CALIBRATING ACTIVATION STEERING WITH SPARSE BEHAVIORAL FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering can control a frozen language model through an additive direction, yet a coefficient that works for one checkpoint can produce little change in another. We study how to turn this model dependence into a calibration problem. Across 15 instruction-tuned models and five personality attributes, residual-stream scale predicts sensitivity to a shared coefficient grid: a mid-layer RMS feature achieves a held-out-family rank correlation of 0.83. Within each model, this scale is stable across attributes, suggesting a reusable measurement for identifying candidate dose ranges. Behavioral feedback then selects a directional operating point within a range. In a retrospective evaluation of measured grids on five models, six rated responses from one calibration scenario select a dose that transfers to the remaining scenarios. At the same feedback budget, expanding the candidate grid raises the mean held-out signed contrast from 0.24 on the original grid to 1.27, with improvements in all 25 model–attribute averages. Six ratings retain 89.6% of the effect obtained using 30 ratings on the same scenario. These results separate model-level sensitivity from attribute-level dose selection and establish sparse behavioral calibration as a practical way to recover useful steering from an existing direction.

1 INTRODUCTION

Inference-time activation steering offers a direct way to change how a frozen language model responds. A vector extracted from contrasting examples is added to an internal state, providing control without updating model weights (Zou et al., 2023; Li et al., 2024; Rimsky et al., 2024). For personality and style, this makes it possible to adapt an existing checkpoint through a compact intervention (Feng et al., 2026; Chen et al., 2025). The practical question is how to choose an intervention that produces the intended behavioral change on the receiving model.

A steering direction and its coefficient play different roles. The direction specifies a contrast in representation space. The coefficient determines how strongly that contrast enters a particular residual stream. Applying the same coefficient to every checkpoint conflates direction quality with model sensitivity. Conversely, increasing the coefficient without measuring directionality can move the model beyond a useful operating region. An effective calibration procedure must therefore answer two questions: where to look for a useful dose, and which dose to use within that range.

We connect these questions through a study of *scale and sparse feedback* (Fig. 1). Activation scale provides a coarse measurement of a model's response regime. Behavioral feedback identifies a useful signed contrast for a particular attribute. The first measurement is reusable across attributes; the second resolves the shape of the dose-response curve. This division is useful when a practitioner has already extracted steering directions and wants to obtain effective control with few additional judged generations.

The evidence has two complementary levels. In a 15-model survey, the same raw grid produces markedly different effect ranges. A scale measurement taken at the middle of each network predicts the ordering of these effects when an entire model family is excluded from fitting. At that fixed measurement site, RMS varies little across the five attributes of a checkpoint, with a median within-model coefficient of variation of 1.03%. Scale thus provides a model-level descriptor even when the requested attribute changes.

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.

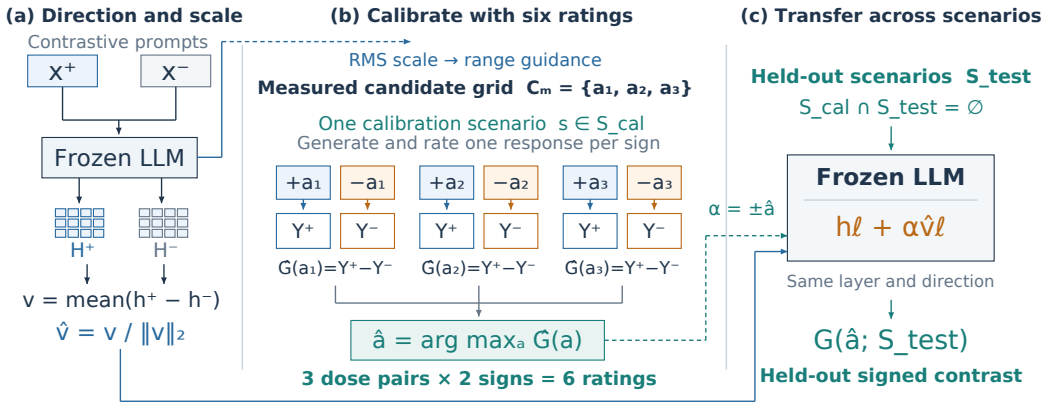

Figure 1: **From a steering direction to a transferable operating point.** (a) Contrastive activations provide a unit direction and a residual-scale measurement. Scale guides the choice of a model-specific range; the present evaluation uses established, measured candidate grids. (b) On one calibration scenario, each sign of three candidate amplitudes produces one rated response. Their signed contrasts select $\hat{a}$. (c) The same direction and layer, with coefficients $\pm\hat{a}$, are evaluated on disjoint scenarios. Solid arrows carry activations or responses; dashed arrows indicate range guidance and dose choice. The six-rating budget covers selection within the given grid.

The intervention study fixes each model's direction and injection layer and compares the original and expanded dose grids on five checkpoints. Expansion increases the available behavioral range, but the best directional operating point depends on the response curve. We evaluate a simple selection rule: rate one response at each sign of three candidate amplitudes, then choose the pair with the largest positive contrast. We repeat this evaluation for every available calibration scenario and repetition, measuring performance on disjoint scenarios. The resulting six-rating selector achieves a mean signed contrast of 1.27 on the expanded grids, compared with 0.24 on the original grid. Every model–attribute average improves under this matched-budget comparison.

This result identifies a useful research objective beyond maximizing the largest observed change: *obtain a transferable, directed operating point with a small feedback budget*. A response range can be large even when the strongest dose ceases to follow the requested direction. The Qwen3 dose curves demonstrate why selecting an operating point matters, while the other calibrated models show how enlarging the candidate range itself can supply much of the gain.

Our contributions are:

1. **A cross-model account of steering sensitivity.** Across 15 checkpoints and five personality attributes, a standardized mid-layer RMS feature predicts held-out-family effect ordering with Spearman $\rho = 0.83$, while remaining stable across attributes within each checkpoint.
2. **A matched-budget evaluation of transferable dose selection.** On five checkpoints, selecting among measured dose pairs with six ratings on one scenario yields positive held-out signed effects and improvements over the original-grid selector in all 25 model–attribute averages.
3. **An empirical separation of range discovery and operating-point selection.** Full dose curves, fixed-amplitude comparisons, and feedback-budget analyses identify when expanding the range is sufficient and when behavioral selection supplies additional value.

2 Related Work

Activation directions for behavioral control. Representation Engineering, Inference-Time Intervention, and Contrastive Activation Addition use internal representations to guide frozen language models (Zou et al., 2023; Li et al., 2024; Rimsky et al., 2024). Personality steering extends this approach to sustained behavioral attributes and persona control (Feng et al., 2026; Chen et al., 2025).

We use mean-difference directions to study the calibration problem that arises after direction extraction: how to obtain a useful signed effect on a receiving checkpoint.

Intervention scale and location. Activation scaling studies learned intervention magnitudes (Stoehr et al., 2024). Norm-based steering uses activation norms to express intervention strengths across models and layers (Korznikov et al.), and LayerNavigator uses activation measurements to select intervention locations (Sun et al., 2025). These approaches motivate measuring the receiving state. We show that an inexpensive scale feature predicts cross-family sensitivity, and we quantify the behavioral feedback needed to select among candidate doses. Its central comparison is between dose ranges under the same selection budget, with layer and direction fixed within each checkpoint. Scale-based parameterization and feedback-based selection answer complementary questions. A norm measurement expresses how large a perturbation is relative to a receiving state; the signed behavioral objective determines which measured amplitude produces the requested ordering of outputs. This distinction connects activation-side diagnostics to an operational choice that can be evaluated on new scenarios.

Measuring expressed personality. Questionnaire-based and behavioral evaluations study personality expression in language models (Serapio-García et al., 2023; Jiang et al., 2024; Zhu et al., 2025). We use scenario responses scored with explicit Big Five rubrics. This setting supplies five related control tasks on every checkpoint and makes cross-attribute reuse of activation scale measurable. The selection objective is a signed difference in expressed behavior on held-out scenarios.

3 From Scale Measurement to a Directed Operating Point

3.1 Directions, Injection, and Residual Scale

For a model m, attribute t, and layer ℓ, let (x_i^+, x_i^-) be a contrastive prompt pair. Last-token hidden states define the mean-difference direction

$$v_{m,t,\ell} = \frac{1}{K}\sum_{i=1}^{K}\big[h_\ell(x_i^+) - h_\ell(x_i^-)\big], \qquad \hat{v}_{m,t,\ell} = v_{m,t,\ell}/\|v_{m,t,\ell}\|_2. \tag{1}$$

During generation, we add $\alpha\hat{v}_{m,t,\ell}$ to the output of the selected transformer block. The hook broadcasts the vector over token positions; generation recomputes the prefix with caching disabled. The stored activation collections use either 20 paired rows or 160 rows constructed from ten scenarios and four high/low formulations per scenario. Appendix A specifies these sampling regimes and the actual injection sites.

We measure coordinate RMS over the contrastive activation collection $H = \{h_i\}_{i=1}^n$, including both signs:

$$R(H) = \sqrt{\frac{1}{nd}\sum_{i=1}^{n}\|h_i\|_2^2}, \qquad N(H) = \sqrt{d}\,R(H). \tag{2}$$

Here $N(H)$ is the root mean squared sample norm. The normalized quantity $a/N(H)$ describes a candidate amplitude relative to this activation collection. A measurement at $\ell = \lfloor L/2 \rfloor$ gives a common depth rule across models. We use Openness activations at that site for the main cross-model predictor, and all five traits to assess within-model scale stability.

Why scale is informative. Consider the scalar normalization part of RMSNorm, $F(\alpha) = (h + \alpha\hat{v})/R(h + \alpha\hat{v})$, with $R(h) = \sqrt{\|h\|_2^2/d + \epsilon}$. Its derivative is

$$F'(0) = \frac{\hat{v}}{R(h)} - \frac{\langle h, \hat{v}\rangle h}{dR(h)^3}. \tag{3}$$

The first term scales the injected component by the inverse residual RMS; the second accounts for its alignment with the receiving state. This local relationship motivates scale as a sensitivity feature. Downstream layers and finite-dose response determine the eventual behavioral effect, which is measured directly in the calibration stage. Appendix B.1 derives Eq. 3.

3.2 Separating Response Range from Directional Control

Let $Y_{t,s,r}(\alpha) \in \{1, \ldots, 5\}$ be the trait rating for scenario s, repetition r, and coefficient α. The mean rating over scenarios and repetitions is $\bar{Y}_t(\alpha)$. For a coefficient grid $\mathcal{A}$, the response range is

$$\Delta_t(\mathcal{A}) = \max_{\alpha \in \mathcal{A}} \bar{Y}_t(\alpha) - \min_{\alpha \in \mathcal{A}} \bar{Y}_t(\alpha). \tag{4}$$

This measures the available spread of expressed behavior. To choose a directional operating point, we instead evaluate symmetric signed pairs. For positive amplitude a, define

$$G_t(a; S) = \frac{1}{|S|} \sum_{s \in S} \left[\frac{1}{R_s} \sum_r Y_{t,s,r}(+a) - \frac{1}{R_s} \sum_r Y_{t,s,r}(-a) \right]. \tag{5}$$

A positive value means the two signs move behavior in the intended order. Its range is $[-4, 4]$. The symmetric contrast is directly relevant when the application requires both high and low expression of an attribute. In contrast, Δ_t can be large because of two coefficients with an unintended order. We therefore use Δ_t for the sensitivity survey and G_t for dose selection and transfer.

The contrast is computed within each scenario before averaging. This pairs the high- and low-expression interventions under the same prompt, so differences in a scenario's default tendency enter both sides of the comparison. For example, a scenario that generally elicits cooperative responses need not have the same baseline score as a conflict scenario for their signed steering effects to be compared. The test then asks whether the selected amplitude preserves useful separation across those contexts. Because the objective compares both poles, a high score requires separation in the requested order rather than simply a high rating for one generated response.

3.3 Selecting a Dose with Sparse Feedback

For a candidate set $\mathcal{C}_m = \{a_1, a_2, a_3\}$ and calibration scenarios S_{cal}, the selected amplitude is

$$\hat{a}_t = \arg\max_{a \in \mathcal{C}_m} \widehat{G}_t(a; S_{\text{cal}}), \qquad U_t = G_t(\hat{a}_t; S_{\text{test}}), \quad S_{\text{cal}} \cap S_{\text{test}} = \varnothing. \tag{6}$$

Ties select the smaller amplitude. One repetition at each sign of three candidates uses six response ratings. Two or five repetitions use 12 or 30 ratings. The same model-level candidate set is reused for every trait; the selected operating point is attribute-specific.

We evaluate this rule retrospectively on measured dose grids. For one-scenario calibration, every scenario serves as the calibration scenario in turn, and all remaining scenarios form the test set. For one or two repetitions, we enumerate all subsets of the five stored repetition indices, using the same index subset at each candidate and sign. Test scores average all five repetitions on the held-out scenarios. This enumerates every available calibration choice. We first average over these choices within each model–trait pair, then equally over traits and models.

The comparison holds the feedback budget fixed while changing the candidate range. Both grids contain three positive amplitudes and their negatives. Additional controls choose the largest expanded amplitude without feedback or choose uniformly among expanded candidates. These controls distinguish the value of expanding the range from the value of selecting within it.

3.4 Calibration Procedure and Feedback Allocation

The procedure separates reusable model configuration from attribute-specific selection. The checkpoint and injection layer define the receiving interface. Contrastive examples supply a direction for each attribute, while the measured candidate amplitudes specify where to query that interface. Once these quantities are fixed, calibration evaluates a small discrete choice.

Sparse calibration within a measured candidate range

Given: frozen model m, layer ℓ, unit direction $\hat{v}_{t,\ell}$, ordered positive candidates $\mathcal{C}_m$, calibration scenarios S_{cal}, and r repetitions.

1. For every $s \in S_{\text{cal}}$ and $a \in \mathcal{C}_m$, generate r responses at each coefficient $+a$ and $-a$, using the same layer and direction.
2. Rate the responses for attribute t and compute the mean signed contrast $\widehat{G}_t(a; S_{\text{cal}})$ for every candidate.
3. Return the smallest amplitude maximizing this contrast: $\hat{a}_t = \min \arg\max_a \widehat{G}_t(a; S_{\text{cal}})$.
4. Fix the resulting pair $(+\hat{a}_t, -\hat{a}_t)$ and evaluate $G_t(\hat{a}_t; S_{\text{test}})$ on scenarios excluded from selection.

Selection cost: $2|\mathcal{C}_m|\,|S_{\text{cal}}|\,r$ response ratings.

For three candidates, one scenario, and one repetition, this cost is six generations and six corresponding ratings. Comparing candidate means requires a linear scan over the candidate set. Neither direction extraction nor model parameters change during selection. Selecting a symmetric pair requires no zero coefficient, because both requested poles are already observed at every candidate amplitude; the zero coefficient serves only as a reference for describing the full dose-response curve. The selected pair can then be reused across subsequent scenarios without additional selection feedback.

Two sources of variation motivate distinct uses of an additional budget. Repetitions refine the contrast estimate on the same scenario, whereas additional scenarios change the contexts informing the choice. We evaluate both using the existing response records. The main six-, twelve-, and thirty-rating comparison varies repetitions at fixed scenario count, making the cost of estimating the operating point directly visible. The multiple-scenario analysis evaluates the complementary use of feedback for broader contextual coverage.

When a choice transfers. The selector benefits when calibration preserves the ordering of useful candidates. To state this precisely, let $\varepsilon = \max_{a \in \mathcal{C}_m} |\widehat{G}_t(a; S_{\text{cal}}) - G_t(a; S_{\text{test}})|$ for a fixed split. The selected pair satisfies

$$G_t(\hat{a}_t; S_{\text{test}}) \geq \max_{a \in \mathcal{C}_m} G_t(a; S_{\text{test}}) - 2\varepsilon. \tag{7}$$

This deterministic relation identifies the relevant quantity: agreement between calibration and test contrasts across the available candidates. The empirical transfer analysis measures the resulting held-out effect directly. Appendix C gives the short derivation and the exhaustive split construction.

4 Experimental Setting

Cross-model survey. The survey covers 15 instruction-tuned checkpoints from the Qwen, LLaMA, Mistral, Gemma, and Phi families, including DeepSeek-R1-Distill-Qwen in the Qwen group. Every checkpoint is evaluated on Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The original coefficient grid is $\{0, \pm 2, \pm 4, \pm 6\}$. Seven scenarios for Openness and eight for each other trait, with five repetitions per coefficient, yield 20,475 scored responses. Sampling uses temperature 0.8, top-p 0.9, and at most 200 new tokens. The benchmark contains both trait-specific probe-selected sites and fixed interior sites; Appendix Table 3 lists every site. The cross-model predictor estimates response under these evaluated policies.

Dose intervention. Five checkpoints have an additional measured grid at the same injection layer as their original-grid run. Positive amplitudes are $\{10, 20, 30\}$ for Qwen2.5-1.5B, 7B, and 14B; $\{18, 36, 54\}$ for Qwen3-0.6B; and $\{46, 92, 138\}$ for Gemma-2-2B. Their negatives and zero complete each grid. These model-specific ranges are fixed across traits and contribute 6,825 additional rated responses. These ranges are empirically configured. Scale motivates their magnitude, and the sparse-feedback evaluation conditions on the available candidates. The six-rating budget covers operating-point selection within a fixed candidate grid; constructing that grid is separate prior work. Model, attribute, injection layer, scenarios, repetition count, and decoding settings match between the two grids.

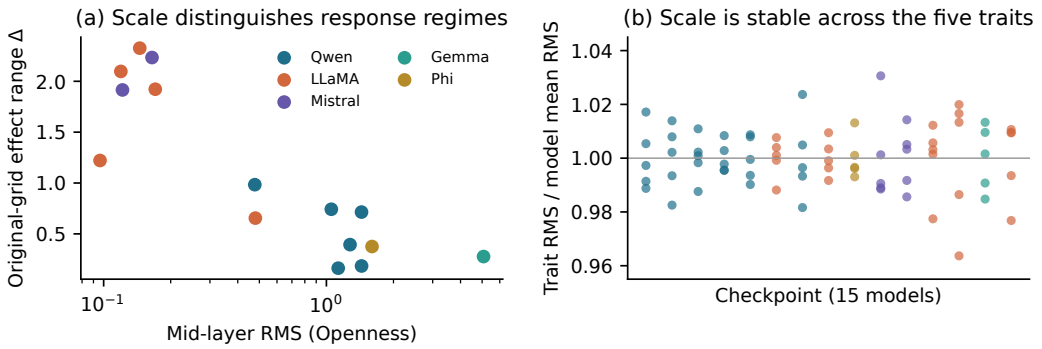

Figure 2: **Scale varies across checkpoints and is stable across attributes.** (a) Openness RMS measured at the standardized middle layer versus the original-grid mean response range. Each point is a checkpoint; the plot shows the measured values, while reported held-out correlations use predictions fitted without the evaluated family. (b) RMS for each of five traits, divided by that checkpoint's mean. Each column is a model. The median within-model coefficient of variation is 1.03%.

Judging and statistical units. Gemma-2-2B-it scores each response using trait-specific high/low indicators and a 1–5 rubric. Ratings quantify expressed personality. General task capability and fluency are separate outcomes from the control objective studied here. A second judge, Qwen2.5-7B-Instruct, scores a 1,260-response subset of the original-grid survey. Appendix E specifies its sampling and the rating parser. For calibration, the unit of aggregation is the model–trait pair, with equal model weighting. A model bootstrap uses 20,000 resamples of the five calibrated checkpoints, retaining all their traits, to quantify uncertainty conditional on the measured scenarios and candidate grids.

5 Results

5.1 Residual Scale Identifies Different Sensitivity Regimes

Under the shared coefficient grid, the mean response range spans a factor of more than fourteen, from 0.163 for Qwen2.5-14B to 2.326 for Llama-3-8B. Figure 2a relates these outcomes to a single mid-layer RMS measurement. Checkpoints with larger residual scale generally express smaller behavioral ranges under the shared raw coefficients. This relationship gives an actionable interpretation of a small observed steering effect: the intervention range is a variable to calibrate alongside the extracted direction.

For held-out prediction, we fit $\log \bar{\Delta}_m = \beta_0 + \beta_1 \log R_m$ using the other model families, then predict every checkpoint in the excluded family. Pooling predictions across the five held-out families gives Spearman $\rho = 0.829$ and median absolute error 0.169 rating points. Leave-one-model-out prediction gives $\rho = 0.739$. The main predictor uses a fixed measurement-depth rule. Probe-site and injection-site measurements supply complementary variants, with held-out-family correlations of 0.846 and 0.871 respectively (Table 1). This analysis supports model prioritization under the measured benchmark policies.

At the standardized middle layer, the within-model RMS coefficient of variation across traits ranges from 0.50% to 2.17%, with median 1.03% (Fig. 2b). Thus the broad scale regime can be measured on one attribute and reused as context for another, while the operating dose remains attribute-specific because scale does not specify the entire dose-response curve. This lets the five control tasks share one candidate range, with the model supplying common information about the receiving state and each attribute contributing its own behavioral evidence. The following paired intervention tests the resulting choice at fixed model and layer.

Table 1: **Alternative scale features.** All features use Openness activations. RMS norm is $\sqrt{d}$ times coordinate RMS. LOMO and LOFO denote leave-one-model-out and leave-one-family-out prediction. The fixed middle-layer feature is used in the main text.

Feature	LOMO ρ	LOFO ρ	LOFO median absolute error
Mid-layer RMS	0.739	0.829	0.169
Probe-layer RMS	0.771	0.846	0.344
Injection-layer RMS	0.793	0.871	0.200
Probe-layer RMS norm	0.875	0.825	0.343
Injection-layer RMS norm	0.854	0.886	0.213

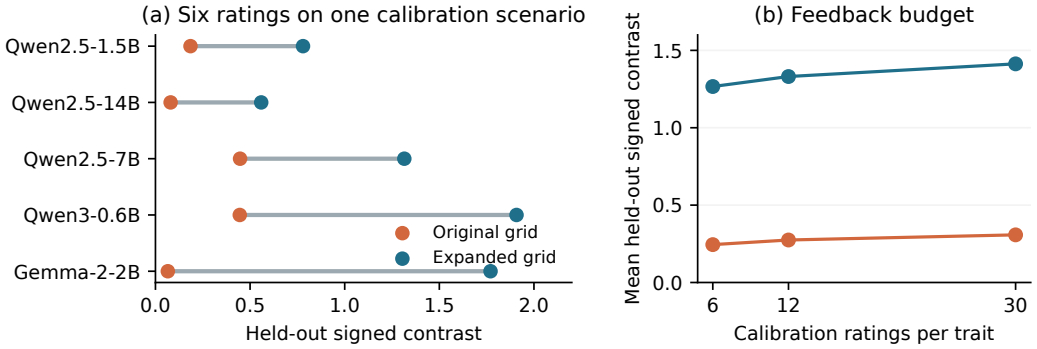

Figure 3: **A useful dose transfers beyond its calibration scenario.** (a) Six response ratings select one of three symmetric dose pairs on one scenario. Points report the mean signed contrast on the other scenarios, averaged over five traits and every calibration-scenario and repetition choice. Original and expanded grids receive identical feedback budgets. (b) Increasing repetitions on the calibration scenario gives a gradual improvement; six ratings recover 89.6% of the expanded-grid effect obtained with 30. These are retrospective evaluations of fixed measured grids.

5.2 Six Ratings Select a Dose that Transfers Across Scenarios

Figure 3 and Table 2 report the main matched-budget comparison. With six calibration ratings, the expanded grids reach a mean held-out signed contrast of 1.266, up from 0.245 for the original-grid selector. The paired mean increase is 1.022 rating points, with model-bootstrap 95% interval $[0.603, 1.441]$. Each of the 25 model–trait averages has a positive expanded-grid effect and improves over its original-grid counterpart. These averages integrate all available choices of the calibration scenario and stored repetition.

The gain appears across different checkpoints: Qwen2.5-14B reaches 0.559, Qwen2.5-7B reaches 1.315, and Gemma-2-2B reaches 1.771. The same candidate range is used for all five traits of a model, so this comparison measures the value of a reusable model-level range followed by inexpensive attribute-level selection. Full per-trait results appear in Appendix Table 5.

The improvement also spans the five control objectives. Averaging the six-rating results over models gives expanded-grid effects of 1.409 for Openness, 1.492 for Conscientiousness, 1.338 for Extraversion, 0.951 for Agreeableness, and 1.142 for Neuroticism. Their original-grid counterparts lie between 0.175 and 0.312. Thus the aggregate gain reflects several kinds of expressed behavior, including exploratory, organized, social, cooperative, and stress-reactive responses. The per-trait view complements the per-model table: a shared range supports multiple attributes, while selection adapts the operating point within that range.

The zero-feedback largest-dose control clarifies the source of these gains. On four checkpoints, the largest expanded amplitude gives a higher mean effect than the six-rating selector, showing that establishing an effective range already supplies substantial control. On Qwen3, selection raises the mean effect from 0.556 at the largest amplitude to 1.908. Across the five models, the corresponding means are 1.172 and 1.266. The transferable advantage over the original grid holds across all models,

Table 2: **Held-out directional control under explicit feedback budgets.** Values average five traits and all calibration splits. The first two columns apply the same six-rating selector to different candidate grids. The 30-rating column uses five repetitions on the same calibration scenario. The last column always chooses the largest expanded amplitude and uses no selection feedback. All columns evaluate the signed contrast, not the unsigned response range.

Model	Original 6 ratings	Expanded 6 ratings	Expanded 30 ratings	Expanded largest dose
Qwen2.5-1.5B	0.185	0.780	0.987	0.893
Qwen2.5-14B	0.080	0.559	0.691	0.839
Qwen2.5-7B	0.447	1.315	1.445	1.499
Qwen3-0.6B	0.446	1.908	2.001	0.556
Gemma-2-2B	0.065	1.771	1.941	2.070
Mean	0.245	1.266	1.413	1.172

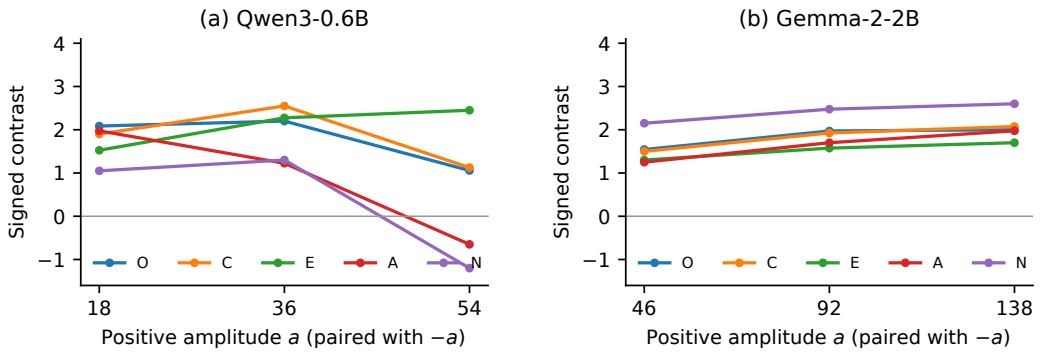

Figure 4: **Different dose curves call for different operating points.** Mean signed contrast for each of the five traits on the expanded grids. O/C/E/A/N denote the Big Five attributes. Qwen3 has useful interior amplitudes and non-monotonic curves; Gemma's larger amplitudes generally support stronger directed effects. These descriptive curves use all scenarios. The selection results in Table 2 hold test scenarios out of dose choice.

while the additional benefit of choosing within the expanded grid is concentrated in a checkpoint with a non-monotonic directional response.

5.3 Dose Curves Explain the Value of an Operating Point

Expanding the grid raises the five-model average response range from 0.458 to 1.719. This establishes that the existing directions can access a substantially broader range of expressed behavior. A directional operating point asks a more specific question: which positive/negative pair should be deployed? Figure 4 shows two instructive cases using every trait and measured positive amplitude.

For Qwen3, the expanded-grid range is 2.630, but the signed contrast at the largest pair is 0.556. Several trait curves reach their strongest directed effect at an interior amplitude, and some reverse order at the largest pair. A single calibration scenario provides enough information for the selector to retain a mean held-out effect of 1.908. For Gemma, the largest pair gives a strong mean signed contrast of 2.070, and sparse selection reaches 1.771. These different curve shapes explain why range expansion and operating-point selection should be evaluated separately.

These curves show why the intervention sign belongs in the objective: selecting and testing a symmetric pair evaluates whether a high/low attribute switch transfers in the requested direction.

5.4 Additional Feedback Refines an Already Useful Choice

Using two repetitions per candidate sign increases the budget from six to 12 ratings and raises the expanded-grid mean effect from 1.266 to 1.331. Five repetitions use 30 ratings and achieve 1.413. Thus six ratings retain 89.6% of the 30-rating effect with one fifth of its selection feedback. The original-grid selector reaches 0.245, 0.275, and 0.308 under the same budgets. Additional repetitions improve estimation within each range; they do not substitute for access to an effective range.

Calibration on multiple scenarios gives further improvement. Using all five repetitions on two or three calibration scenarios costs 60 or 90 ratings, respectively; Appendix C reports these results alongside uniform candidate selection. With two calibration scenarios, the expanded-grid mean reaches 1.450; three scenarios give 1.480. Uniform selection among the three expanded candidates yields an expected contrast of 1.187 without feedback. Together with the fixed-largest-dose control, these values locate the six-rating result between an unconditioned choice and more extensively informed selection. The comparison measures the value of task-relevant feedback within the same candidate set.

The budget curve concerns selecting a reusable control setting. Its cost is paid when calibrating a model–attribute pair, while the held-out evaluation tests whether that setting remains useful elsewhere. This makes the six-rating operating point attractive when many subsequent scenarios share the same requested attribute. The experiment measures that reuse across six or seven test scenarios, and additional repetitions progressively refine the initial choice.

5.5 Supporting Checks on Representation and Measurement

Scenario-grouped probe checks assess whether the contrastive prompts encode readable attributes. For the 160-row collections, all 16 paired rows from one scenario are held out together. Across 45 available model–trait collections, the mean accuracy is 0.965. The 24 available 20-row collections yield mean pair-held-out accuracy 0.918. These checks use the existing selected layers; Appendix D reports each score and its grouping rule. They establish that the activation collections contain readable trait information while the intervention experiments determine useful control settings.

The second judge gives sample-level Pearson and Spearman correlations of approximately 0.77 with the primary judge, with 93.4% of scores within one point. On this subset, the two judges' model-level effect-range rankings have Spearman correlation 0.951. This check supports the broad ordering of original-grid model sensitivity. Its sampling uses the first scenario and three repetitions at each nonzero coefficient; the calibrated-grid transfer results use the primary judge. Excluding parser-fallback ratings yields an expanded-grid six-rating effect of 1.283, with all 25 pair averages still improving over the original-grid selector. Appendix E gives the protocol and conservative rating-uncertainty bounds.

6 Implications

The results support a concrete workflow: measure the checkpoint's scale, establish a candidate range, and select a signed operating pair with task-relevant feedback. A shared scale regime supports reuse across attributes, while the selected dose follows the attribute's response curve. Our transfer evaluation covers selection within established grids. Extending the workflow to unseen checkpoints calls for evaluating range discovery as well as operating-point selection, with application-specific quality constraints alongside the signed control objective.

7 Conclusion

Activation scale and behavioral feedback resolve different parts of steering calibration. A standardized scale measurement predicts cross-model sensitivity and is stable across attributes within a checkpoint. Given measured candidate grids, six ratings on one scenario select a directional dose that transfers to other scenarios, improving every model–attribute average over the original-grid selector. The dose curves explain where range expansion supplies control and where an interior operating point matters. Together, these findings make low-feedback calibration a concrete route from an extracted direction to useful behavioral steering.

AI USE STATEMENT

Generative AI tools assisted with retrospective analysis design and implementation, interpretation of experimental results, manuscript organization and revision, and figure preparation. Language models generated and rated synthetic scenario responses using the protocol described in the paper. Reported measurements are computed from experimental records. The authors are responsible for the analyses, interpretations, and final manuscript.

ETHICS STATEMENT

This work studies expressed personality in responses to synthetic scenarios using public language-model checkpoints. It involves no human participants or personal data. Personality labels describe the response rubric and do not imply psychological properties of a model. Behavioral steering can affect how users interpret a system's responses; applications should communicate when such controls are active.

REPRODUCIBILITY STATEMENT

The appendix specifies the model inventory, injection layers, coefficient grids, rating protocol, selection rule, and aggregation procedure. Accompanying materials provide scenario definitions, experiment configurations, and source code for generating model responses and computing response summaries, exhaustive calibration splits, and figures. The sparse-feedback results are retrospective analyses of stored responses; no new model generations were required for this evaluation.

REFERENCES

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025. doi: 10.48550/arXiv.2507.21509. URL https://arxiv.org/abs/2507.21509.

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. In *International Conference on Learning Representations (ICLR)*, 2026.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Ma, and Yixin Liang. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, O Rogov, Ivan Oseledets, and Elena Tutubalina. The rogue scalpel: Activation steering compromises llm safety, 2026. *URL https://arxiv.org/abs/2509.22067*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Alexander Matt Turner, et al. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

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

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8189–8200, 2024. arXiv:2410.04962.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. Layernavigator: Finding promising intervention layers for efficient activation steering in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2408.11779.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Model Inventory and Generation Protocol

Table 3 lists all checkpoints and zero-based injection sites. The middle measurement site is $\lfloor L/2 \rfloor$. As specified in Sec. 3, a unit direction is added at the selected block output on each forward pass.

Table 3: **Complete primary-survey inventory.** K is the pair count recorded in the scale profiles (20/160 denotes variation across traits); L is depth and d is hidden width. Injection sites are in A/C/E/N/O order. RMS is measured for Openness at the standardized middle layer. $\bar{\Delta}$ averages five traits on the original grid. The base Llama-2 checkpoint is excluded from this instruction-tuned survey.

Checkpoint	K	L	d	Injection layers	RMS	$\bar{\Delta}$
Qwen2.5-0.5B	20	24	896	11,15,5,15,11	0.477	0.984
Qwen2.5-1.5B	160	28	1536	14,14,14,14,14	1.277	0.394
Qwen2.5-14B	160	48	5120	24,24,24,24,24	1.130	0.163
Qwen2.5-7B	160	28	3584	14,14,14,14,14	1.052	0.742
Qwen3-0.6B	20/160	28	1024	14,14,14,14,14	1.437	0.714
TinyLlama-1.1B	20	22	2048	21,14,13,19,13	0.096	1.221
DeepSeek-R1-Distill-14B	160	48	5120	24,24,24,24,24	1.438	0.184
Llama-2-7B	160	32	4096	16,16,16,16,16	0.480	0.655
Phi-3-mini	160	32	3072	16,16,16,16,16	1.602	0.376
Mistral-7B-v0.1	160	32	4096	8,4,13,8,5	0.165	2.234
Mistral-7B-v0.2	160	32	4096	16,16,16,16,16	0.121	1.915
Llama-3.1-8B	160	32	4096	12,4,14,10,14	0.171	1.922
Llama-3.2-1B	20	16	2048	8,15,6,7,6	0.119	2.097
Gemma-2-2B	20	26	2304	13,13,13,13,13	5.075	0.277
Llama-3-8B	160	32	4096	12,13,13,12,14	0.145	2.326

Checkpoint identifiers. `Qwen/Qwen2.5-{0.5B,1.5B,7B,14B}-Instruct`
`Qwen/Qwen3-0.6B`
`TinyLlama/TinyLlama-1.1B-Chat-v1.0`
`deepseek-ai/DeepSeek-R1-Distill-Qwen-14B`
`meta-llama/Llama-2-7b-chat-hf`
`microsoft/Phi-3-mini-128k-instruct`
`mistralai/Mistral-7B-Instruct-v0.{1,2}`
`unsloth/Llama-3.1-8B-Instruct`
`unsloth/Llama-3.2-1B-Instruct`
`unsloth/gemma-2-2b-it`
`unsloth/llama-3-8B-Instruct`

Contrastive collections. The 160-row collections form a Cartesian product within each of ten scenarios: four high formulations and four low formulations generate 16 paired rows. Consequently, a positive activation is repeated across the four low formulations, and a negative activation across the four high formulations. Mean-difference extraction weights each formulation equally. The 20-row collections use 20 stored high/low pairs. The Qwen3 Openness scale profile records 160 pairs, whereas its available raw array contains 20 pairs; the probe check uses those 20 rows and the scale analysis uses the stored profile. We retain the actual collection size for each checkpoint and distinguish scenario-grouped from pair-held-out diagnostics in Appendix D.

Behavioral scenarios. The behavioral evaluation uses seven Openness scenarios and eight scenarios for each other trait. Responses are generated at temperature 0.8, top-p 0.9, with a 200-new-token cap and five repetitions per coefficient and scenario. The scenario prompts are distinct from the contrastive examples used to extract the direction. The accompanying `evaluation_prompts.json` supplies the exact synthetic scenarios and rubric indicators. Scenario holdouts in the calibration analysis refer to these behavioral scenarios.

Table 4: **Dose intervention configuration.** All traits of a checkpoint share the same candidate grid. Zero is measured for descriptive response ranges and is not a candidate in symmetric-pair selection. Original and expanded runs use the same listed injection layer.

Model	Layer	Original positive amplitudes	Expanded positive amplitudes
Qwen2.5-1.5B	14	$2, 4, 6$	$10, 20, 30$
Qwen2.5-7B	14	$2, 4, 6$	$10, 20, 30$
Qwen2.5-14B	24	$2, 4, 6$	$10, 20, 30$
Qwen3-0.6B	14	$2, 4, 6$	$18, 36, 54$
Gemma-2-2B	13	$2, 4, 6$	$46, 92, 138$

B Cross-Model Prediction and Scale Definitions

Let $x_m = \log_{10} R_m$ and $y_m = \log_{10} \bar{\Delta}_m$. For each held-out model or family, ordinary least squares fits $y = \beta_0 + \beta_1 x$ on the remaining models. Predictions are transformed back to rating units as $10^{\hat{y}}$. We compute Spearman correlation across all 15 out-of-fold predictions and the corresponding measured effects. Median absolute error uses these transformed predictions. Family groups are Qwen (including the DeepSeek Qwen distillation), LLaMA (including TinyLlama), Mistral, Gemma, and Phi. Model count, rather than response count, determines the number of prediction units.

The injection-site feature uses the site's recorded RMS; the probe-site feature uses the existing Openness probe-selected site. These alternatives measure different locations in the same receiving network. The fixed-middle feature avoids attribute-specific site selection in the measurement itself. The behavioral outcome retains the benchmark's actual injection policies listed in Table 3. Among the nine models whose primary behavior was measured at a fixed middle site for all traits, the inverse-RMS rank correlation is 0.60. This policy-restricted subset separates the breadth of the full survey from evidence under a uniform depth rule.

Within-model stability. For each checkpoint, we compute $\mathrm{std}_t(R_{m,t}) / \mathrm{mean}_t(R_{m,t})$ over five traits at the same middle layer, using population standard deviation. The observed interval is 0.0050–0.0217, with median 0.0103. This is a property of the contrastive collections at a fixed site. Measuring different trait-specific layers would combine attribute variation with layer variation.

B.1 Local Response of RMS Normalization

Write $z(\alpha) = h + \alpha\hat{v}$ and $R(\alpha) = \sqrt{z(\alpha)^\top z(\alpha)/d + \epsilon}$. Differentiation gives

$$R'(\alpha) = \frac{z(\alpha)^\top \hat{v}}{dR(\alpha)}. \tag{8}$$

Using the quotient rule for $F(\alpha) = z(\alpha)/R(\alpha)$,

$$F'(\alpha) = \frac{\hat{v}}{R(\alpha)} - \frac{z(\alpha)\, z(\alpha)^\top \hat{v}}{dR(\alpha)^3}. \tag{9}$$

Evaluating at zero yields Eq. 3. If the learned RMSNorm gain is g, the derivative is multiplied elementwise by g. When h and $\hat{v}$ are orthogonal, the alignment term vanishes and the pre-gain derivative norm is $1/R(h)$ for a unit direction. This is a local calculation at one normalization operation. The empirical scale predictor tests whether a related association appears in the complete model, and the measured dose curves characterize the finite amplitudes used for control.

C Exhaustive Calibration Splits and Additional Budgets

Selection and transfer. Let a^* maximize $G_t(a; S_{\text{test}})$ over the candidate set. By the definition of ε in Sec. 3.4,

$$\begin{aligned} G_t(a^*; S_{\text{test}}) &\le \widehat{G}_t(a^*; S_{\text{cal}}) + \varepsilon \\ &\le \widehat{G}_t(\hat{a}_t; S_{\text{cal}}) + \varepsilon \le G_t(\hat{a}_t; S_{\text{test}}) + 2\varepsilon. \end{aligned}$$

The middle inequality follows from the calibration argmax. This proves Eq. 7 for the fixed split, with no distributional assumption. The empirical results evaluate the selected effect rather than estimating this worst-candidate deviation.

Calibration configurations. The six-rating analysis has $5S_t$ calibration configurations per model–trait pair, where S_t is seven or eight: choose one scenario and one of its five stored repetition indices. Each configuration evaluates three positive/negative coefficient pairs and selects the largest estimated signed contrast. The 12-rating analysis enumerates $\binom{5}{2}S_t$ configurations. The 30-rating analysis uses all five repetitions and has S_t configurations. Numerical ties within 10^{-12} choose the smaller amplitude.

Every configuration evaluates the selected amplitude on all remaining scenarios using their five stored repetitions. The reported score is the mean across calibration configurations. Thus a scenario can serve as training data in one configuration and test data in another, but its outcomes are never used for selection in the configuration where it is held out. Configurations overlap and are not independent samples; we aggregate them before uncertainty estimation.

Table 5: **Every model–trait result.** Signed contrast on scenarios excluded from selection. O/C/E/A/N denote Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each cell averages all available calibration configurations.

		6 calibration ratings		30 calibration ratings	
Model	Trait	Original	Expanded	Original	Expanded
Qwen2.5-1.5B	O	0.290	0.979	0.390	1.367
Qwen2.5-1.5B	C	0.101	1.470	0.246	1.686
Qwen2.5-1.5B	E	0.005	0.416	-0.054	0.500
Qwen2.5-1.5B	A	0.139	0.269	0.171	0.382
Qwen2.5-1.5B	N	0.391	0.764	0.350	1.000
Qwen2.5-14B	O	0.014	0.853	-0.005	1.124
Qwen2.5-14B	C	0.034	0.362	0.029	0.389
Qwen2.5-14B	E	0.230	0.866	0.186	0.925
Qwen2.5-14B	A	0.064	0.201	0.086	0.289
Qwen2.5-14B	N	0.059	0.511	0.075	0.729
Qwen2.5-7B	O	0.449	1.402	0.529	1.567
Qwen2.5-7B	C	0.211	1.561	0.268	1.889
Qwen2.5-7B	E	0.835	1.726	1.061	1.711
Qwen2.5-7B	A	0.239	0.868	0.271	1.075
Qwen2.5-7B	N	0.502	1.019	0.625	0.982
Qwen3-0.6B	O	0.486	2.115	0.686	2.114
Qwen3-0.6B	C	0.451	2.330	0.571	2.439
Qwen3-0.6B	E	0.399	2.209	0.536	2.318
Qwen3-0.6B	A	0.488	1.817	0.496	1.975
Qwen3-0.6B	N	0.404	1.068	0.518	1.157
Gemma-2-2B	O	0.031	1.697	0.167	1.971
Gemma-2-2B	C	0.079	1.735	0.111	1.889
Gemma-2-2B	E	-0.056	1.471	-0.075	1.511
Gemma-2-2B	A	0.071	1.602	0.100	1.854
Gemma-2-2B	N	0.203	2.347	0.357	2.479

Multiple calibration scenarios. For $k = 2$ or 3, we enumerate every subset of k scenarios, use all five repetitions at each candidate sign, and evaluate on the complement. The budget is $30k$ ratings. For the final row of Table 6, selection uses all but one scenario, costing 180 ratings for Openness and 210 for other traits. All rows use three symmetric candidate pairs. Different rows also change the number of held-out scenarios, and their score averages are reported for their respective exhaustive splits.

Uncertainty. For the six-rating comparison, we first subtract original-grid performance from expanded-grid performance within each model–trait pair and average the five traits within each model. We resample these five model-level differences with replacement 20,000 times. The percentile interval is $[0.603, 1.441]$ for a mean difference of 1.022. The random seed is 20260926 in the

Table 6: **Additional calibration budgets and fixed-grid controls.** Signed held-out contrast, averaged equally over model–trait pairs. Uniform selection is the expected score of choosing one of three expanded candidates at random.

Selection rule	Feedback ratings	Original	Expanded
1 scenario	30	0.308	1.413
2 scenarios	60	0.336	1.450
3 scenarios	90	0.348	1.480
All but one scenario	180/210	0.360	1.537
Largest amplitude	0	0.406	1.172
Uniform expanded candidate	0	–	1.187

accompanying analysis script, which also specifies the order of the budget analyses. This interval conditions on the observed scenarios, candidate grids, and model set; it does not treat the thousands of ratings or overlapping calibration configurations as independent model replications.

D Representation Diagnostics with Scenario Grouping

We recompute linear probe accuracy at each available collection's existing selected layer using logistic regression with $C = 0.01$, the liblinear solver, a maximum of 1,000 iterations, and random seed 42. For 160-row collections, the repeated positive/negative activation structure is checked before forming ten scenario groups. An entire 16-row paired block is excluded per fold, keeping its high and low formulations out of training together. For 20-row collections, one high/low pair is held out per fold. No layer is reselected inside these checks, so the scores diagnose readability at the selected sites rather than nested generalization of a layer-selection procedure.

There are 45 available scenario-grouped collections, with mean accuracy 0.965 and range 0.825–1.000, and 24 pair-held-out collections, with mean 0.918 and range 0.750–1.000. Six collections lack saved arrays for this recomputation. Table 7 gives the full inventory; these absent diagnostic scores do not affect the recorded behavioral results.

Table 7: **Probe checks at existing selected layers.** Group identifies the unit excluded in each fold. A dash indicates that a saved activation array was unavailable for recomputation.

Model	Group	O	C	E	A	N
Qwen2.5-0.5B	pair	0.800	0.925	0.850	0.875	0.850
Qwen2.5-1.5B	scenario	0.975	0.863	0.912	0.975	0.975
Qwen2.5-7B	scenario	1.000	0.938	0.963	0.825	0.988
Qwen2.5-14B	unavailable	–	–	–	–	–
Qwen3-0.6B	pair	0.775	0.975	1.000	0.975	0.925
TinyLlama-1.1B	pair	–	0.750	0.750	0.800	0.800
DeepSeek-R1-Distill-14B	scenario	0.988	0.975	0.950	0.938	0.988
Llama-2-7B	scenario	1.000	1.000	1.000	1.000	0.988
Phi-3-mini	scenario	0.912	0.900	0.875	0.963	0.925
Mistral-7B-v0.1	scenario	0.925	0.975	0.988	0.975	0.975
Mistral-7B-v0.2	scenario	1.000	0.988	1.000	1.000	1.000
Llama-3.1-8B	scenario	1.000	0.887	0.988	1.000	0.938
Llama-3.2-1B	pair	1.000	1.000	1.000	1.000	1.000
Gemma-2-2B	pair	1.000	1.000	1.000	1.000	0.975
Llama-3-8B	scenario	1.000	1.000	1.000	0.975	0.988

E Judging Protocol and Independent Scoring

The primary judge is Gemma-2-2B-it. It receives the generated response and the target trait's description and high/low indicators. It does not receive the coefficient, the model identity, or the original behavioral scenario in the rating prompt. The rating prompt uses the following template:

You are a psychologist evaluating a person's response for the personality trait "[Trait]".
TRAIT: [Trait]
Description
HIGH [TRAIT] BEHAVIORS:
High indicators
LOW [TRAIT] BEHAVIORS:
Low indicators
PERSON'S RESPONSE:
"[Response]"
Rate how much this response reflects [Trait] on a 1–5 scale:
1 = Very Low (clearly shows LOW indicators)
2 = Low (mostly shows LOW indicators)
3 = Moderate (mixed or neutral)
4 = High (mostly shows HIGH indicators)
5 = Very High (clearly shows HIGH indicators)
First write one sentence of reasoning, then output your rating on a new line.
Format: RATING: [1–5]

The judge samples at temperature 0.1 with a 100-new-token cap. The parser first extracts the digit following `RATING:`, then attempts a rating/score expression, and finally searches the last 50 characters for a 1–5 digit. If no match is found, it returns 3. Stored numerical ratings therefore include the parser's neutral fallback. All 20,475 primary-survey ratings have an explicit rating marker. Of 6,825 expanded-grid ratings, 6,761 have an explicit marker and 64 use the neutral fallback.

Parser sensitivity. We repeat the six-rating analysis after excluding calibration configurations that contain any fallback among their six required ratings, and excluding fallback ratings from each held-out scenario's endpoint means. This retains 917 of 975 calibration configurations across the 25 model–trait pairs. We apply the same retained calibration configurations to the original-grid selector. The resulting mean signed contrasts are 1.283 and 0.247, respectively, and all 25 pair averages improve with the expanded grid. A second analysis allows every fallback to take any value in $[1, 5]$. For each calibration configuration, we identify every amplitude that could maximize the training contrast under those intervals, respecting the smaller-amplitude tie rule. We then take the smallest possible held-out contrast over those amplitudes, assigning uncertain positive-endpoint ratings to 1 and negative-endpoint ratings to 5. Averaging these per-configuration bounds gives a conservative expanded-grid mean lower bound of 1.194, compared with 0.245 for the original-grid selector; 24 of 25 pairwise lower bounds exceed their original-grid scores. The bound permits independently adverse assignments across overlapping configurations and is therefore conservative for any single assignment to the shared records. The accompanying `parser_sensitivity.py` reproduces both checks.

Trait indicators. Openness contrasts curiosity, abstraction, and creative exploration with routine, concrete detail, and preference for proven methods. Conscientiousness contrasts planning, detail, and follow-through with procrastination, improvisation, and reduced emphasis on commitments. Extraversion contrasts social enthusiasm and high energy with solitude and reserved expression. Agreeableness contrasts empathy, cooperation, and trust with competition, skepticism, and confrontation. Neuroticism contrasts worry, volatility, and stress sensitivity with calm, resilience, and self-assurance. The exact indicator lists accompany the scenario definitions in the analysis materials.

Second judge. Qwen2.5-7B-Instruct independently scores 1,260 responses: 14 models, five traits, six nonzero original-grid coefficients, and the first three stored repetitions of the first scenario. Mistral-v0.2 is outside this subset. The response-level Pearson correlation is approximately 0.774, Spearman correlation 0.772, mean absolute difference 0.708, and within-one-point agreement 0.934. For each judge, we compute an effect range across the six coefficients for each model–trait pair and average five traits per model. The resulting model rankings agree at Spearman $\rho = 0.951$. These results concern original-grid sensitivity on the specified scenario subset.