trait, the model reasons through each scenario and makes a choice. Before admitting samples to the dataset, we apply a consistency filter that removes those lacking discriminative power.

We then use a prompt-aligned evaluator to isolate deliberation spans within the whole thinking trace. The thinking trace is the model's entire chain of thought, whereas deliberation spans are the segments within it where the model reveals its own preferences while weighing trade-offs, the places where the target psychological construct is expressed. The model's autoregressive reasoning usually mixes objective factual statements and descriptions of the environment, so extracting from the whole trace would contaminate the direction with irrelevant semantics. We therefore annotate only the set of token positions $\tilde{S}$ where the model reveals its own preferences while weighing trade-offs. Isolating these deliberation spans filters out background noise from the scenario, and our later evaluations show that this step substantially improves the selectivity of the extracted vectors.

At each candidate layer l, we aggregate hidden states over the deliberation spans and compute paired differences. For each valid scenario i, we mean-pool the hidden states within $\tilde{S}$ separately under the high- and low-pole conditions. We then average the differences across scenarios to obtain the direction vector v_l as follows:

$$z_l^{(i,c)} = \frac{1}{|\tilde{S}_{i,c}|} \sum_{t \in \tilde{S}_{i,c}} h_{l,t}^{(i,c)}, \qquad v_l = \frac{1}{N} \sum_{i=1}^{N} \left(z_l^{(i,\text{high})} - z_l^{(i,\text{low})} \right), \tag{1}$$

where c denotes the high- or low-polarity condition and N is the number of valid scenarios.

3.4 TRAIT-INJECTION: CONTEXT-ADAPTIVE DOSING

We propose CtxDose, a dynamic injection mechanism in which a lightweight dose network sets the intervention strength adaptively from contextual hidden states.

The dose network models the input-specific intervention strength as a global prior baseline plus a contextual compensation term. For each input text, we take the hidden state z_i averaged over all prompt tokens at the end of prefill, read at the injection layer with no intervention applied. We reduce this state with principal component analysis, concatenate it with the continuously sampled target score T, and add cross-interaction terms. The result is fed into a linear map Δ_θ, which predicts the compensation dose.

We adopt a linear rather than neural form for Δ_θ based on a first-order analysis. If item i's response curve is approximately the mean curve shifted by u_i, expanding at $\alpha_0(T)$ gives $\alpha_i - \alpha_0(T) \approx -u_i / \bar{s}'(\alpha_0(T))$, and because u_i is linearly predictable from the hidden state, the correction is linear in z_i. A neural network adds parameters with no counterpart in this structure, which must be estimated from the same hundred-odd training items and thus inflate variance.

The final intervention dose applied at the injection layer is

$$\alpha_i = \text{clip}\left(\alpha_0(T) + \Delta_\theta(z_i, T),\, \alpha_{\min},\, \alpha_{\max}\right), \tag{2}$$

where $\alpha_0(T)$ is a warm-start dose obtained by inverting the averaged response surface.

We train the dose network against a differentiable surrogate response surface fitted offline, with supervision coming directly from the external psychometric standard. Piecewise-linear interpolation then turns these discrete scores into a continuous response function $s_i(\alpha)$ that is differentiable almost everywhere. The loss is the mean squared error between each group's predicted inventory score and the prescribed target score T as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{G,T}\left[\left(\frac{1}{|G|} \sum_{i \in G} s_i(\alpha_i) - T \right)^2 \right], \tag{3}$$

where G denotes a randomly sampled subset of assessment items.

3.5 TRAIT-EVALUATION: SELECTIVE AND ROBUST EVALUATION

Selectivity. We measure whether the other traits stay within the range that measurement noise and inter-trait correlation can explain while the target trait is being steered.

We propose the selective operating amplitude metric A_{sele}. To capture statistical uncertainty under different intervention states, we build a 95% confidence noise band from the variances of two independent samples of the base model, one at the unintervened baseline and one under intervention with the candidate direction applied:

$$b_k(\alpha) = 1.96 \cdot \sqrt{\frac{\sigma_{\alpha,k}^2}{N} + \frac{\sigma_{0,k}^2}{N}}, \tag{4}$$

where k indexes any personality dimension other than the target trait, N is the number of items in each trait's test, and $\sigma_{\alpha,k}$ and $\sigma_{0,k}$ are the sample standard deviations under intervention and at baseline, respectively. For each non-target trait, we first subtract the co-movement expected from its population correlation with Extraversion and then test whether the remaining deviation falls within the noise band that we define as follows:

$$|\Delta_k(\alpha) - r_k\, G(\alpha)| \le b_k(\alpha), \tag{5}$$

where $\Delta_k(\alpha)$ is the change in trait k relative to baseline, $G(\alpha)$ is the gain in Extraversion relative to baseline, and r_k is the IPIP-questionnaire subgroup estimate from the meta-analysis of van der Linden et al. (van der Linden et al., 2010), which are 0.45 for Openness, 0.18 for Conscientiousness, 0.45 for Agreeableness, and -0.37 for Neuroticism. The criterion is two-sided. A non-target trait that drifts beyond what correlation predicts, or that fails to change along with Extraversion, counts as a breach, because an intervention on a personality construct should reproduce the covariance structure of traits in human populations. Starting from zero dose and extending continuously in both directions, we define the clean interval C as the range over which every non-target trait satisfies this condition, meaning that variable isolation holds. A_{sele} is the span of target-trait score gains reachable within C, expressed in scale-score points and invariant to any reparameterization of the dose axis.

To select, among candidate representations, the personality vectors that are both practically useful and construct-pure, we then define a composite score with a saturating gate and asymmetric conditional semantics in the following composite form:

$$S = A_{\text{sele}} \cdot \frac{A_{\text{eff}}}{A_{\text{eff}} + \kappa}, \tag{6}$$

where A_{eff} is the maximum control amplitude over which the vector moves the target trait within its monotonic effective range, and $\kappa > 0$ is a characteristic constant that sets the saturation scale of the gate. The gate vetoes vectors that cannot steer the trait and saturates toward one once A_{eff} is adequate, ensuring that S is driven mainly by A_{sele}. We set κ to the largest effective amplitude observed in the candidate set, $\max A_{\text{eff}}$.

Robustness. We measure whether a prescribed modulation level takes effect consistently across behavioral scenarios. Under our protocol, the controller first sets its dose for each target value T using only the calibration dataset. With all parameters then frozen, it performs injection on an unseen, independent test set. Taking the actual official inventory score $s(T)$ on the independent test set as ground truth, we use the root mean squared error to measure its deviation from each prescribed target value T belonging to the target set $\mathcal{T}$ as follows:

$$E_{\text{rob}} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{T \in \mathcal{T}} (s(T) - T)^2}. \tag{7}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Target trait and model. The target trait is Extraversion, one of the Big Five and among the personality constructs most widely invoked in simulations of social interaction. The steered model is

Qwen3.8-27B, which reliability and validity screening confirmed to possess a measurable synthetic personality; the screening criteria and results are reported in Appendix C.2.

Data and evaluation protocol. Primary evaluation employs the complete IPIP-300 inventory, whose 300 items span all five Big Five dimensions, with Extraversion serving as the target trait and the remaining four as non-target traits. The items used for dose learning and calibration are likewise drawn from the IPIP psychometric item pool and remain strictly disjoint from the final evaluation set; their composition and scale are detailed in Appendix C.1.

Implementation details. We sweep intervention strength across a non-uniform grid of 15 doses spanning -32 to $+64$, collecting $n = 60$ independent scores per construct at each dose. For robustness, we sample 4 target scores uniformly within the linear regime of the intervention curve, repeat each 4 times, and measure E_{rob} on held-out Extraversion items that are kept apart from all training and calibration throughout. All experiments run on a single NVIDIA RTX PRO 6000 with 96 GB of memory, with the model served through vLLM 0.29 and EasySteer (Xu et al., 2025) at generation temperature 0, seed 42, and thinking mode reasoning_effort=medium.

4.2 Selectivity and Construct Purity

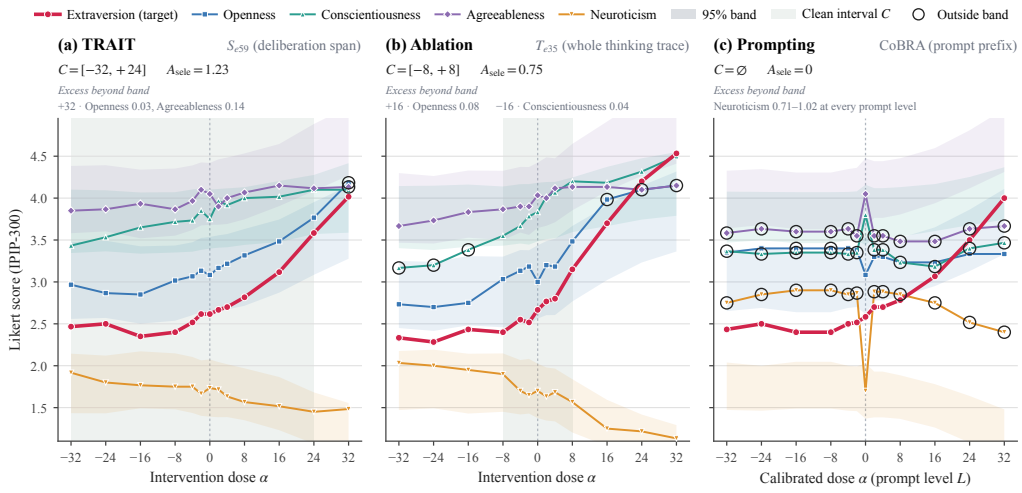

Figure 3: Selectivity on the IPIP-300 test set. Shaded bands span the range explained by construct correlation and noise, and the green region marks the clean interval C. (a) DeliSpan keeps every non-target trait inside the band over the widest interval. (b) Whole-thinking-trace extraction breaches at ± 16. (c) CoBRA prompting breaches at every level.

We sweep the intervention dose on an independent IPIP-300 test set covering the four non-target dimensions. Table 1 summarizes the results for representative vectors.

The principal vector S_{e59}, extracted from self-synthesized decision-thinking segments, achieves the best construct isolation. After construct correlation is subtracted, its clean interval spans -32 to $+24$, which corresponds to a selective operating amplitude of 1.233 points. It also preserves a monotonic effective control span of 2.183 points, and **its composite score of 0.563 is the highest of all candidates**. This result shows that isolating the segments of the chain of thought in which the model states its preferences and weighs trade-offs keeps behavioral control while minimizing aberrant spillover into non-target traits.

To probe the mechanistic contribution of segment isolation, we run an ablation across extraction positions. The whole-thinking-trace vector T_{e35}, extracted without segment selection, has a much narrower clean interval of -8 to $+8$ and a selective amplitude of only 0.750 points. This indicates that the factual statements and environmental reasoning mixed into an unisolated autoregressive chain of thought readily introduce semantic contamination into multiple non-target dimensions. Appendix C.4 reports the first breach of each candidate vector and the screening of Persona Vectors.

Table 1: Selectivity of representative candidate vectors on the independent IPIP-300 test set, sorted in ascending order of the composite score S. C is the clean interval after construct correlation is subtracted, A_{eff} is the effective control amplitude, and A_{sele} is the selective operating amplitude, all in scale points. The two deliberation-span vectors rank first.

Vector	Extraction	C	A_{eff}	A_{sele}	S
S_{e15}	Deliberation span	$[-32, +8]$	2.600	0.550	0.275
T_{e35}	Whole thinking trace	$[-8, +8]$	2.250	0.750	0.348
T_{e31}	Whole thinking trace	$[-16, +16]$	1.650	1.083	0.421
S_{e43}	Deliberation span	$[-16, +16]$	2.100	1.133	0.506
$\boldsymbol{S_{e59}}$	**Deliberation span**	$\boldsymbol{[-32, +24]}$	**2.183**	**1.233**	**0.563**

4.3 Robustness and Dynamic Dosing

On an independent test item bank, we set four continuous target scores and measure E_{rob}, the root-mean-square error between the realized official scale scores and the prescribed targets. Figure 4 compares how well each control strategy transfers across contexts.

The static dose shows marked alignment distortion on unseen item sets, with an E_{rob} of 0.225 points. AdaSteer does not solve the problem either. Its E_{rob} is 0.193 points over four passes. Closer analysis shows that the norm of the model's hidden states at the injection layer is nearly constant, with a coefficient of variation of only about 1%. The projection inner product therefore carries no meaningful variance that reflects context-specific responsiveness, and the method in effect degenerates into static control with no context adaptivity.

CtxDose, by contrast, generalizes adaptively across contexts. Across the four continuous targets, **it reduces E_{rob} to 0.142 points**. Structural ablation further reveals that nonlinearity confers no benefit. Replacing the linear compensation term with a single-hidden-layer MLP raises E_{rob} from 0.142 to 0.220, close to the static dose, which supports the linear mapping in Section 3.4. Appendix C.5 reports the per-target errors and analyzes the context information that the dose network exploits.

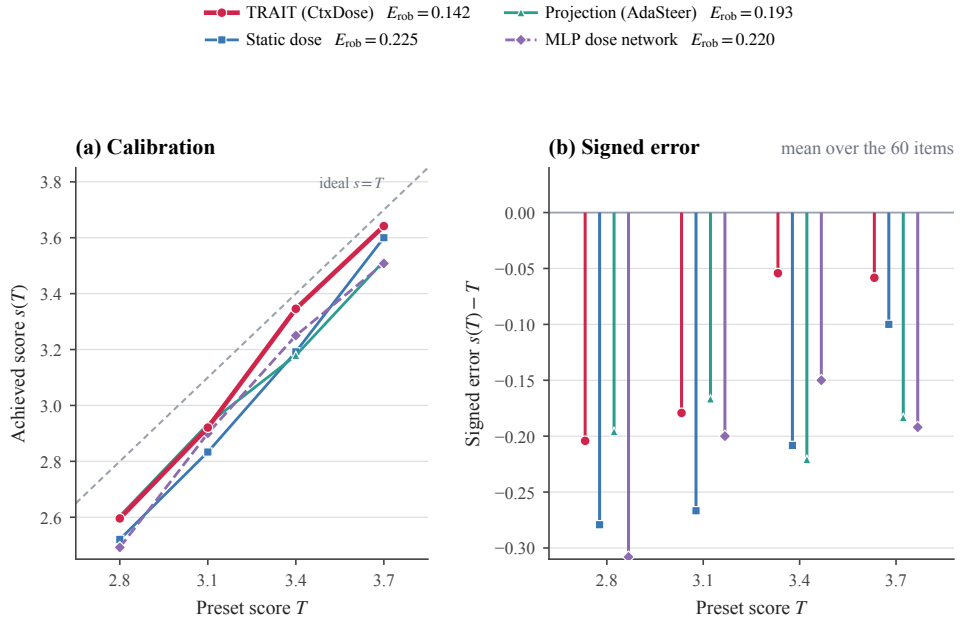

Figure 4: Robustness on the held-out official 60 Extraversion items. (a) Achieved versus preset score. (b) Signed error at each target. CtxDose stays closest to the preset level.

4.4 Overall Comparison and Experimental Validity

Table 2 reports effective control amplitude, selective operating amplitude, mean non-target trait drift, and E_{rob} for each method on the IPIP-300 benchmark.

The CoBRA baseline moves Extraversion to some degree and is highly robust, but **its selectivity is very poor**. On the official 60 items its E_{rob} is 0.081 points, far below the 0.142 achieved by TRAIT. Pure prompting methods therefore very readily introduce confounding variables when controlling the personality constructs of agents, undermining experimental validity. Appendix C.6 details the drift of CoBRA at each prompt level.

Persona Vectors extracts its direction from the final response and was eliminated at the vector screening stage because of its weak control effect.

These results show that TRAIT combines a wide modulation range, rigorous multivariable isolation, and stable targeting across scenarios. It gives social science experiments an independent-variable manipulation technique that achieves both selectivity and robustness, laying a foundation for the validity of causal inference.

Table 2: Overall comparison on the IPIP-300 benchmark. A_{eff} and A_{sele} are in scale points, mean drift is the mean absolute non-target drift within the clean interval, and E_{rob} is measured on the held-out official 60 Extraversion items over four passes. CoBRA is robust, but its selective amplitude of zero leaves it without practical value.

Method	A_{eff}	A_{sele}	Mean drift	E_{rob}
Persona Vectors[b]	*no effective vector extracted*			
AdaSteer[a]	–	–	–	0.193
CoBRA[c]	2.117	0	–	0.081
TRAIT	**2.183**	**1.233**	**0.151**	**0.142**

[a]Projection-based dosing borrowed from AdaSteer and applied to the TRAIT vector S_{e59}, differing only in how the dose is set. A_{eff} and A_{sele} are defined on a fixed-dose axis and are not applicable to input-adaptive dosing, and sharing TRAIT's vector they would duplicate its amplitudes. [b]The Persona Vectors extraction scheme (response-token mean difference) on our self-synthesized data fails the dose screening, so no effective vector is available and no metric is defined. [c]The clean interval is empty, so drift within it is undefined. Its A_{eff} is the span over the 10 prompt levels.

5 Limitations

Personality-control robustness still requires improvement. It still lags well behind prompting methods, and it is harder to maintain over multi-turn interactions.

Validation covers only Extraversion in Qwen3.8-27B. Other traits and models may differ in representations and intervention responses, requiring renewed extraction and calibration.

6 Conclusion

We present TRAIT for selective and robust personality control in social simulation. DeliSpan extracts personality directions only from the deliberation spans of the target model's self-synthesized reasoning, and CtxDose sets the injection dose from context under supervision by post-injection scale scores. On Extraversion, TRAIT sets a new state of the art among representation engineering methods on all three metrics, E_{rob}, A_{sele}, and S.

TRAIT gives simulated social science experiments an automated and scalable way to manipulate personality as an independent variable selectively and robustly, laying a foundation for the validity of these experiments and the conclusions drawn from them.

AI USE STATEMENT

We used generative AI tools, chiefly programming agents built on large language models such as Claude Code and Codex, at the following stages of this work. They helped implement code for experiments, analysis, and plotting. They launched and monitored GPU experiments, cleaned and assembled datasets, and supported statistical analysis and the interpretation of results. They translated Chinese drafts written by the authors into English and produced preliminary versions of figures, which the authors subsequently vectorized and revised. They also assisted with literature searches, citation verification, and editing for readability. The method and experimental design were developed by the authors themselves through discussion with these tools. A large language model is also a component of the method itself, as described in Section 3. The model under study, Qwen3.8-27B, synthesizes decision scenarios, generates reasoning traces, and annotates deliberation spans, while a separate large language model filters the synthesized scenarios according to fixed quality criteria. The evaluation uses no LLM judge; every item is scored by the option to which the model assigns the highest probability. This paper contains no proofs. The authors reviewed all work completed with AI assistance and checked every reported number against the original outputs. We take full responsibility for the accuracy, originality, and compliance of all content, including any text, code, data, and claims produced with the aid of generative AI.

ETHICS STATEMENT

This study involves no human participants. IPIP-300 and BFI are publicly available inventories, and all other experimental data were synthesized by the model itself. These data contain no personal information and pose no ethical risks concerning privacy or human subjects. The personality control technique presented here is meant to give social science simulation a calibratable way to manipulate variables. We also recognize that deliberately shaping a model's personality carries a potential risk of misuse for deception or manipulation. To reduce this risk, we anchor control in publicly available standardized psychometric instruments, so that the target and magnitude of each intervention remain quantifiable and auditable. We will also include usage guidelines with the released code and data. We emphasize that a synthetic personality is a stable pattern of responses that the model shows under a particular prompt distribution and is not equivalent to human personality. Conclusions about social simulation drawn with this technique should therefore not be extrapolated directly to real populations. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

All components needed to reproduce the results of this paper are provided in the main text and appendices. The extraction pipeline, the criteria for scenario synthesis and deliberation span annotation, and the filtering gates are described in Section 3.3 and Appendix B.2. The dose network and its training protocol are presented in Section 3.4 and Appendices B.3 and B.4. The evaluation protocol is set out in Section 3.5 and Appendix C.1. The IPIP-300 and BFI inventories and the base model are all publicly available, and the source code will be released to the public.

REFERENCES

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371, 2023.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2. URL https://doi.org/10.1017/pan.2023.2.

Donald T. Campbell and Julian C. Stanley. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (ed.), *Handbook of Research on Teaching*, pp. 171–246. Rand McNally, Chicago, 1963.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, 2025a. URL https://arxiv.org/abs/2507.21509.

Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. Seal: Steerable reasoning calibration of large language models for free, 2025b.

Yilin Geng, Omri Abend, Eduard Hovy, and Lea Frermann. Implicit personality representations in humans and llms, 2026. URL https://arxiv.org/abs/2609.12704.

Maik Larooij and Petter Törnberg. Validation is the central challenge for generative social simulation: a critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*, 59(1):15, 2026. doi: 10.1007/s10462-025-11412-6.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Haoze Liu, Run Liu, Haiying Xu, Jiahui Han, Siyuan Fang, Siyu Yan, Huiqi Deng, Guanchu Wang, and Na Zou. Your llm, your style: Behavioral mode axes for llm behavioral control, 2026a. URL https://arxiv.org/abs/2608.10703.

Xuan Liu, Haoyang Shang, and Haojian Jin. CoBRA: Programming cognitive bias in social agents using classic social science experiments. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 2026b. doi: 10.1145/3772318.3790804.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt 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)*, 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://doi.org/10.1038/s42256-025-01115-6.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston, 2002.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Dimitri van der Linden, Jan te Nijenhuis, and Arnold B. Bakker. The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, 44(3):315–327, 2010. doi: 10.1016/j.jrp.2010.03.003.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Stephen W. Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, 2024. doi: 10.18653/v1/2024.findings-acl.878.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, 2025.

Haolei Xu, Xinyu Mei, Yuchen Yan, Rui Zhou, Wenqi Zhang, Weiming Lu, Yueting Zhuang, and Yongliang Shen. EasySteer: A unified framework for high-performance and extensible LLM steering, 2025. Accepted to EMNLP 2026 System Demonstrations.

Weixiang Zhao, Jiahe Guo, Yulin Hu, Yang Deng, An Zhang, Xingyu Sui, Xinyang Han, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. AdaSteer: Your aligned LLM is inherently an adaptive jailbreak defender. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24559–24577, 2025. URL https://aclanthology.org/2025.emnlp-main.1248/.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A Additional Related Work

Later methods shifted to extraction from model responses. CAA differences activations at the option-letter position in answers to human-authored multiple-choice questions (Rimsky et al., 2024); Persona Vectors uses differences in mean activations across complete open-ended responses (Chen et al., 2025a); SEAL computes differences in group means from activations at the boundaries of self-generated reasoning (Chen et al., 2025b).

Generative agent-based models use LLMs as their core, allowing macro-level social patterns to emerge from micro-level behavior through agent actions and interactions. A recent systematic review finds that generative social simulations have extremely low validity and can scarcely produce reliable social science experimental results (Larooij & Törnberg, 2026). One important reason is the difficulty of controlling independent variables. Specifically, manipulating an independent variable must satisfy two conditions. First, steering one trait must leave the other traits unchanged beyond what measurement noise and their correlations with the target trait can explain; otherwise, the observed effect cannot be attributed to the independent variable itself. We call this selectivity, corresponding to confounding in experimental validity theory (Campbell & Stanley, 1963; Shadish et al., 2002). Second, preset adjustments must remain stable across behavioral contexts, with the nominal and actual levels of the manipulation of the independent variable aligned. We call this robustness (Shadish et al., 2002).

Campbell and Stanley noted that failing to separate the effects of extraneous variables from those of the experimental stimulus threatens internal validity and makes results difficult to interpret (Campbell & Stanley, 1963). Shadish, Cook, and Campbell termed the inclusion of non-target constructs in a manipulation construct confounding, listing it as a formal threat to construct validity (Shadish et al., 2002). Shadish et al. listed partial implementation of standardized treatments as a formal threat to statistical conclusion validity because it underestimates treatment effects and undermines comparability across implementations (Shadish et al., 2002). Serapio-García et al. used 104 trait adjectives to control target traits at 9 levels; 11 of 12 tested models achieved a mean Spearman correlation of at least 0.80 between preset levels and scale scores (Serapio-García et al., 2025). CoBRA exemplifies this approach by converting classic social science experiments into behavioral measurement tasks to achieve quantitative control (Liu et al., 2026b). From these segments, we extract vectors that are uncontaminated by other constructs, match the distribution of the model, and can actually change its reasoning and behavior. Indiscriminate injection can also elicit target behavior for irrelevant inputs, impairing usability (Lee et al., 2025).

B Method Details

B.1 Control Objectives, Evaluation Framework, and Model Selection

Our method is designed for any target personality dimension and controls one dimension at a time. All operations and numerical values reported in this paper instantiate this general method for Extraversion. A change in the target dimension requires only replacing the corresponding scales and item sources and repeating the same procedure.

The target model is denoted by M, and its residual stream activation at layer l and position t is $h_{l,t} \in \mathbb{R}^d$. The model produces an answer y for a scenario x. The answer is evaluated with an item group adapted from the psychometric scale, yielding the observed score $\hat{T}$. The control objective is to bring $\hat{T}$ close to the preset score T. The current implementation uses Likert response options from 1 to 5, and T is the mean score across the scale group. The method consists of an extraction layer and an injection layer. The extraction layer constructs a personality vector v from the generation traces of the target model and freezes it. The injection layer decides for each answer during generation how much to inject. A dose network performs this function.

Our evaluation follows the psychometric framework of Serapio-García et al. (Serapio-García et al., 2025). This framework adapts public personality assessment items from the International Personality Item Pool IPIP. It evaluates whether models exhibit personality through reliability, convergent validity, discriminant validity, and criterion validity.

Our initial tests assess the candidate models Qwen3-8B, Qwen3.8-27B, and gpt-oss-120b to ensure that the target model exhibits synthetic personality. Qwen3.8-27B is the only model that passes both the convergent validity and discriminant validity gates for the tested dimension of Extraversion. It is selected as the target model.

These items have been calibrated through long psychometric practice. This ensures that its latent space already contains a stable, intrinsic synthetic personality representation with room for modulation, rather than producing random responses to the assessment.

We take Extraversion as the control target, since it is the most central trait in social simulation, and build the seed pool from the corresponding items of IPIP-300 and the BFI.

The data comprise three parts with no item overlap, a dose learning item pool, held-out behavioral scenarios, and a final test. The dose learning item pool contains 139 authentic scale items for the target dimension, finalized after exclusions. The entire pool is used without splitting. The 72 held-out behavioral scenarios are registered separately and excluded from direction extraction and dose learning. The final test comprises IPIP-300 and BFI with their official configurations. These are excluded from training and model selection.

B.2 Extraction Layer

The extraction layer selects behavioral and affective items associated with the target dimension from the public IPIP item pool. It excludes items that overlap with the final evaluation or have equivalent meanings to form the seed data. The selection for Extraversion yields 228 items from 3,320 items, with 137 items retained after exclusions.

Our scenario generation procedure consists of two steps. The first step generates a decision outline that specifies the central tradeoff for the dimension, such as spending time alone or with others. The second step generates concrete scenarios using 12 predefined material libraries, such as choosing between spending time alone and meeting friends after work on a Friday. The 137 seed data items for Extraversion yield 772 scenarios that enter the collection after quality checks. The quality checks are performed by an LLM against fixed quality rules, and this screening is aligned with human expert judgments. Each scenario is paired with the two extreme levels of the personality-control preamble from Serapio-García et al. as personality conditions (Serapio-García et al., 2025). Level 1 represents an extremely low level, and level 9 represents an extremely high level. The model is instructed to reason and make a decision within the scenario.

Repeated sampling from the target model for Extraversion yields 4,632 reasoning traces. The target model serves as the annotator and is calibrated by humans. It annotates sentence-level segments in the reasoning traces that express the personality preferences of the model during deliberation. The annotation criteria distinguish facts within the scenario, judgments about other people or things, and the choices of the model. Only the third category is marked as candidate segments. Empty segments are allowed, and both extremes use the same criteria. The annotation covers 4,227 traces, which are divided into high and low groups according to the actual choices of the model within the scenarios. The extraction retains 1,026 traces, including 402 under the L1 condition and 624 under the L9 condition.

The high and low conditions for the same scenario form paired samples. The personality vector is the mean of the paired differences.

$$z_l^{(i,c)} = \frac{1}{|\tilde{S}_{i,c}|} \sum_{t \in \tilde{S}_{i,c}} h_{l,t}^{(i,c)}, \qquad v_l = \frac{1}{N} \sum_{i=1}^{N} \left(z_l^{(i,\text{high})} - z_l^{(i,\text{low})} \right),$$

$\tilde{S}_{i,c}$ contains all positions covered by the annotated segments for scenario i under condition c. The layer index and normalization method for the vector are selected from candidate sets during development and frozen. The same set of traces also yields three control arms using the whole thinking trace, the full answer, and the boundary before reasoning. These control arms support experimental comparisons of extraction positions.

The extraction scheme is determined through three screening stages with progressively increasing cost and accuracy. The first stage is a two-gate battery with zero generation cost. The first check assesses consistency between methods. The cosine similarity between the DiffMean direction and the linear probe direction must be at least 0.6 for at least three consecutive layers within a layer band. The next check assesses transfer across scenarios. Extraction and testing use two disjoint halves of the scenarios, with 1000 label permutations required to yield $p \leq 0.01$. This gate has sole veto authority. The second stage is a small-scale layer sweep. The finalized candidate vector and layer pairs number no more than 5. The validation item group tests whether injection produces an actual steering effect. The third stage is a full intensity grid test. It selects the vector and injection layer with the widest adjustment range and the effect closest to linearity. The readout scan covers 16 full-attention layers and selects a layer band instead of a single layer. The injection layer is determined through empirical testing of each layer within the band.

It synthesizes decision data that respects psychological construct boundaries and follows the model's own intrinsic distribution, and it distills from the model's chain of thought personality vectors that are both selective and effective at steering behavior. Because the scenarios are self-synthesized from inventory seeds, they have clear construct boundaries while staying close to the model's intrinsic behavioral distribution. This avoids the distributional shift that data generated by an external model would introduce. This confirms that the representation has a linearly separable geometric structure and rules out statistical artifacts caused by outlying samples. This rules out binding to the narrative content of particular scenarios. Together, the two gates ensure that each surviving vector encodes a genuine, robust, and context-independent personality representation.

Candidate personality vectors must pass two quality gates. The first is a cross-method consistency test, which requires the mean-difference direction and the linear-probe direction to have high cosine similarity across consecutive layers. The second is a cross-scenario generalization test. We partition the scenarios into two disjoint subsets and require that a vector extracted from one subset significantly separates the high and low poles of the other, with significance established by a random permutation test.

B.3 Injection Layer

Injection uses bounded addition $h' = h + \alpha v$. The signed scalar α is mapped through a bounded parameterization to an interval determined by development data.

The dose network f_θ takes the scenario state z_i and preset score T as input and outputs $\alpha_i = f_\theta(z_i, T)$. z_i is the hidden state at the injection layer averaged over all input positions at the end of prefill without injection. Each answer receives one dose decision at the end of prefill. This decision remains in effect throughout the answer without being recomputed for each token. The injection rule is $h'_{l,j} = h_{l,j} + M_j \alpha_i v_l$, with M_j denoting the token mask for injection positions. Each training sample is an item group and target score pair $(\mathcal{Q}, T)$. Each item within the group is answered independently. For each training item i, the hard score takes the highest-probability option among the K valid options, and reverse-scored items use $r_i(k) = K + 1 - k$. We record this score on a fixed grid of doses and interpolate piecewise-linearly to obtain the response curve $s_i(\alpha)$. The group score is the mean over the group.

$$\bar{s}_\mathcal{Q} = \frac{1}{|\mathcal{Q}|} \sum_{i \in \mathcal{Q}} s_i(\alpha_i), \qquad \mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(\mathcal{Q},T) \in \mathcal{B}} (\bar{s}_\mathcal{Q} - T)^2 + \lambda \|W\|^2,$$

where W denotes the weights of Δ_θ. Development and evaluation use the same hard scores.

This lets the model converge reliably on the prescribed personality target score across contexts, so that control remains robust across diverse scenarios. At inference time, the system computes and applies the dynamic dose as soon as prefill of the prompt is complete. It reads the prompt-averaged hidden state z_i at the injection layer, with no intervention applied, and passes it through the dose network to obtain the context-specific dose α_i. This dose is added to the residual stream of the target layer in bounded form and held fixed for the rest of autoregressive generation. The intervention therefore stays stable throughout generation and corrects the personality drift that arises from differences in contextual representation.

The dose network is trained in a simulated environment. First, we apply a discrete grid of interventions to the training items and record the resulting assessment scores. The network is optimized end to end in this simulated environment, one item group at a time.

B.4 Training and Evaluation Configuration

The primary model selection metric is the mean absolute error between the hard group mean scores of the validation groups and the preset scores. Scale reliability and validity metrics follow the original criteria of Serapio-García et al. Reliability must be at least 0.70. The convergent correlation must be at least 0.80. The difference between convergence and discrimination must exceed 0.40. The criterion direction must be consistent. The control correlation must be at least 0.80. Selectivity is evaluated via the selective operating amplitude A_{sele} and the composite score S, assessing construct purity under a two-sample 95% confidence noise band. Robustness is evaluated via the root mean squared error E_{rob} under the calibration-to-transfer protocol across held-out behavioral scenarios and item pools. A static dose, linear conditional regression, and an MLP controller are compared using the same training data, target set, and optimization budget.

B.5 Rationale of the Evaluation Metrics

We propose the following standardized evaluation protocol and metrics for measuring the selectivity and robustness of personality-construct control. In human populations, the Big Five traits naturally show moderate construct correlations. Extraversion correlates moderately and positively with Openness and Agreeableness, and negatively with Neuroticism. Requiring non-target traits to stay perfectly constant would therefore contradict established findings in personality psychology. In addition, any measurement based on a finite number of items carries a sampling standard error, and intervention can shift the model's generation probability distribution. Principled selective control thus requires that fluctuations in non-target traits stay strictly within the bounds allowed by construct correlation and sampling error. Because different intervention strengths can change how dispersed the model's outputs are, the two-sample variance lets the decision boundary adapt to the current dose, so the evaluation scale stays consistent across intervention states. The score treats effectiveness as a nonlinear admission gate and makes the selective amplitude A_{sele} its main component. When A_{eff} is severely inadequate, the gating factor decisively vetoes weak vectors that cannot steer the trait. Once effectiveness passes the critical threshold, the gating factor converges quickly toward its saturation limit of one, and the score comes to depend almost entirely on selectivity, A_{sele}. This uses the modulation ceiling observed in the model's latent space as an unbiased normalization reference, so the gate is half-saturated at that ceiling. By construction, this caps the gain a vector can earn from amplitude alone and keeps selectivity the dominant factor in the composite score. This tests whether injecting the personality-construct vector still achieves the intended effect in contexts it has never seen. Occasional runaway deviations in extreme scenarios could invalidate the experimental variable in a social simulation, so we penalize them heavily. This cross-context transfer protocol tests whether the control method can hold an agent reliably to its prescribed personality baseline in unfamiliar environments.

C Experimental Details and Additional Results

C.1 Subject Model, Evaluation Protocol, and Baseline Implementation

We carry out a systematic empirical evaluation of how well the TRAIT framework controls personality. The experiments focus on three questions. First, we test whether personality representations extracted from chain-of-thought segments can shift the target trait substantially while strictly suppressing aberrant drift in non-target traits. This establishes the selectivity of control. Second, we test whether CtxDose can overcome contextual variation when transferred across item sets and hold realized scale scores steadily at their prescribed levels. This establishes the robustness of control. Finally, we compare TRAIT against mainstream baselines to assess its experimental validity as an instrument for precisely manipulating independent variables in social simulation experiments.

The final systematic evaluation uses the complete IPIP-300 inventory, whose 300 items cover the five Big Five dimensions, including the target trait of Extraversion and the four non-target traits. The BFI

serves only in base-model admission, where it provides the second inventory for the cross-inventory tests.

In its default configuration, the model reasons through a chain of thought, weighing trade-offs and preferences carefully before it commits to a discrete decision or writes a final response. This makes it an ideal substrate for studying how representation-level interventions shape both the model's deeper cognition and its overt behavior. The protocol treats the LLM as an independent test-taker, assessed in its natural state without any inducing persona. It then applies genuine psychometric tests of reliability and validity to decide whether a stable, measurable synthetic personality has emerged endogenously in the model's latent space.

1. **CoBRA.** CoBRA (Liu et al., 2026b) controls the strength of a cognitive bias through numerical prompts that specify a percentage level. We instead specify the level of Extraversion, a personality construct, and use the two poles of the trait as anchors. The prompt prefix is prepended to each inventory item within the same user message.
2. **Persona Vectors.** The key of Persona Vectors (Chen et al., 2025a) is to average activations over the model's response tokens and to take the difference between the mean activations of the two poles as the trait direction. We implement this method on the same synthesized scenarios and filtered data as TRAIT, using the tokens of the final response after $<$/think$>$. Its injection strength is fixed and does not depend on the input.
3. **AdaSteer.** AdaSteer (Zhao et al., 2025) sets the steering coefficient from the projection of the input hidden state onto the steering vector; it targets the binary choice between refusal and compliance and maps the projection to a coefficient by regression. We borrow only its idea of setting the injection strength from the projection. We read the projection of the last prompt token onto this vector before injection and calibrate the relationship between the post-injection projection length and the preset value. At the injection stage, we measure the projection length in advance and set the injection strength according to the preset projection length.

C.2 Psychometric Assessment of the Target Model

Following the protocol of Serapio-García et al. (Serapio-García et al., 2025), we screened three candidate models psychometrically with no intervention applied.

We conduct initial assessments of Qwen3-8B, Qwen3.8-27B, and gpt-oss-120b. Full direct-response assessments and thinking-mode assessments are administered separately. Because our method extracts directions from natural thinking, model selection relies primarily on an Extraversion-focused assessment in thinking mode. This assessment comprises 124 items and 250 administration combinations, producing 31,000 item-level records per model.

Table 3 reports Extraversion reliability, convergent validity, and discriminant validity. The Cronbach α exceeds 0.70 for both IPIP and BFI Extraversion in all three models. For Qwen3.8-27B, the two coefficients are 0.961 and 0.907, the convergent correlation is 0.917, and the discriminant-validity difference is 0.638, satisfying the current selection criteria. It passes all four checks of Extraversion reliability, convergent validity, discriminant validity, and criterion validity.

Intervening on an LLM's personality presupposes that a stable, measurable synthetic personality representation has already emerged endogenously in its latent space. Qwen3.8-27B achieves internal-consistency reliability coefficients of 0.961 on IPIP Extraversion and 0.907 on BFI Extraversion. Its cross-inventory convergent correlation is 0.917, well above the admission threshold of 0.80. The gap between its convergent validity and its discriminant validity against non-target traits is 0.638, far above the criterion of 0.40. The model therefore passes the full suite of admission tests. By contrast, Qwen3-8B was excluded because its discriminant-validity gap was only 0.368. The third candidate, gpt-oss-120b, also fell short, with a cross-inventory convergent correlation of only 0.750.

Qwen3-8B achieves a convergent correlation of 0.838 but a discriminant-validity difference of 0.368. This indicates that correlation between the two Extraversion measures alone is insufficient to satisfy the current model-selection criteria. gpt-oss-120b achieves a discriminant-validity difference of 0.509 and a convergent correlation of 0.750. Based on the complete thinking-mode Extraversion assessment, we select Qwen3.8-27B for subsequent experiments.

Table 3: Extraversion psychometrics in thinking mode.

Model	IPIP E α	BFI E α	Convergent r	Discriminant Δ	Outcome across four checks
Qwen3-8B	0.897	0.715	0.838	0.368	Fails discriminant validity
Qwen3.8-27B	0.961	0.907	0.917	0.638	Passes all checks
gpt-oss-120b	0.837	0.781	0.750	0.509	Fails convergent validity

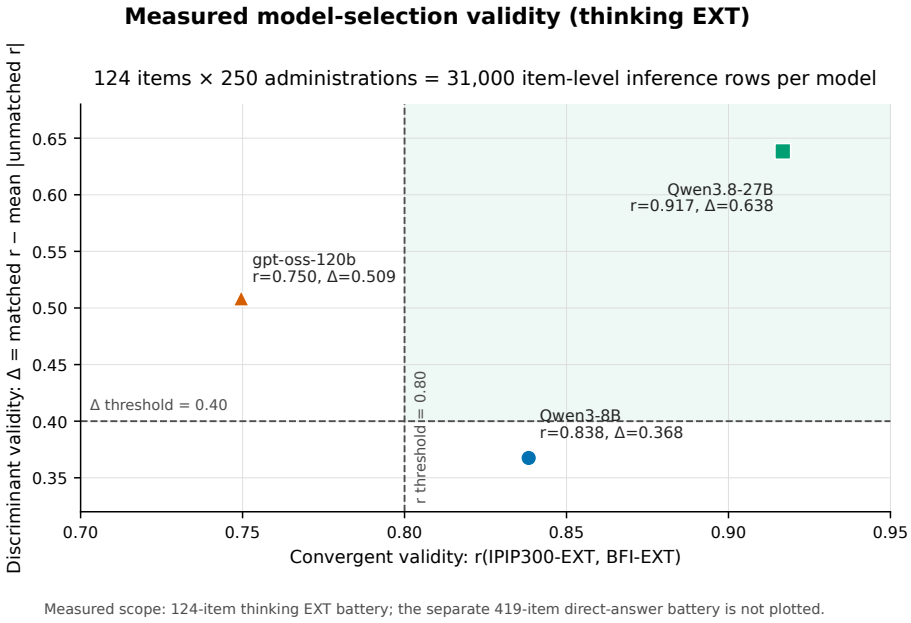

Figure 5: Basis for model selection. The horizontal axis shows the correlation between IPIP and BFI Extraversion scores. The vertical axis shows the difference between this matched-trait correlation and the mean absolute correlation with unmatched traits. Dashed lines mark the selection thresholds of 0.80 and 0.40, respectively. The upper-right region satisfies both criteria. Outcomes for all four checks appear in Table 3 and the original reports.

C.3 Extraction Data and Cross-Scenario Representation Diagnostics

Using 137 standard Extraversion items as seeds, we prompt the model to synthesize 772 first-person decision scenarios that involve behavioral conflict. After strict personality-preference annotation and behavioral-consistency filtering, we retain 1,026 high-quality decision traces for representation extraction.

The 772 scenarios produce 4,632 trajectories, of which 4,227 have been annotated and 1,026 retained for extraction. Among the retained trajectories, 402 originate from low-Extraversion conditions and 624 from high-Extraversion conditions. These source conditions describe the sampling composition; actual grouping follows annotations of the behavior and preferences expressed in each trajectory. Table 4 summarizes the extraction pipeline.

On the admitted Qwen3.8-27B, we apply a scale-seeded self-synthesis extraction pipeline. Under bipolar personality prompts, this yields 4,632 self-generated chain-of-thought reasoning traces.

We extract mean-difference vectors of deliberation spans at the model's 16 full-attention layers and pass them through two quality gates:

- **Cross-method consistency gate.** A candidate layer must reach a cosine similarity of at least 0.6 between the mean-difference direction and a linear probe across three consecutive layers. This confirms that the representation has a robust, linearly separable geometric structure.
- **Cross-scenario generalization gate.** We split the scenario data into two disjoint subsets. Each candidate vector must significantly separate personality polarity on the held-out subset, reaching $p \leq 0.01$ under a 1,000-permutation test.

Together, the two gates filter out overfitted vectors tied to specific narrative content as well as artifacts caused by outlier samples. The result is 47 candidate vectors that are geometrically consistent and generalize across scenarios, giving a reliable representational foundation for the dose analysis and precise intervention that follow.

We compare the four extraction positions S, T, R, and B0′ on the same trajectories. The left panel of Figure 6 shows the AUROC of mean-difference directions transferred between the two scenario halves; the right panel shows agreement between mean-difference and linear-probe directions. S, T, and R all pass cross-scenario screening, retaining 15, 16, and 16 extraction layers, respectively. B0′ does not pass the full screening procedure.

Complete responses R show strong separability and agreement between methods in the existing representation diagnostics. S and T also preserve the distinction between high and low Extraversion in unseen scenarios. These results provide candidates for subsequent intervention at all three extraction positions. Although some later layers of B0′ achieve high mean-difference AUROC, B0′ still fails the complete screen, illustrating that a single metric from one method captures a narrower aspect of the evidence than the screening procedure as a whole.

Representation diagnostics establish whether hidden states contain transferable differences in Extraversion. Effective control further requires examining target and off-target changes after injecting the same vector. Subsequent experiments retain S, T, and R and compare their attainable ranges and drift to assess how span selection influences behavioral control.

C.4 Vector Selection: Effectiveness, Selectivity, and a Composite Criterion

Figure 3 panels. (a) The deliberation-span vector S_{e59} holds every non-target trait inside the band over the clean interval $C = [-32, +24]$. (b) The whole-thinking-trace vector T_{e35} breaches at $+16$ and at -16. (c) Under CoBRA prompting, Neuroticism exceeds the band at every prompt level, so the clean interval is empty.

Cross-scenario screening retains more candidate vectors than downstream experiments can carry forward, so the behavioral sweep doubles as a selection stage. This subsection defines the selection criteria and applies them to five candidate Extraversion vectors. Three come from deliberation spans (S_{e15}, S_{e43}, S_{e59}) and two from the whole thinking trace (T_{e31}, T_{e35}), where the suffix denotes the extraction layer. Each vector is injected at fifteen doses $\alpha \in \{-32, -24, -16, -8, -4, -2, 0, 2, 4, 8, 16, 24, 32, 48, 64\}$, with $n = 60$ items per dose and construct, and each dose is scored independently; interventions do not compound across doses.

All band comparisons below use the same two-sample 95% noise band. For a score mean $\bar{s}_\alpha$ at dose α with standard deviation σ_α and the no-injection baseline $\bar{s}_0$ with σ_0, the band is

$$b(\alpha) = 1.96\sqrt{\sigma_\alpha^2/60 + \sigma_0^2/60}.$$

A deviation within the band is treated as measurement noise; a deviation beyond it is treated as a real change.

Effectiveness. The valid region of a direction is the largest dose interval containing zero in which the Extraversion response stays functional. Moving outward from zero along each arm, a dose is invalid if the Extraversion gain falls significantly below the running extreme of that arm, that is, if the response reverses direction beyond the band. Let $G(\alpha)$ denote the Extraversion gain at dose α relative to baseline. The effective amplitude is the span of Extraversion scores reachable within the valid region,

$$A_{\text{eff}} = \max_{\alpha \in V} G(\alpha) - \min_{\alpha \in V} G(\alpha),$$

where V is the valid region.

Table 4: Extraction data counts.

Stage	Count	Description
Seed items after exclusions	137	Used to generate decision scenarios
Accepted scenarios	772	Scenarios retained through the construction pipeline
Generated trajectories	4,632	Samples under high- and low-personality conditions
Annotated trajectories	4,227	Trajectories with completed span annotations
Trajectories retained for extraction	1,026	Trajectories used for activation extraction
Halves A and B	513 each	Mutually disjoint scenarios
Vectors passing subsequent screening	47	15 layers for S; 16 each for T and R

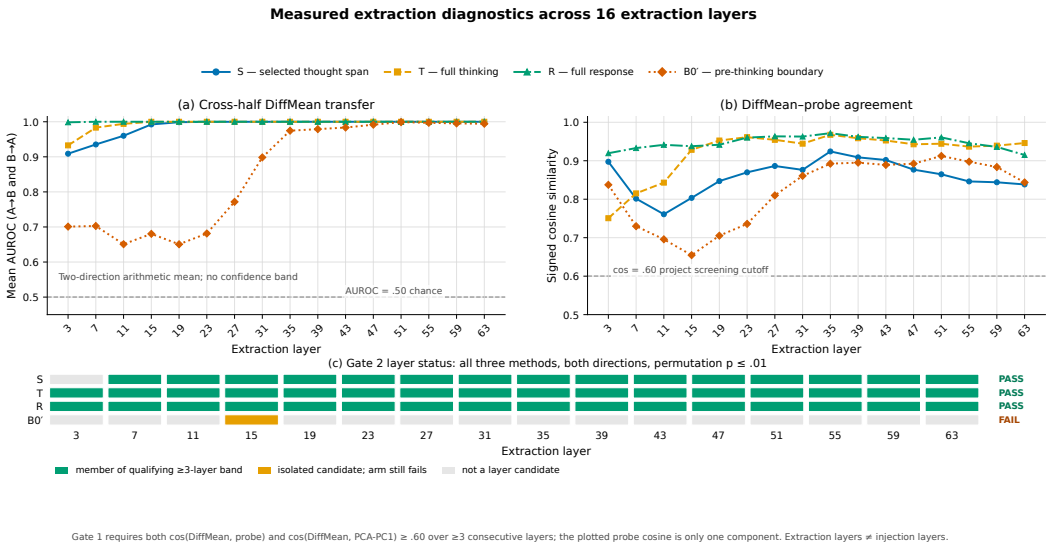

Figure 6: Representation diagnostics across extraction positions and layers. The dashed line at 0.5 in the left panel denotes chance ranking. The dashed line at 0.6 in the right panel marks our threshold for agreement between methods. The four curves correspond to deliberation spans, the whole thinking trace, complete responses, and the pre-thinking boundary. Passing layer intervals are marked according to the full cross-scenario tests. Layer indices refer to extraction layers; actual injection layers are selected in the behavioral experiments.

Selectivity. The clean interval of a direction is the largest dose interval containing zero such that, for every dose in the interval and every off-target trait $k \in \{O, C, A, N\}$, $|(\bar{s}_{\alpha k} - \bar{s}_{0k}) - r_k\, G(\alpha)| \leq b_k(\alpha)$, where r_k is the population correlation between Extraversion and trait k. The first dose beyond the interval at which any off-target trait exceeds the band is the first break. The selective operating amplitude is the span of Extraversion scores reachable within the clean interval,

$$A_{\text{sele}} = \max_{\alpha \in C} G(\alpha) - \min_{\alpha \in C} G(\alpha),$$

where C is the clean interval. Both amplitudes are expressed in scale-score points and are invariant to any reparameterization of the dose axis. This is the quantity defined in Section 3.5.

Permitted drift from construct correlations. Not every change in a non-target trait is undesirable. Measurement noise and drift induced by the correlations among personality constructs are both admissible, and only changes beyond these two sources indicate that the intervention has leaked into unrelated constructs. The clean-interval criterion therefore centers the noise band on the correlation-predicted drift $r_k\, G(\alpha)$. The correlations are the IPIP-questionnaire subgroup estimates of van der Linden et al. (van der Linden et al., 2010) ($N = 5{,}619$), with $r = 0.45$ for Openness, 0.18 for Conscientiousness, 0.45 for Agreeableness, and -0.37 for Neuroticism. The criterion is two-sided, so a trait that fails to co-move with Extraversion also counts as a breach; an intervention on a personality construct should reproduce the covariance structure observed in human populations.

The choice of correlation estimates does not affect the selectivity optimum. Replacing the IPIP subgroup estimates with the pooled meta-analytic correlations of the same study ($N = 144{,}117$; uncorrected r of 0.31, 0.21, 0.18, and -0.26 for Openness, Conscientiousness, Agreeableness, and Neuroticism) or with their corrected values (ρ of 0.43, 0.29, 0.26, and -0.36) leaves S_{e59} at $A_{\text{sele}} = 1.233$ with clean interval $[-32, +24]$ under all three sets. Under the corrected set, S_{e43} rises to 1.183 and the other vectors are unchanged; under the uncorrected pooled set, S_{e43}, T_{e31}, T_{e35}, and S_{e15} reach 0.600, 0.633, 0.750, and 0.550.

One limitation applies. The meta-analytic correlations describe covariation between persons, whereas the intervention moves a single model within itself; the criterion assumes that the within-model covariance under intervention should mirror the between-person structure.

Composite criterion. The two amplitudes are combined by a saturating gate with asymmetric conditional semantics, as defined in Section 3.5,

$$S = A_{\text{sele}} \cdot \frac{A_{\text{eff}}}{A_{\text{eff}} + \kappa},$$

where the characteristic constant κ is set to the largest effective amplitude in the candidate set, $\max A_{\text{eff}} = 2.60$. Effectiveness acts as an admission gate. When A_{eff} is severely inadequate the gating factor vetoes vectors that cannot steer the trait, and once effectiveness passes the critical threshold the gating factor converges toward one, so the score comes to depend almost entirely on selectivity. Setting κ to the observed modulation ceiling makes the gate half-saturated at that ceiling and caps the gain a vector can earn from amplitude alone.

Table 5 reports the measurements, and Figure 7 illustrates the two amplitudes and their geometric meaning on S_{e59}. The ranking under S is $S_{e59} > S_{e43} > T_{e31} > T_{e35} > S_{e15}$. The top two positions are both deliberation-span vectors, and S_{e59} ranks first, attaining the largest selective amplitude and the widest clean interval. Beyond the summary measurements, the two whole-thinking-trace vectors degrade at high doses. The Extraversion gain of T_{e31} falls from 1.25 at $+32$ to 0.78 at $+48$ while three off-target traits break simultaneously, and T_{e35} leaves its valid region at $+64$. No deliberation-span vector shows this reversal within the swept grid.

Deliberation-span versus whole-thinking-trace extraction. The top two composite scores both come from deliberation spans, and the best deliberation-span vector exceeds the best whole-thinking-trace vector on both metrics. At a dose of $+16$, Openness rises by 0.98 points, more than twice the 0.47 points explained by construct correlation; at -16, Conscientiousness falls by 0.45 points, whereas construct correlation predicts only 0.04 points.

First breaches. Quantifying where each vector first breaches the noise band reveals a clear pattern. After construct correlation is subtracted, every candidate vector exceeds the noise band by only 0.02 to 0.25 points at its first breach. Under positive doses, breaches concentrate on Openness and Agreeableness, the two traits most strongly correlated with Extraversion. Openness rises beyond the amount predicted by correlation, whereas Agreeableness fails to rise along with Extraversion. Under negative doses, breaches concentrate on Conscientiousness. The principal vector first breaches only at a dose of $+32$, where Openness exceeds the band by 0.03 points and Agreeableness by 0.14 points. This finding contradicts the hypothesis that intervention triggers systemic representational collapse. Trait shifts beyond the boundary are instead departures from the covariance structure of psychological constructs under strong intervention. DeliSpan effectively contains these departures and keeps a wide clean interval of $[-32, +24]$.

Persona Vectors screening. The Persona Vectors baseline, extracted from the model's final responses, is clearly weaker in control than both families of thinking-based vectors in the vector screening. Its maximum control span is 1.8 points, below 2.0 points for deliberation spans and 2.1 points for whole thinking traces, and its mean downward shift at the same injection layer is only 0.27 points, about half that of the deliberation-span and whole-thinking-trace vectors, so it was excluded at the screening stage. Its extraction position does not isolate the segments of the chain of thought that express preferences and trade-offs, so the resulting direction can hardly change the model's reasoning and behavior effectively, and it therefore did not enter the full evaluation of selectivity and robustness. Its injection strength is fixed, placing it in the same class as the static dose in this paper; even with the vector extracted by our method, fixed-strength injection reaches an E_{rob} of 0.225.

Table 5: Vector selection measurements for the five candidate Extraversion vectors. Measurements are from the behavioral dose sweep ($n = 60$ per dose and construct). Downward valid regions extend to the -32 grid boundary for all five vectors.

Vector	Source	Valid region	A_{eff}	Clean interval	First break	A_{sele}	S
S_{e15}	deliberation span	$[-32, +64]$	2.60	$[-32, +8]$	$+16$ (C)	0.55	0.275
S_{e59}	deliberation span	$[-32, +64]$	2.18	$[-32, +24]$	$+32$ (O, A)	**1.23**	**0.563**
S_{e43}	deliberation span	$[-32, +64]$	2.10	$[-16, +16]$	$+24$ (O), -24 (C)	1.13	0.506
T_{e35}	whole thinking trace	$[-32, +48]$	2.25	$[-8, +8]$	$+16$ (O), -16 (C)	0.75	0.348
T_{e31}	whole thinking trace	$[-32, +64]$	1.65	$[-16, +16]$	$+24$ (A), -24 (C)	1.08	0.421

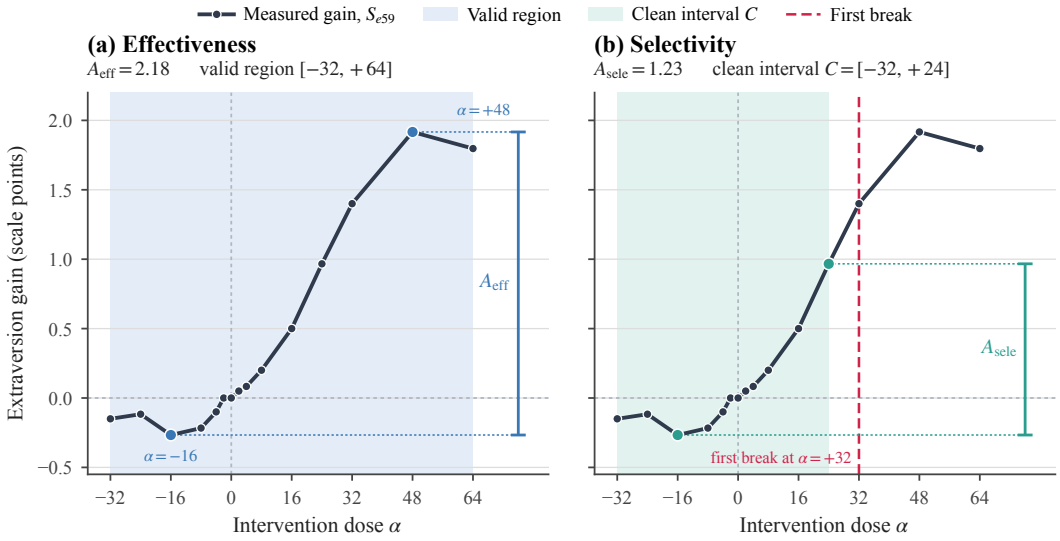

Figure 7: Geometric meaning of the selection amplitudes, illustrated on S_{e59} ($n = 60$ per dose). In panel (a), the valid region, shaded in blue, is where the Extraversion response does not reverse direction beyond the noise band; A_{eff} is the span of reachable gains within it. In panel (b), the clean interval, shaded in green, is where no off-target trait deviates from its correlation-predicted drift by more than the noise band; A_{sele} is the span of reachable gains within it, and the red line marks the first break at $+32$, where Openness and Agreeableness first exceed the band.

Two qualifications bound the strength of these rankings. First, under a 5000-fold bootstrap that resamples each cell as $N(\bar{s}, \sigma^2/60)$, individual pairwise gaps fall within sampling noise at $n = 60$ per cell, so the ranking is reported as point estimates rather than as pairwise significance claims. Second, the maximum gain of S_{e15} lies at the boundary dose $+64$, so its A_{eff}, which also sets κ, is right-censored. The position of S_{e59} as the selectivity optimum is unaffected by either qualification, because S_{e59} has the largest A_{sele} and it ranks first for any value of κ.

C.5 Robustness under Transfer: Per-Target Errors and Mechanism

Textual contexts differ inherently in how difficult and how sensitive they are to express, so no single static dose can hold across all of them. Empirical analysis shows that the hidden state at the final position of the prompt prefill, taken before injection, contains an independent component orthogonal to the personality direction. This component predicts context-induced response shifts with a coefficient of determination between 0.2 and 0.35. The dose network explicitly extracts this orthogonal feature and builds an interaction-based correction from it, which compensates for context-specific differences in sensitivity.

The static dose and AdaSteer-style projection fall below every target.

Table 6: Robustness under transfer to the held-out official 60 Extraversion items. E_{rob} is the root-mean-square of the four signed errors, and MAE is their mean absolute value. Every method is averaged over four passes. The first three rows inject the same vector S_{e59} and differ only in how the dose is set.

				Signed error $s(T) - T$			
Method	Passes	E_{rob}	MAE	$T = 2.8$	$T = 3.1$	$T = 3.4$	$T = 3.7$
Representation engineering							
TRAIT (CtxDose)	**4**	**0.142**	**0.124**	**−0.204**	**−0.179**	**−0.054**	**−0.058**
Static dose	4	0.225	0.214	−0.279	−0.267	−0.208	−0.100
AdaSteer-style projection	4	0.193	0.192	−0.196	−0.167	−0.221	−0.183
Prompting							
CoBRA	4	0.081	0.071	+0.021	−0.054	−0.083	−0.125

C.6 Overall Comparison: CoBRA Details

Finally, we assess how sound each representative method is as an instrument for manipulating independent variables in social simulation experiments. At all 10 prompt levels, Neuroticism rises by 0.70 to 1.20 points and still exceeds the noise band after construct correlation is subtracted; even when the Extraversion score has barely moved, Neuroticism has already risen by about 1 point.