PREDICTIVE ACTIVATION STEERING VIA SOLUTION MAPPING AND STOCHASTIC SPECTRAL EXPANSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering is a promising inference-time intervention that modulates large language models (LLMs) by perturbing their internal activations during generation. However, its practical application requires determining whether an effective steering intervention can be identified and quantifying trade-offs among suitable interventions. In this work, we propose an inexpensive framework for identifying optimized steering interventions subject to specified requirements. We model the steering coefficients as Gaussian random variables and approximate their effects on measured LLM behavior using a second-order Hermite polynomial expansion. Then, optimization is applied to find steering coefficients predicted to jointly satisfy specified targets. The resulting covariance matrix reveals structure within the steering coefficients, including coupling between coefficients and the degree to which they are jointly constrained. We propose a feasibility test that identifies unattainable specifications and systematically determines which requirements must be removed. We demonstrate the benefits of our approach through a case study and show that it performs competitively with existing methods for precise behavioral control of LLMs.

1 INTRODUCTION

Large Language Models (LLMs) have achieved remarkable success across a wide range of disciplines and are being rapidly deployed in critical applications spanning healthcare, education, and scientific research (Lee et al., 2023; Hwang & Chang, 2023; Gottweis et al., 2026). Despite this success, understanding of their internal mechanisms remains incomplete (Ferrando et al., 2024; Bereska & Gavves, 2024), motivating methods to probe the generation process and align model outputs with desired behavior (Zou et al., 2025). Activation steering, an inference-time mechanism for controlling the characteristics of generated text, is one such approach (Subramani et al., 2022; Turner et al., 2024; Zou et al., 2025). Steering often relies on contrasting prompts to elicit steering vectors or latent directions associated with a specific behavior (Rimsky et al., 2024; Arditi et al., 2024). During generation, activations are perturbed along this direction to promote or suppress the corresponding characteristic. This methodology has proven effective for steering text toward arbitrary concepts, including many of direct relevance to AI safety and alignment (Zou et al., 2025; Turner et al., 2024; Arditi et al., 2024; Rodriguez et al., 2024).

However, activation steering can be severely limited by oversteering, the substantial degradation of generated text when activations are perturbed too far, as well as an inability to predict an intervention's impact before it is applied (Vogels et al., 2025; Tan et al., 2024). Parameter sweeps are typically used to evaluate the sensitivity of steering interventions (Rimsky et al., 2024; Konen et al., 2024; Vogels et al., 2025), but such sampling does not supply a principled mapping between desired behavioral outcomes and the coefficients that achieve them. Unintended side effects are also largely neglected, particularly when multiple behaviors are steered jointly, since joint steering introduces interactions between objectives that make it difficult to determine whether a requested combination is achievable. An additional issue, noted by Vogels et al. (2025), Skifstad et al. (2026), and Hao et al. (2025), is that smaller models are more susceptible to oversteering. However, smaller models are desirable for constrained hardware and latency-sensitive applications in which iterative procedures such as prompt tuning are often cost-prohibitive.

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.

Therefore, we investigate small open-weight language models and treat activation steering as a calibration problem. For each model, we construct a low-order solution map relating steering coefficients to measured behavioral outcomes using a second-order Hermite polynomial expansion. This surrogate provides a forward model for predicting behavioral changes resulting from steering interventions. Optimization using this forward model can then determine steering coefficients that simultaneously satisfy multiple behavioral requirements. This procedure also yields a covariance matrix that characterizes coefficient couplings and distinguishability.

The specific problems we address and contributions of the work are summarized as follows:

- A forward map relating steering coefficients to LLM behaviors, combined with penalized optimization, predicts the effect of a steering intervention before it is applied, improving steering reliability.
- A feasibility test reveals and rejects targets that cannot be jointly satisfied, clarifying the limits of steering interventions.
- A covariance matrix of the steering coefficients quantifies correlations and their coupled impact on response quantities. This enables exploration of trade-offs in near-optimal steering interventions.

2 Related Work

2.1 LLM Behavior Control and Activation Steering

The most common alignment approaches, reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022), supervised finetuning (Wei et al., 2022), and preference optimization (Rafailov et al., 2023; Xu et al., 2024; Fan et al., 2025; Liu et al., 2025), require additional training and can unintentionally degrade model capabilities (Kirk et al., 2023; Luo et al., 2025).

Activation steering instead identifies directions in activation space that can be manipulated at inference time, requiring no additional training. Directions can be derived in many ways, including contrastive activation differences (Arditi et al., 2024; Rimsky et al., 2024; Turner et al., 2024; Chen et al., 2025), sparse autoencoders (SAEs) (Gao et al., 2024; Cunningham et al., 2023; Chalnev et al., 2024), and optimization-based extraction (Subramani et al., 2022). The application of activation steering has also been increasingly formulated as a dynamical control problem, including state-dependent, adaptive, and feedback-based interventions (Kong et al., 2024; Cheng & Alonso, 2026; Zhao et al., 2026; Nguyen et al., 2025a; Skifstad et al., 2026). Studies of multi-attribute steering show that combining directions can introduce interference, cancellation, or reduced effectiveness across target behaviors (van der Weij et al., 2024; Nguyen et al., 2025b; Oozeer et al., 2026; Ghasemi et al., 2026; Scalena et al., 2024).

Key limitations of activation steering include performance that cannot be predicted a priori across input prompts and inconsistent generalization beyond the setup in which the directions were extracted (Tan et al., 2024; Braun et al., 2024). Large-scale benchmarking has also found that prompting and finetuning can outperform representation-based steering methods (Wu et al., 2025). In contrast, other studies have shown that activation steering provides precise behavioral control and remains effective alongside prompting and finetuning (Konen et al., 2024; Rimsky et al., 2024). These mixed results indicate that steering performance depends on how the intervention is constructed and subsequently applied. Steering interventions have been observed to produce nonlinear and even non-monotonic changes in model behavior (Taimeskhanov et al., 2026), and several recent methods adapt when, where, or how strongly steering is applied during inference (Hedström et al., 2025; Ferrando et al., 2026; Do et al., 2025). Together, these results suggest that functional activation steering requires understanding how steering coefficients interact and map to downstream model responses.

2.2 Solution Mapping and Stochastic Spectral Expansion

Solution mapping represents model responses as inexpensive algebraic functions of model parameters and uses them for joint parameter analysis and inverse optimization (Frenklach et al., 1992;

Sheen & Wang, 2011). This approach exposes relationships among parameters, provides analytical gradients, and supports calibration, sensitivity analysis, and control (Smith, 2013). Polynomial chaos expansions, a form of stochastic spectral expansion, express responses in polynomials orthogonal under the distribution assumed for the inputs (Xiu & Karniadakis, 2002), can be built directly from model evaluations (Reagan et al., 2003), and yield means, variances, and sensitivities from the coefficients alone (Sheen & Wang, 2011). Together, these techniques allow repeated inverse optimization without further model evaluations.

3 METHODOLOGY

3.1 INPUTS, STEERING INTERVENTIONS, AND SPECIFICATIONS

We regard activation steering as a map from steering coefficients to judged characteristics of the generated responses. We consider decoder-only transformers, where the residual stream at layer k and position t, $\boldsymbol{h}_t^{(k)} \in \mathbb{R}^d$, is updated by each consecutive layer with attention and feedforward contributions. Let $\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_K$ be unit steering vectors, each associated with a behavior i and injected at a layer ℓ_i. Let $\boldsymbol{\gamma} \in \mathbb{R}^K$ be the corresponding steering coefficients. During generation the residual stream at layer ℓ is offset by

$$\tilde{\boldsymbol{h}}_t^{(\ell)} = \boldsymbol{h}_t^{(\ell)} + \sum_{i:\, \ell_i = \ell} \gamma_i \hat{\boldsymbol{v}}_i \tag{1}$$

at every generated token position; prompt positions are not modified. The offset propagates through subsequent layers, thereby altering the next-token prediction and generated output (Elhage et al., 2021). Setting $\boldsymbol{\gamma} = \mathbf{0}$ recovers the unsteered model.

Steering vectors are obtained from contrasting prompts (Rimsky et al., 2024; Turner et al., 2024; Chen et al., 2025). Following the convention of Chen et al. (2025), we use five pairs of contrastive system prompts to elicit a high and a low response to every input prompt and for each behavior. The steering vector is then the normalized mean difference of the residual activations over response tokens and retained pairs (Equation 10). A target layer ℓ_i is selected for each behavior and model from a layer-by-layer sensitivity analysis.

For each direction i, we set $\gamma_i^{\max}$ to be the largest coefficient magnitude at which judged coherence remains within 2 points of the unsteered model for both signs of γ_i. Coefficients are then normalized as

$$\xi_i = \gamma_i / \gamma_i^{\max}, \qquad \boldsymbol{\xi} \in [-1, 1]^K. \tag{2}$$

Appendix A gives details of vector construction, layer selection, and the range estimation.

Each generated response is scored on a set of behaviors $\mathcal{R}$ using an integer 0–10 scale. Because scores vary across prompts and stochastic generations, and are subject to potential judge biases, we model the mean score μ_r of behavior r over a fixed prompt set. A user specification consists of a target score τ_r and a tolerance σ_r for each behavior r in a subset $\mathcal{R}_{\text{spec}} \subseteq \mathcal{R}$ with an acceptable range of $\tau_r \pm 2\sigma_r$. Notably, a target may be specified for a behavior that has no steering direction, as couplings among directions can still influence it.

3.2 STOCHASTIC RESPONSE SURFACES

We approximate μ_r by a total-degree-two polynomial in $\boldsymbol{\xi}$. Steering coefficients are treated as independent Gaussian random variables centered at the unsteered model, with $\pm\gamma_i^{\max}$ representing two standard deviations, so that $\xi_i \sim \mathcal{N}(0, 1/4)$. Writing $z_i = 2\xi_i$, we expand the response in Hermite polynomials, which are orthogonal with respect to the Gaussian distribution,

$$\widehat{\mu}_r(\boldsymbol{\xi}) = a_{r,0} + \sum_{i=1}^{K} a_{r,i}\, He_1(z_i) + \sum_{i=1}^{K} a_{r,ii}\, He_2(z_i) + \sum_{i<j} a_{r,ij}\, He_1(z_i) He_1(z_j), \tag{3}$$

where $He_1(z) = z$ and $He_2(z) = z^2 - 1$. The expansion contains 78 coefficients for $K = 11$ steering directions. The degree-two expansion captures curvature and pairwise interactions with a reasonable number of coefficients. Its adequacy is assessed on held-out configurations (Section 5.1).

Steering configurations are sampled from three space-filling Sobol sequences, each mapped through a zero-mean Gaussian with standard deviation 0.1, 0.3, or 0.5 (Zhang et al., 2023), forming three shells. Configurations outside $[-1, 1]^K$ are rejected, and the outermost shell is given twice as many samples as the inner shells. We supplement these samples with sparse configurations that activate one to four directions, thus supporting identifiability of the coupling terms and enabling specifications in which only a few directions are active (Appendix B.1). For configuration k, N_k judged responses give a mean $\bar{\mu}_{r,k}$ and standard deviation $\sigma_{r,k}$ per behavior. Letting $\mathbf{X}$ contain the basis evaluations of Equation 3 at the sampled configurations, the polynomial coefficients are estimated per behavior by regularized and weighted least squares,

$$\widehat{\mathbf{a}}_r = \left(\mathbf{X}^T\mathbf{W}_r\mathbf{X} + \lambda_r s_r \mathbf{P}_2\right)^{-1}\mathbf{X}^T\mathbf{W}_r\overline{\boldsymbol{\mu}}_r, \qquad \mathbf{W}_r = \operatorname{diag}\Big(N_k \,/\, \max(\sigma_{r,k}, \sigma_{\text{floor}})^2\Big). \tag{4}$$

Each configuration is weighted by the inverse variance of its mean, $N_k/\sigma_{r,k}^2$, with $\sigma_{r,k}$ floored at $1/\sqrt{12}$ to prevent unbounded weights. We place a ridge penalty on the second-order terms because they are least informed by the fitting data and therefore the most susceptible to fitting noise. $\mathbf{P}_2$ selects those terms, and s_r is the average diagonal entry of $\mathbf{X}^T\mathbf{W}_r\mathbf{X}$ that corresponds to them. The hyperparameter λ_r governs the trade-off between fitting the measured data and minimizing the second-order terms; its value is selected using five-fold cross-validation (Appendix B).

3.3 Inverse optimization and feasibility

Each user-specified target is treated as an observation of behavior r with mean τ_r and standard deviation σ_r. The coefficients are determined by minimizing the objective

$$\Phi(\boldsymbol{\xi}) = \sum_{r\in\mathcal{R}_{\text{spec}}} \left[(\widehat{\mu}_r(\boldsymbol{\xi}) - \tau_r)/\sigma_r\right]^2 + 4\|\boldsymbol{\xi}\|_2^2, \qquad \boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}\in\mathbb{R}^K} \Phi(\boldsymbol{\xi}). \tag{5}$$

The first term measures deviation from the targets relative to their tolerances, and the second term penalizes intervention magnitude. In Appendix H, we show that a larger-magnitude steering coefficient vector is associated with reduced model capability. The factor 4 is the precision of the coefficient prior, and it ensures that the same penalty is applied to $|\xi_i| = 1$ as to missing a target by $2\sigma_r$ (Sheen & Wang, 2011). We solve this optimization with BFGS (Nocedal & Wright, 2006) using analytic gradients from multiple initializations (Appendix C).

A user may request an unachievable target or a set of irreconcilable targets. To address this problem, we employ a feasibility test (Sheen & Wang, 2011). We normalize each target's residual at the solution by its acceptable deviation,

$$F_r = \frac{\tau_r - \widehat{\mu}_r(\boldsymbol{\xi}^*)}{2\sigma_r}, \qquad r \in \mathcal{R}_{\text{spec}}. \tag{6}$$

A target with $|F_r| \le 1$ is achievable within its specified range and a target with $|F_r| > 1$ is not. When one or more targets are inconsistent, the one that most influenced the solution is removed first, identified using a weighted statistic:

$$W_r = S_r F_r, \qquad S_r = \frac{\nabla\widehat{\mu}_r(\boldsymbol{\xi}^*)^T\boldsymbol{\xi}^*}{\|\nabla\widehat{\mu}_r(\boldsymbol{\xi}^*)\|\,\|\boldsymbol{\xi}^*\|}. \tag{7}$$

S_r measures the alignment between the sensitivity of the target and the intervention. Among the inconsistent targets, the one with the largest $|W_r|$ is removed, the optimization is re-solved, and the test repeats until all remaining targets are jointly achievable.

3.4 Coefficient covariance and propagation to behaviors

We characterize variation around the optimized coefficients by linearizing the response surfaces, yielding a locally quadratic approximation to the objective. Defining $\mathbf{g}_r \equiv \nabla_{\boldsymbol{\xi}}\widehat{\mu}_r(\boldsymbol{\xi}^*)$, this approximation corresponds to a Gaussian distribution centered at $\boldsymbol{\xi}^*$ with covariance

$$\boldsymbol{\Sigma} = \left(4\mathbf{I} + \sum_{r\in\mathcal{R}_{\text{spec}}} (\mathbf{g}_r\mathbf{g}_r^T) \,/\, \sigma_r^2\right)^{-1}, \tag{8}$$

where $4\mathbf{I}$ comes from the prior, and the summation from the targets. The diagonal entries are the variances of the individual coefficients under the specification, and the off-diagonal entries are their covariances, which encode how the coefficients are coupled. Propagating $\mathbf{\Sigma}$ through the response surface of a behavior r' produces the spread

$$\sigma_{r'}^{\text{spec}} = \sqrt{\mathbf{g}_{r'}^T \mathbf{\Sigma}\, \mathbf{g}_{r'} + \tfrac{1}{2}\,\text{tr}((\mathbf{H}_{r'}\mathbf{\Sigma})^2)}, \qquad \mathbf{H}_{r'} = \nabla^2_{\boldsymbol{\xi}} \widehat{\mu}_{r'}, \tag{9}$$

that quantifies the variation in behavior r' among the interventions compatible with the specification (Appendix D).

4 Experimental setup

We calibrate four open-weight language models: Phi-3.5-mini (Abdin et al., 2024), Ministral 8B (Mistral AI, 2024), Llama 3 (8B) (Grattafiori et al., 2024), and Qwen3-8B (Yang et al., 2025) in non-thinking mode. Qwen3-8B is used for the worked examples in the main text; the other three models test the generality of the approach across model families. We consider 13 behaviors: *complexity*, *figurative language*, *formality*, *hedging*, *humor*, *optimism*, *politeness*, *readability*, *refusal*, *sycophancy*, *verbosity*, *helpfulness*, and *coherence*, among which the first 11 are steered. *Helpfulness* and *coherence*. The first 11 are steered; *helpfulness* and *coherence* are judged but have no steering direction.

Evaluation prompts are sourced from eight public datasets separated into six categories: 225 prompts are used for extracting steering directions, a fixed subset of 100 fitting prompts is scored at every design configuration to calibrate the response surfaces, and 132 held-out prompts are reserved for validation. behavior per judge call on an integer 0–10 scale against a fixed rubric by Qwen2.5-14B-Instruct (Team, 2024). We use PubMedQA (Jin et al., 2019) and ARC-Challenge (Clark et al., 2018) as a measure of general model capability and oversteering. We evaluate 400 questions from each dataset and apply steering while the model reasons to an answer.

For each model, we evaluate 249 steering configurations: 149 from three Sobol designs, 99 sparse configurations, and the unsteered baseline. Of these, 201 are used for fitting and 48 for validation. Selected layers and steering ranges are reported in Table 2. Generation uses each model's default sampling settings with a 2048-token limit. Calibration requires $\sim 7 \times 10^4$ generations and $\sim 5 \times 10^5$ judge calls per model. Once the surfaces are calibrated, solving a specification requires no further generations or judge calls.

5 Results

5.1 Held-out predictive accuracy

We first assess the accuracy of the fitted response surfaces. Figure 1 compares predicted and judged scores for the 48 held-out configurations across the 13 behaviors and four LLMs. The mean absolute error, averaged over the 13 behaviors, is 0.32–0.37 on the fitting prompts and 0.34–0.44 on prompts not used in fitting. The few large misses occur at configurations near the edge of the calibrated region, where *coherence* is low and the judged response departs from the quadratic approximation.

5.2 Coupling among steering directions

An analytical response surface enables us to probe the sensitivity of the judged scores to changes in the steering coefficients. Figure 2 shows the matrix of these sensitivities, $\partial\widehat{\mu}_r/\partial\xi_i$, evaluated at the unsteered Qwen3-8B ($\boldsymbol{\xi} = \mathbf{0}$), including the two behaviors without a steering direction (*coherence* and *helpfulness*). Interestingly, for only 6 of the 11 steering directions is the behavior most affected by the one it was extracted for, namely *complexity*, *figurative language*, *formality*, *hedging*, *optimism*, and *verbosity*. For example, the *readability* direction moves *complexity* more than it moves *readability*, with $\partial\widehat{\mu}_{\text{complexity}}/\partial\xi_{\text{readability}} = 2.52$ compared to $\partial\widehat{\mu}_{\text{readability}}/\partial\xi_{\text{readability}} = 1.92$. The same holds for the other three LLMs, where only 6 or 7 of the 11 directions most affect their own behavior (Figure 8). Behaviors without a steering direction of their own can nonetheless be controlled. Illustratively, *helpfulness* exhibits a sensitivity of -0.77 to the *refusal* coefficient, as one might expect, and, perhaps surprisingly, *coherence* exhibits an appreciable sensitivity (0.88) to the

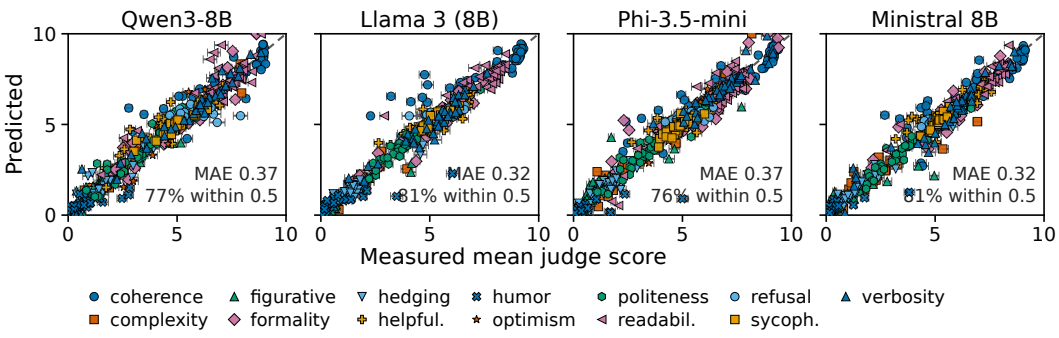


Figure 1: Predicted versus judged scores for the 48 held-out configurations, each averaged over the 100 fitting prompts, across 13 scored behaviors, for each LLM considered. Error bars are standard errors of the judged means.

optimism coefficient. These observations indicate that LLM responses to steering involve a high degree of coupling.

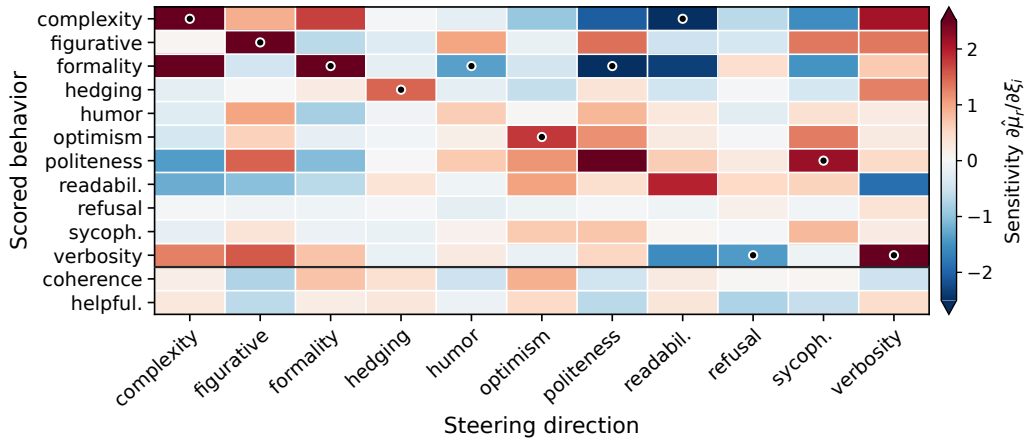


Figure 2: Sensitivity matrix ($\partial\widehat{\mu}_r/\partial\xi_i|_{\boldsymbol{\xi}=0}$) of the 13 judged behaviors with respect to the 11 steering coefficients, computed for unsteered Qwen3-8B. Color scale indicates the sensitivity value, and the dot marks the largest-magnitude entry in each column.

5.3 A CASE STUDY IN REDUCED DIMENSIONS

To illustrate many of the useful features, we devise a test case in reduced dimensionality of the steering space. We steer Qwen3-8B, starting from a joint consideration of five behaviors: *politeness* from an unsteered, judged score of 4.05 to a steered, target score of 5.16, *verbosity* from 6.04 to 4.52, *figurative language* from 1.20 to 3.17, *complexity* from 2.29 to 6.00, and *readability* from 7.76 to 9.80. Figure 3 shows the judged score values of the unsteered model (circles) and targeted scores (red vertical bars) along with their tolerances (orange bands). The targets were chosen well away from the unsteered scores, and the tolerances follow Appendix C. The feasibility test (Equation 6) identifies the targeted *complexity* score as unattainable within the given set of target behaviors ($F_{complexity} = 1.33$), and, having the largest $|W|$, it is removed first. Resolving, we now find *readability* to be unattainable ($F_{readability} = 1.03$) within the remaining target set and it is subsequently removed. On the next attempt, the feasibility test indicates the remaining constraints (*figurative language*, *politeness*, and *verbosity*) are jointly attainable; they are fulfilled with $\|\boldsymbol{\xi}^*\| = 0.65$ and $\max|\xi_i^*| = 0.43$ following optimization. Here, both refusals are sensible, since the 19 of 249 measured configurations with *complexity* above 6.0 have *readability* between 0.49 and 5.27. Further, the

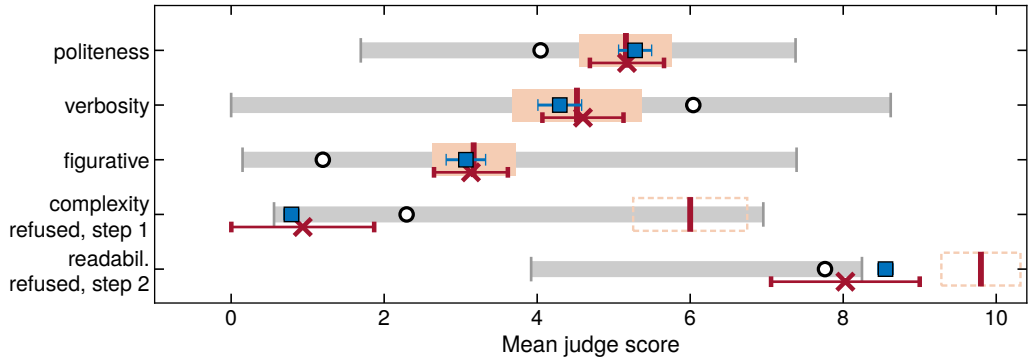


Figure 3: A test case on five steered behaviors of Qwen3-8B. Open circles mark the mean judged scores of the unsteered LLM; gray bars show the range of scores the fitted surface predicts when the behavior's own coefficient alone is varied over $\pm\gamma_i^{\max}$; red ticks mark the targets and the orange bands around them the $\pm 2\sigma_r$ tolerances; red crosses show the polynomial-predicted scores after optimization, in which *complexity* and *readability* are rejected (see text), with error bars spanning one standard deviation of the prediction over the interventions compatible with the request, propagated from $\boldsymbol{\Sigma}$; blue squares with standard errors show the means over 132 held-out prompts evaluated directly from Qwen3-8B.

ten configurations with *readability* above 8.5 have *complexity* no higher than 1.17. This example demonstrates the utility of the feasibility test in identifying contradictory or unachievable requests.

All three surviving targets are reached, both as predicted by the response surfaces (red crosses in Figure 3) and as measured directly on the 132 held-out prompts (blue squares). *Figurative language* is measured at 3.07 ± 0.26 against the 3.17 target, *politeness* at 5.28 ± 0.22 against 5.16, and *verbosity* at 4.30 ± 0.29 against 4.52. The corresponding correlation matrix is shown in Figure 9. This three-target specification is used for the analyses below.

5.4 Constraint structure of the solution

The eigenvectors of $\boldsymbol{\Sigma}$ define the principal axes of steering-coefficient variation, and the square roots of the eigenvalues give the standard deviations along those axes. A specification with three targets can constrain at most three such axes (Equation 8). For the case study of Section 5.3, the three retained targets reduce the standard deviation along three eigenvectors of $\boldsymbol{\Sigma}$ from the prior value of 0.5 to 0.031, 0.066, and 0.104, while the remaining eight stay at 0.5. Figure 4 shows the resulting joint distribution. The specification contracts the prior sphere into an ellipsoid whose size and orientation depend on the targets and their tolerances. The strongest coupling is between the *politeness* and *verbosity* coefficients, with a correlation of -0.53; *verbosity* and *complexity* are less correlated.

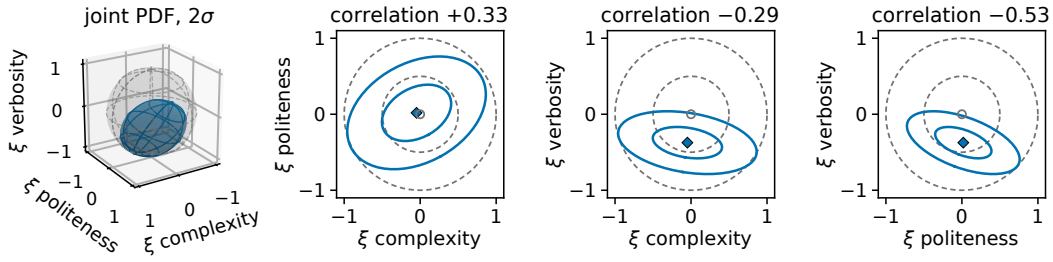


Figure 4: *Left:* the joint 2-standard-deviation surfaces of the prior and posterior probability distributions calculated from $\boldsymbol{\Sigma}$ for the case study. *Remaining panels:* the three pairwise 2D projections of the 3D plot on the *left* at 1 and 2 standard deviations (solid lines). Dashed circles are the prior, and the diamond symbols are the optimal solution $\boldsymbol{\xi}^*$.

5.5 Behavioral trade-offs in unrequested behaviors

Because $\boldsymbol{\Sigma}$ describes a distribution of interventions, many interventions satisfy the specification to within its tolerances, and they differ in the behaviors that were not requested. To second order, moving t standard deviations from $\boldsymbol{\xi}^*$ along any eigenvector of $\boldsymbol{\Sigma}$ raises Φ by t^2; all points on the shell at fixed t achieve the same objective value. We test this along one of the unconstrained directions, whose standard deviation remains at the prior value of 0.5. We measure Qwen3-8B at nine interventions along this direction, spaced 0.5 standard deviations (0.25 in coefficient norm) apart and indexed $\alpha = 1, \ldots, 9$ (Figure 5). To within judging noise, the three targets remain satisfied at all nine interventions. Within one standard deviation of the solution ($\alpha = 3$ to 7), targeted behaviors change very little (center panel). Untargeted behaviors move more. *Helpfulness* and *readability* move the most (right panel), and the response surfaces accurately predict the direction of every change and its size within one standard deviation of the solution (lines against symbols). The covariance therefore separates the directions in coefficient space that the specification pins down from those it leaves free. Along the free directions, a user can move untargeted behaviors without violating the specification.

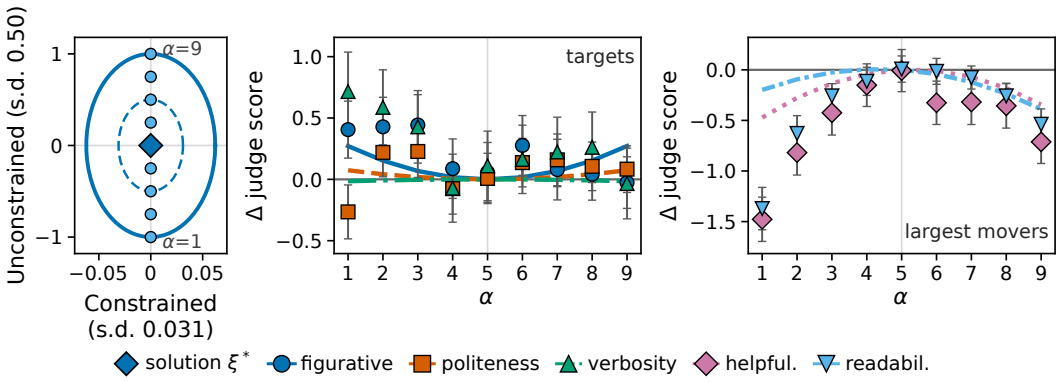


Figure 5: The nine interventions along the unconstrained direction are indexed from $\alpha = 1$ at the lowest point on the ellipse to $\alpha = 9$ at the highest, spaced 0.25 apart in $\|\boldsymbol{\xi} - \boldsymbol{\xi}^*\|_2$. *Left:* the plane of two eigenvectors of $\boldsymbol{\Sigma}$ with the largest and the smallest standard deviation. The dashed and solid contours are 1 and 2 standard deviations from $\boldsymbol{\xi}^*$. Circles are the nine interventions tested along the unconstrained direction. Changes in mean judged scores along the unconstrained direction are assessed on 132 held-out prompts for the target set (*center*) and two untargeted behaviors with the largest change in their mean scores (*right*). Symbols with associated standard errors are measured directly from the model; lines are the response-surface predictions (see text).

5.6 Comparison with other methods

We compare against Pre-Control (Zhang et al., 2026), which was shown to outperform fixed-direction steering (ITI; Li et al., 2023), sparse orthogonal multi-attribute steering (MAT-Steer; Nguyen et al., 2025b), and test-time optimization of the hidden state (RE-Control; Kong et al., 2024) on its target-reaching benchmark. A target is a vector of five integer scores from 0 to 4. On Code-UltraFeedback (Weyssow et al., 2024), these targets cosist of *complexity and efficiency*, *style*, *explanation*, *instruction-following*, and *readability*. A response succeeds if all five of its scores match the target, and ℓ_1 is its distance from the target. Table 1 reports the results. Base and Prompting are the unsteered model without and with an instruction template; grid selection picks, among the measured design configurations, the one whose judged scores are closest to the target, without response surfaces or optimization; our method is run with tolerances $\sigma_r = 0.5$ and 0.25 for every target and, in a third variant, combined with prompting. Despite our best efforts, we could not reproduce the Base and Prompting numbers reported by Zhang et al. (2026), whose rows are therefore listed as published and marked †; our Base ℓ_1 is larger and our Prompting baseline considerably stronger than theirs, which prevents an exact comparison. Within that caveat, our method performs comparably. All three variants exceed Pre-Control's success rate on both Phi-4-mini targets, while Pre-Control attains a smaller ℓ_1 on both Llama-3.2-3B targets. Our method achieves this without

the inference-time optimization that RE-Control and Pre-Control perform, since a single $\boldsymbol{\xi}^*$ per target is applied unchanged to every prompt. HelpSteer2 (Wang et al., 2024b) comparisons are in Appendix G.

Table 1: Code-UltraFeedback benchmark on two LLMs, where ℓ_1 is the distance to the target and SR is the success rate (%). The best score within the top section of the table is shown in bold and the second best is underlined. † marks rows reproduced from Zhang et al. (2026).

	Llama-3.2-3B				Phi-4-mini			
	[3,3,3,3,3]		[2,2,2,2,2]		[3,3,3,3,3]		[2,2,2,2,2]	
Method	$\ell_1 \downarrow$	SR↑	$\ell_1 \downarrow$	SR↑	$\ell_1 \downarrow$	SR↑	$\ell_1 \downarrow$	SR↑
Base	3.18	–	3.43	–	2.83	–	3.92	–
Prompting	<u>2.70</u>	<u>21.5</u>	3.50	8.1	2.20	41.3	3.56	11.9
Grid selection	3.07	9.9	2.66	30.6	1.96	46.6	3.14	28.3
Grid selection + Prompting	2.75	20.9	2.87	27.7	**1.64**	**57.1**	3.06	25.2
This work (± 0.5)	3.00	13.4	<u>2.57</u>	**35.5**	2.35	28.8	**2.45**	**33.9**
This work (± 0.25)	2.80	18.8	**2.55**	<u>31.1</u>	2.11	37.8	<u>2.90</u>	<u>31.8</u>
This work (± 0.25) + Prompting	**2.42**	**30.5**	2.89	26.2	<u>1.74</u>	<u>52.8</u>	3.68	24.7
Base†	2.29	–	2.95	–	1.57	–	3.65	–
Prompting†	2.21	6.8	2.85	6.1	1.47	9.5	3.64	2.2
Pre-Control †	2.08	17.5	1.95	30.7	1.42	26.2	2.80	22.3

6 Discussion and limitations

The framework extends readily to secondary objectives: among interventions that satisfy the primary targets, a user may select the one predicted to be most coherent or helpful (Section 5.5). Because the calibration cost is paid once per model, it is also suited to serving many requests, such as personalizing a model per user, each at the cost of one optimization.

While our results demonstrate the effectiveness of the proposed method, several limitations merit discussion. First, the precision of target matching is limited by the accuracy of the response surfaces tested thus far, and tighter control may require a more accurate surrogate model (Appendix C). Additionally, the intervention penalty expresses a preference for a small normalized steering coefficient vector but does not guarantee minimal LLM capability degradation or consider the different unnormalized magnitudes of each steering coefficient. However, we confirm that, on all four LLMs tested and over the PubMedQA and ARC-Challenge datasets, accuracy decreases with the magnitude of the steering coefficient vector (Appendix H). Another limitation worth noting is that we observe the accuracy of our surrogate model, and hence of the predicted interventions, to decrease near the limits of the steering range. Further, the assumptions concerning the Gaussian prior require further examination. Additional considerations include the use of a mean judged score over a prompt set, whereas responses to individual prompts under a steering intervention can be more diverse.

7 Conclusion

We present a framework for precise, predictive control over the behavior of LLM responses. Our methodology identifies optimized steering interventions that match a specified behavioral request. We demonstrate the ability of the approach to detect and systematically reject infeasible or conflicting requests. We extend this analysis by quantitatively characterizing the trade-offs among alternative steering interventions that satisfy the same behavioral specification. Compared with available methods, our approach exhibits strong performance, especially when combined with prompting. Beyond performance gains, the framework is inexpensive to deploy. It also provides an interpretable view of the structure of the steering problem, revealing relationships, redundancies, and coupling among candidate interventions.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with code development, implementation of our framework, and for verifying the interpretation of results. We did not use generative AI tools for research ideation or experimental design. Additionally, we used generative AI tools for assistance in revising and polishing the manuscript. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Activation steering can alter behaviors such as *refusal*, and large interventions reduce model performance (Appendix H). These risks should be considered before deployment. To evaluate *refusal*, the `benign_unsafe` category contains benign prompts that resemble unsafe requests, drawn from XSTest (Röttger et al., 2024) and OR-Bench (Cui et al., 2025). As a precaution, generated responses to these prompts will not be released.

REPRODUCIBILITY STATEMENT

The appendices contain the details necessary to reproduce our results, including the prompt sources and splits, the construction and calibration of the steering directions, the experimental design and fitting procedure, the optimizer settings, and the judging procedure. The code and all artifacts will be made available to the reviewers and area chairs through an anonymous repository during the discussion period and released publicly upon publication.

REFERENCES

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahmoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Xihui (Eric) Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Olatunji Ruwase, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone. Technical Report MSR-TR-2024-12, Microsoft, August 2024. URL `https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/`.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. November 2024. URL `https://openreview.net/forum?id=pH3XAQME6c&noteId=fLFFFEFoFE`.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022. URL `http://arxiv.org/abs/2204.05862`. arXiv:2204.05862 [cs.CL].

Leonard Bereska and Stratis Gavves. Mechanistic Interpretability for AI Safety - A Review. *Transactions on Machine Learning Research*, April 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=ePUVetPKu6.

Joschka Braun, Dmitrii Krasheninnikov, Usman Anwar, Robert Kirk, Daniel Tan, and David Scott Krueger. A sober look at steering vectors for LLMs, 2024. URL https://www.alignmentforum.org/posts/QQP4nq7TXg89CJGBh/a-sober-look-at-steering-vectors-for-llms. Alignment Forum.

Sviatoslav Chalnev, Matthew Siu, and Arthur Conmy. Improving steering vectors by targeting sparse autoencoder features, 2024. URL https://arxiv.org/abs/2411.02193.

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

Emily Cheng and Carmen Amo Alonso. Liseco: Linear semantic control for language generation, 2026. URL https://arxiv.org/abs/2405.15454.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-bench: An over-refusal benchmark for large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=CdFnEu0JZV.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Van Dai Do, Quan Hung Tran, Svetha Venkatesh, and Hung Le. Dynamic steering with episodic memory for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13731–13749, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.706. URL https://aclanthology.org/2025.findings-acl.706/.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Yuchen Fan, Yuzhong Hong, Qiushi Wang, Junwei Bao, Hongfei Jiang, and Yang Song. Preference-Oriented Supervised Fine-Tuning: Favoring Target Model over Aligned Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(22):23859–23867, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i22.34558. URL https://ojs.aaai.org/index.php/AAAI/article/view/34558.

Alex Ferrando, Xavier Suau, Jordi Gonzàlez, and Pau Rodriguez. Dynamically scaled activation steering, 2026. URL https://openreview.net/forum?id=wMtS3brldm.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. A Primer on the Inner Workings of Transformer-based Language Models, October 2024. URL http://arxiv.org/abs/2405.00208. arXiv:2405.00208 [cs.CL].

Michael Frenklach, Hai Wang, and Martin J. Rabinowitz. Optimization and analysis of large chemical kinetic mechanisms using the solution mapping method—combustion of methane. *Progress in Energy and Combustion Science*, 18(1):47–73, January 1992. ISSN 0360-1285. doi: 10.1016/0360-1285(92)90032-V. URL https://www.sciencedirect.com/science/article/pii/036012859290032V.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Narges Ghasemi, Amir Ziashahabi, Salman Avestimehr, and Jonathan May. Orbit: Training-free multi-attribute behavioral steering via orthogonal subspace rotation, 2026. URL https://arxiv.org/abs/2606.22357.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with Co-Scientist. *Nature*, 655(8122):487–496, July 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10644-y. URL https://www.nature.com/articles/s41586-026-10644-y.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi,

Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ra-

maswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Yixiong Hao, Ayush Panda, Stepan Shabalin, and Sheikh Abdur Raheem Ali. Patterns and mechanisms of contrastive activation engineering. March 2025. URL https://openreview.net/forum?id=FZk9oWvZm2.

Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To Steer or Not to Steer? Mechanistic Error Reduction with Abstention for Language Models. June 2025. URL https://openreview.net/forum?id=fUCPq5RvmH.

Gwo-Jen Hwang and Ching-Yi Chang. A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, 31(7):4099–4112, October 2023. ISSN 1049-4820. doi: 10.1080/10494820.2021.1952615. URL https://doi.org/10.1080/10494820.2021.1952615. _eprint: https://doi.org/10.1080/10494820.2021.1952615.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019. URL https://arxiv.org/abs/1909.06146.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the Effects of RLHF on LLM Generalisation and Diversity. October 2023. URL https://openreview.net/forum?id=PXD3FAVHJT.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style Vectors for Steering Generative Large Language Models. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 782–802, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.52. URL https://aclanthology.org/2024.findings-eacl.52/.

Lingkai Kong, Haorui Wang, Wenhao Mu, Yuanqi Du, Yuchen Zhuang, Yifei Zhou, Yue Song, Rongzhi Zhang, Kai Wang, and Chao Zhang. Aligning large language models with representation editing: A control perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=yTTomSJsSW.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023. URL https://arxiv.org/abs/2304.07327.

Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*, 388(13):1233–1239, March 2023. ISSN 0028-4793. doi: 10.1056/NEJMsr2214184. URL https://www.nejm.org/doi/full/10.1056/NEJMsr2214184.

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, pp. 41451–41530. Curran Associates, Inc., 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Xinxin Liu, Ming Li, Zonglin Lyu, Yuzhang Shang, and Chen Chen. Learning from Noisy Preferences: A Semi-Supervised Learning Approach to Direct Preference Optimization. October 2025. URL https://openreview.net/forum?id=rRc04jyoAk.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-Tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. ISSN 2998-4173. doi: 10.1109/TASLPRO.2025.3606231. URL https://ieeexplore.ieee.org/document/11151751/.

Mistral AI. Un ministral, des ministraux. Mistral AI news post, October 2024. URL https://mistral.ai/news/ministraux/.

Dung Viet Nguyen, Yen Nhi Pham, Hieu M. Vu, Lei Zhang, and Tan Minh Nguyen. Activation Steering with a Feedback Controller. October 2025a. URL https://openreview.net/forum?id=vzkEX2SwFD.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-attribute steering of language models via targeted intervention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20619–20634, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1007. URL https://aclanthology.org/2025.acl-long.1007/.

Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2006. ISBN 9780387400655. URL https://books.google.com/books?id=VbHYoSyelFcC.

Narmeen Oozeer, Luke Marks, Shreyans Jain, Fazl Barez, and Amirali Abdullah. Beyond linear steering: Unified multi-attribute control for language models, 2026. URL https://arxiv.org/abs/2505.24535.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. October 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. November 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Matthew T. Reagan, Habib N. Najm, Roger G. Ghanem, and Omar M. Knio. Uncertainty quantification in reacting-flow simulations through non-intrusive spectral projection. *Combustion and Flame*, 132(3):545–555, 2003. ISSN 0010-2180. doi: https://doi.org/10.1016/S0010-2180(02)00503-5. URL https://www.sciencedirect.com/science/article/pii/S0010218002005035.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via Contrastive Activation Addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Pau Rodriguez, Arno Blaas, Michal Klein, Luca Zappella, Nicholas Apostoloff, Marco Cuturi, and Xavier Suau. Controlling Language and Diffusion Models by Transporting Activations. October 2024. URL https://openreview.net/forum?id=l2zFn6TIQi.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. Multi-property steering of large language models with dynamic activation composition. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 577–603. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.34. URL http://dx.doi.org/10.18653/v1/2024.blackboxnlp-1.34.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

David A. Sheen and Hai Wang. The method of uncertainty quantification and minimization using polynomial chaos expansions. *Combustion and Flame*, 158(12):2358–2374, December 2011. ISSN 0010-2180. doi: 10.1016/j.combustflame.2011.05.010. URL https://www.sciencedirect.com/science/article/pii/S0010218011001568.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local Linearity of LLMs Enables Activation Steering via Model-Based Linear Optimal Control. June 2026.

Ralph C. Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013. doi: 10.1137/1.9781611973228. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973228.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting Latent Steering Vectors from Pretrained Language Models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Magamed Taimeskhanov, Samuel Vaiter, and Damien Garreau. Towards understanding steering strength. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=A52LNgpGKy.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 139179–139212. Curran Associates, Inc., 2024. doi: 10.52202/079017-4417. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fb3ad59a84799bfb8d700e56d19c231b-Paper-Conference.pdf.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models with Activation Engineering. October 2024. URL https://openreview.net/forum?id=2XBPdPIcFK.

Teun van der Weij, Massimo Poesio, and Nandi Schoots. Extending activation steering to broad skills and multiple behaviours, 2024. URL https://arxiv.org/abs/2403.05767.

Arthur Vogels, Benjamin Wong, Yann Choho, Annabelle Blangero, and Milan Bhan. In-Distribution Steering: Balancing Control and Coherence in Language Model Generation, October 2025. URL http://arxiv.org/abs/2510.13285. arXiv:2510.13285 [cs.CL].

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*, 2024a.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=PvVKUFhaNy.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. 2022.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Martin Weyssow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. Codeultrafeedback: An llm-as-a-judge dataset for aligning large language models to coding preferences, 2024. URL https://arxiv.org/abs/2403.09032.

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, 2025. URL https://arxiv.org/abs/2501.17148.

Dongbin Xiu and George Em Karniadakis. Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Computer Methods in Applied Mechanics and Engineering*, 191(43):4927–4948, 2002. ISSN 0045-7825. doi: https://doi.org/10.1016/S0045-7825(02)00421-8. URL https://www.sciencedirect.com/science/article/pii/S0045782502004218.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55204–55224. PMLR, July 2024. URL https://proceedings.mlr.press/v235/xu24t.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Rongzhi Zhang, Liqin Ye, Yuzhao Heng, Xiang Chen, Tong Yu, Lingkai Kong, Sudheer Chava, and Chao Zhang. Precise attribute intensity control in large language models via targeted representation editing, 2026. URL https://arxiv.org/abs/2510.12121.

Yue Zhang, Wendi Dong, Laurien A. Vandewalle, Rui Xu, Gregory P. Smith, and Hai Wang. Neural network approach to response surface development for reaction model optimization and uncertainty minimization. *Combustion and Flame*, 251:112679, 2023. ISSN 0010-2180. doi: https://doi.org/10.1016/j.combustflame.2023.112679. URL https://www.sciencedirect.com/science/article/pii/S0010218023000640.

Hongjue Zhao, Haosen Sun, Jiangtao Kong, Xiaochang Li, Qineng Wang, Liwei Jiang, Qi Zhu, Tarek Abdelzaher, Yejin Choi, Manling Li, and Huajie Shao. Odesteer: A unified ode-based steering framework for llm alignment, 2026. URL https://arxiv.org/abs/2602.17560.

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, March 2025. URL http://arxiv.org/abs/2310.01405. arXiv:2310.01405 [cs.LG].

A Steering-direction construction and calibration

A.1 Generation, filtering, and activation read

Prompts are drawn from the `extract` split of Appendix E, which is separate from the fitting and held-out prompts. We screen response pairs by checking that the positive response scores higher than the negative response using the LLM judge.

Activations are collected with the eliciting system prompt in context and averaged over generated response positions restricted to a maximum of 256 tokens. Writing $\bar{\boldsymbol{h}}^{(\ell)}(c) = |T_c|^{-1} \sum_{t \in T_c} \boldsymbol{h}_t^{(\ell)}(c)$ for the mean residual activation over the generated response positions, the steering vector for behavior r at layer ℓ is the normalized mean difference over pairs,

$$\boldsymbol{v}_r^{(\ell)} = \frac{\sum_{(c^+,c^-)} \left[\bar{\boldsymbol{h}}^{(\ell)}(c^+) - \bar{\boldsymbol{h}}^{(\ell)}(c^-)\right]}{\left\| \sum_{(c^+,c^-)} \left[\bar{\boldsymbol{h}}^{(\ell)}(c^+) - \bar{\boldsymbol{h}}^{(\ell)}(c^-)\right] \right\|_2}. \tag{10}$$

A.2 Layer and range selection

We restrict the layers at which steering can be applied to the middle half of the network (relative depths 0.25–0.75), so that an intervention is placed neither at the very start nor at the very end of the forward pass. To choose among them, we use a subset of 24 fitting prompts per behavior, spread evenly across the prompt groups of Appendix E, from which steering directions for each behavior are extracted. We apply steering to every considered layer by adding and subtracting the unnormalized difference-in-means vector at each layer because activation magnitudes vary by orders of magnitude at different depths. We define the sensitivity of a layer to be the average change in the judged score of the behavior between positively and negatively steered models, restricted to layers whose *coherence* stays within 2.0 points of the unsteered model at both signs. The most sensitive layers are carried into the coefficient sweep, which makes the final choice as described next; Figure 6 shows the sensitivity of every layer and the layer selected for each behavior.

Once the candidate layers are fixed, we approximate $\gamma_i^{\max}$ on a probe set of 40 prompts drawn in the same way. The coefficient on the normalized steering vector is increased in steps in both the positive and negative directions until mean judged *coherence* falls more than 2.0 points below that layer's own unsteered *coherence*. $\gamma_i^{\max}$ is the largest magnitude at which *coherence* holds in both directions. The final layer selected for a behavior is the candidate with the largest change in the behavior between $-\gamma_i^{\max}$ and $+\gamma_i^{\max}$. Table 2 reports the selected layer and $\gamma_i^{\max}$ for each behavior and model. Activation scales differ across models and layers, so $\gamma_i^{\max}$ is not directly comparable across models. Table 2 also presents the span of each behavior, defined as the tenth-to-ninetieth percentile range of the score predicted by the fitted response surface over coherent settings in the calibrated region. These spans set the tolerances and the target ranges used in Appendix C.

B Response-surface details

B.1 Design

The 149 Sobol configurations are split 37, 37, and 75 across the three shells. The 99 sparse configurations comprise 36, 27, 19, and 17 configurations with exactly one, two, three, and four active

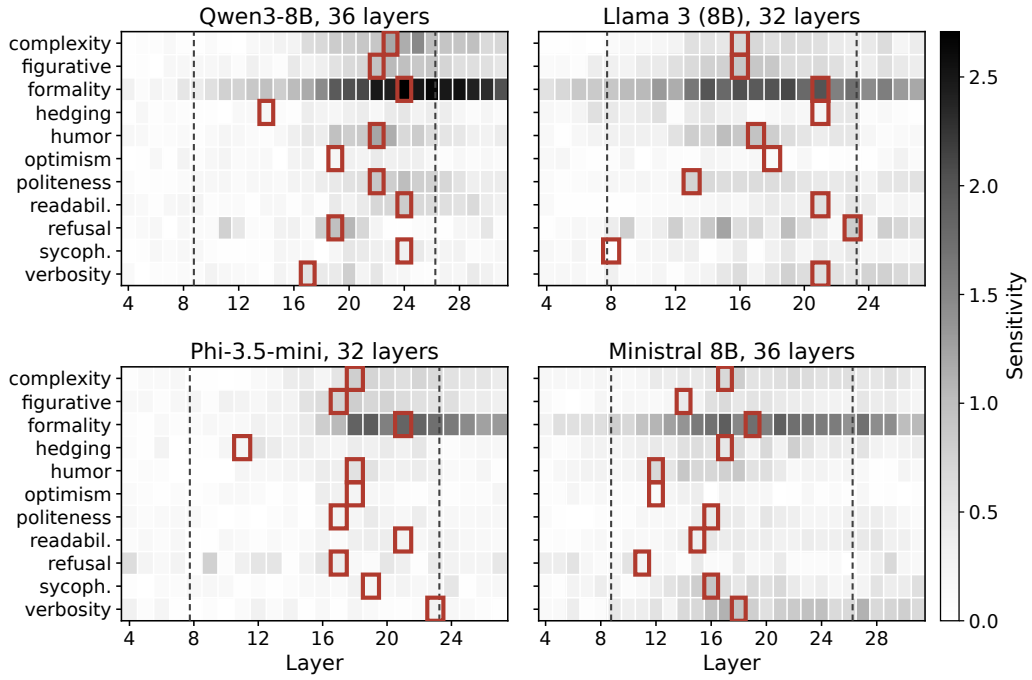


Figure 6: Layer sensitivity of each steered behavior across the four models. Sensitivity is defined as the difference in average judged score when adding versus subtracting its unnormalized difference-in-means steering vector at a given layer. Dashed lines bound the middle half of the network, from which layers are selected, and red boxes indicate the layer chosen for each behavior.

Table 2: Selected layer, steering range, and span for each behavior and model. Layer is the injection layer; $\gamma^{\max}$ is the largest coefficient magnitude at which mean judged *coherence* stays within 2.0 points of the unsteered model in both directions. Span is the tenth-to-ninetieth percentile range of the fitted response surface.

Behavior	qwen3-8b			llama3-8b			phi3.5-mini			ministral-8b		
	layer	$\gamma^{\max}$	span	layer	$\gamma^{\max}$	span	layer	$\gamma^{\max}$	span	layer	$\gamma^{\max}$	span
complexity	23	66.2	5.96	16	3.7	4.30	18	33.9	6.12	17	3.6	4.36
figurative lang.	22	73.8	4.36	16	4.6	3.57	17	27.5	4.37	14	3.2	2.35
formality	24	86.6	6.83	21	6.0	6.10	21	49.1	6.01	19	3.3	5.74
hedging	14	33.6	2.73	21	9.4	2.30	11	19.0	2.36	17	6.5	3.93
humor	22	43.2	1.66	17	4.4	1.98	18	30.2	1.22	12	2.4	1.21
optimism	19	38.3	3.00	18	4.8	2.28	18	43.8	3.04	12	2.4	2.62
politeness	22	74.4	4.85	13	3.9	4.06	17	38.4	3.70	16	4.1	3.31
readability	24	90.1	4.14	21	7.7	2.91	21	51.1	4.19	15	3.3	3.25
refusal	19	43.7	1.09	23	9.0	0.58	17	41.6	0.79	11	3.1	0.73
sycophancy	24	123.3	1.50	8	3.8	1.04	19	34.9	1.37	16	3.8	1.38
verbosity	17	57.8	6.79	21	4.8	5.46	23	99.6	3.91	18	5.1	7.25

directions, respectively, in which each active coordinate is set to ± 0.5 or ± 1. The 48 held-out configurations contain 7, 7, and 15 from the shells at 0.1, 0.3, and 0.5, respectively, and 6, 7, 4, and 2 from the sparse sets with one, two, three, and four active directions.

B.2 PRIOR MOMENTS

Under the prior $\xi_i \sim \mathcal{N}(0, 1/4)$, the rescaled variable $z_i = 2\xi_i$ is standard normal. The Hermite polynomials therefore satisfy $\mathbb{E}[He_1] = \mathbb{E}[He_2] = 0$, $\mathbb{E}[He_1^2] = 1$, and $\mathbb{E}[He_2^2] = 2$. Due to orthogonality, the variance of the predicted mean score $\widehat{\mu}_r(\boldsymbol{\xi})$ under the Gaussian coefficient prior is a weighted sum of squared expansion coefficients,

$$\mathrm{Var}[\widehat{\mu}_r] = \sum_{i=1}^{K} a_{r,i}^2 + 2\sum_{i=1}^{K} a_{r,ii}^2 + \sum_{i<j} a_{r,ij}^2. \tag{11}$$

This is the spread of the predicted mean score of behavior r over interventions drawn from the prior. The intercept $a_{r,0}$ is the prior mean of $\widehat{\mu}_r$, and the predicted score of the unsteered model is

$$\widehat{\mu}_r(\mathbf{0}) = a_{r,0} - \sum_i a_{r,ii}.$$

B.3 REGULARIZATION

During fitting, we apply regularization to the second-order terms. The matrix $\mathbf{P}_2$ is defined such that it has zero entries for the intercept and linear terms and ones for the quadratic and interaction terms. The scale s_r is the mean diagonal entry of $\mathbf{X}^T\mathbf{W}_r\mathbf{X}$ over these second-order columns, so that λ_r is dimensionless.

We treat λ_r as a hyperparameter selected by five-fold cross-validation over steering configurations. Fitting configurations that pass the *coherence* screen are randomly partitioned into five groups. Then, for each λ in the grid $\{0, 10^{-3}, 3\cdot10^{-3}, 10^{-2}, 3\cdot10^{-2}, 0.1, 0.3, 1\}$, the surface is fitted on four groups and evaluated on the fifth by the mean absolute error between predicted and measured mean judged scores. The λ_r that minimizes this error is selected; it is nonzero for all thirteen behaviors on all four models. Regularization lowers the cross-validation error by about a third relative to the unregularized surface (Table 3). With limited and noisy measurements, regularization prevents the curvature and interaction terms from overfitting.

Table 3: Five-fold cross-validation error of the degree-two response surfaces, in judge points, averaged over the thirteen behaviors, without regularization ($\lambda_r = 0$) and with λ_r selected per behavior.

	unregularized	regularized	reduction
Qwen3-8B	0.54	0.37	33%
Llama 3 (8B)	0.55	0.35	37%
Phi-3.5-mini	0.56	0.37	34%
Ministral 8B	0.51	0.33	34%

B.4 COHERENCE SCREENING AND PREDICTION CLIPPING

Configurations with mean judged *coherence* below 5.0 are excluded from all response-surface fits except *coherence* itself. Of the 201 fitting configurations, 177, 187, 182, and 182 satisfy this requirement for Qwen3-8B, Llama 3 (8B), Phi-3.5-mini, and Ministral 8B, respectively. *Coherence* is fit on all 201 configurations.

It is possible for predictions from a polynomial response surface to fall outside the judge's $[0, 10]$ scale. Predictions are therefore clipped to this range only when compared with judged observations. Across the four LLMs, clipping affects 1 to 4% of the held-out predictions.

B.5 TWO FITTED SURFACES AGAINST AN INDEPENDENT GRID

Figure 7 shows the fitted expansion for two behaviors of Qwen3-8B. A design configuration is displayed on the plot only if each of its other nine coefficients has magnitude below 0.10. This leaves 8 and 11 of the 249 configurations in the left and right surfaces of Figure 7. To better visualize and verify the fit performance, we supplement with additional measurements in a 5×5 grid across

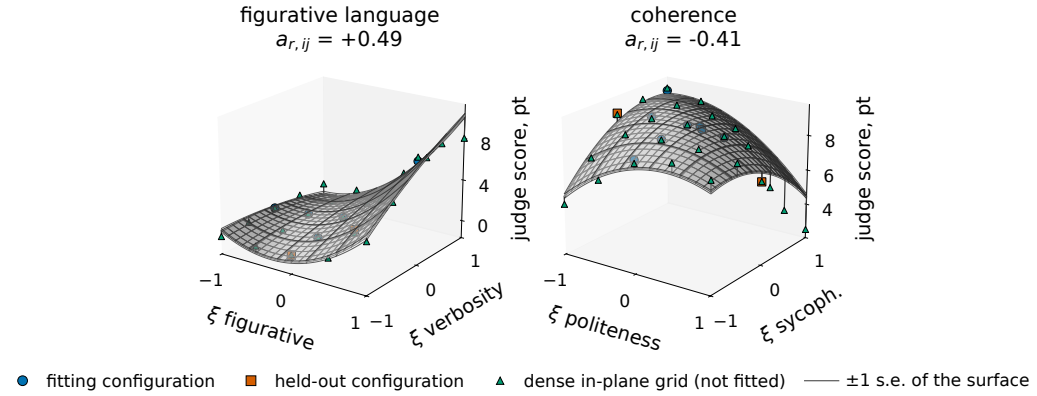


Figure 7: Fitted response surfaces of Qwen3-8B in two planes. *Left:* predicted mean judge score for *figurative language* over the *figurative-language* and *verbosity* coefficients. *Right:* predicted *coherence* over the *politeness* and *sycophancy* coefficients. The envelope around the response surfaces is ± 1 standard error of the fit. Circles and squares are fitting and held-out design configurations, triangles are the 25 grid measurements, and stems join each measurement to the surface.

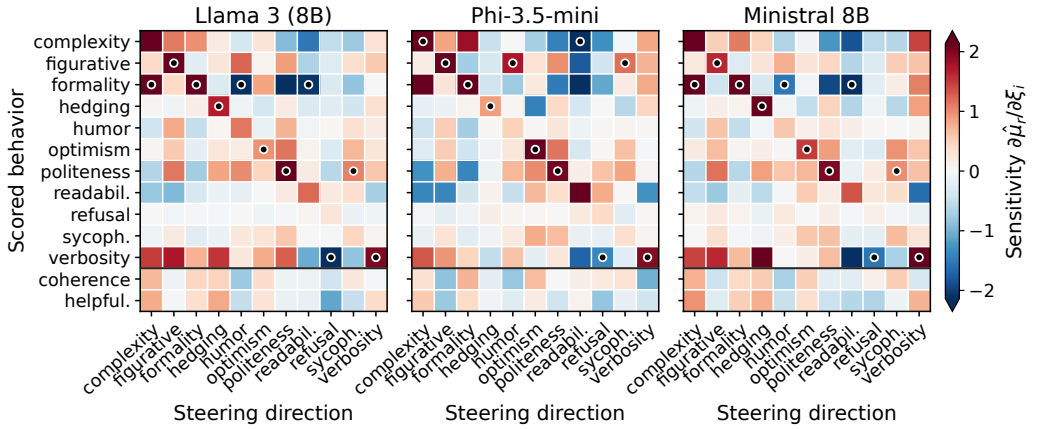


Figure 8: Sensitivity $\partial\widehat{\mu}_r/\partial\xi_i$ of each scored behavior to each steering direction at the unsteered model. This is measured in judge points per unit coefficient. The rows correspond to the thirteen scored behaviors and the columns to the eleven steering directions. The dot marks the largest-magnitude entry in each column.

this 2D plane. The surfaces follow the data with mean absolute errors of 0.42 and 0.26 points on the design configurations they were fitted to and 0.49 points on the independent grid in both planes.

Figure 8 shows the sensitivity matrices of Llama 3 (8B), Phi-3.5-mini, and Ministral 8B, computed from their fitted surfaces as in Figure 2.

C Inverse-solve details

C.1 Target tolerances

For behavior r, the tolerance is $\sigma_r = \max(0.125, w_r/16)$, where w_r is the span of Table 2. The acceptable band $\tau_r \pm 2\sigma_r$ is then one quarter of the range over which the behavior can be shifted while remaining coherent. Thus, an unsteered model satisfies roughly a quarter of uniformly drawn targets. The floor of 0.125 keeps the band wider than measurement noise: the mean judge score at a steering configuration carries a standard error of about 0.2 points. Behaviors whose span is small

relative to this band, defined as the unsteered model satisfying more than 35% of uniformly drawn targets, are not targeted in sampled specifications, although they remain scored and predicted. This excludes *refusal* on Qwen3-8B, *refusal* and *sycophancy* on Llama 3 (8B), and *humor*, *refusal*, and *sycophancy* on Phi-3.5-mini and Ministral 8B.

C.2 Optimization

Optimization is performed using BFGS (Nocedal & Wright, 2006), as implemented in SciPy's `minimize` with its default tolerances and supplied with the analytic gradient of Equation 5. Because the objective is a quartic in $\boldsymbol{\xi}$ and in general nonconvex, each solve uses sixteen starting points, namely the unsteered model $\boldsymbol{\xi} = \mathbf{0}$ and fifteen samples from the coefficient prior $\mathcal{N}(\mathbf{0}, \mathbf{I}/4)$. The minimizer with the lowest objective value is retained. The optimization is unconstrained, so a solution may fall outside $[-1, 1]^K$. Such solutions are extrapolations beyond the calibrated region.

D Covariance and propagation

D.1 Covariance

The objective in Equation 5 defines a density over the coefficients,

$$p(\boldsymbol{\xi}) \propto \exp\left[-\frac{1}{2}\Phi(\boldsymbol{\xi})\right].$$

We approximate this density by a Gaussian centered at the solution $\boldsymbol{\xi}^*$ and whose precision, the inverse of its covariance, is the Hessian of $\Phi/2$ at $\boldsymbol{\xi}^*$. The Hessian $\mathbf{H}_r$ is constant for a degree-two surface, and the Hessian of $\Phi/2$ at $\boldsymbol{\xi}^*$ is

$$4\mathbf{I} + \sum_{r \in \mathcal{R}_{\text{spec}}} \left[\frac{\mathbf{g}_r \mathbf{g}_r^T}{\sigma_r^2} + \frac{\widehat{\mu}_r(\boldsymbol{\xi}^*) - \tau_r}{\sigma_r^2}\mathbf{H}_r\right].$$

We omit the second term, giving the covariance as

$$\boldsymbol{\Sigma}^{-1} = 4\mathbf{I} + \sum_{r \in \mathcal{R}_{\text{spec}}} \frac{\mathbf{g}_r \mathbf{g}_r^T}{\sigma_r^2}. \tag{12}$$

When no targets are specified, the sum is empty and $\boldsymbol{\Sigma} = \mathbf{I}/4$ as expected. The omitted term is proportional to the residual $\widehat{\mu}_r(\boldsymbol{\xi}^*) - \tau_r$ of each retained target, which the feasibility test bounds by $2\sigma_r$ and which is exactly zero when every target is met exactly.

D.2 Constraint structure

Figure 9 shows the correlation matrix induced by the request of Section 5.3. The strongest correlation is between the *politeness* and *verbosity* coefficients at -0.53.

D.3 Propagation through the quadratic response surface

Let

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\xi}^*, \boldsymbol{\Sigma})$$

and write $\boldsymbol{\xi} = \boldsymbol{\xi}^* + \boldsymbol{\delta}$ with $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Because $\widehat{\mu}_r$ is quadratic,

$$\widehat{\mu}_r(\boldsymbol{\xi}) = \widehat{\mu}_r(\boldsymbol{\xi}^*) + \mathbf{g}_r^T \boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}^T \mathbf{H}_r \boldsymbol{\delta} \tag{13}$$

Taking moments yields

$$\mathbb{E}[\widehat{\mu}_r] = \widehat{\mu}_r(\boldsymbol{\xi}^*) + \frac{1}{2}\operatorname{tr}(\mathbf{H}_r \boldsymbol{\Sigma}), \tag{14}$$

and

$$\operatorname{Var}[\widehat{\mu}_r] = \mathbf{g}_r^T \boldsymbol{\Sigma} \mathbf{g}_r + \frac{1}{2}\operatorname{tr}\big((\mathbf{H}_r \boldsymbol{\Sigma})^2\big), \tag{15}$$

whose square root gives Equation 9. The expectation describes the mean score obtained when coefficient settings are drawn from the local Gaussian approximation.

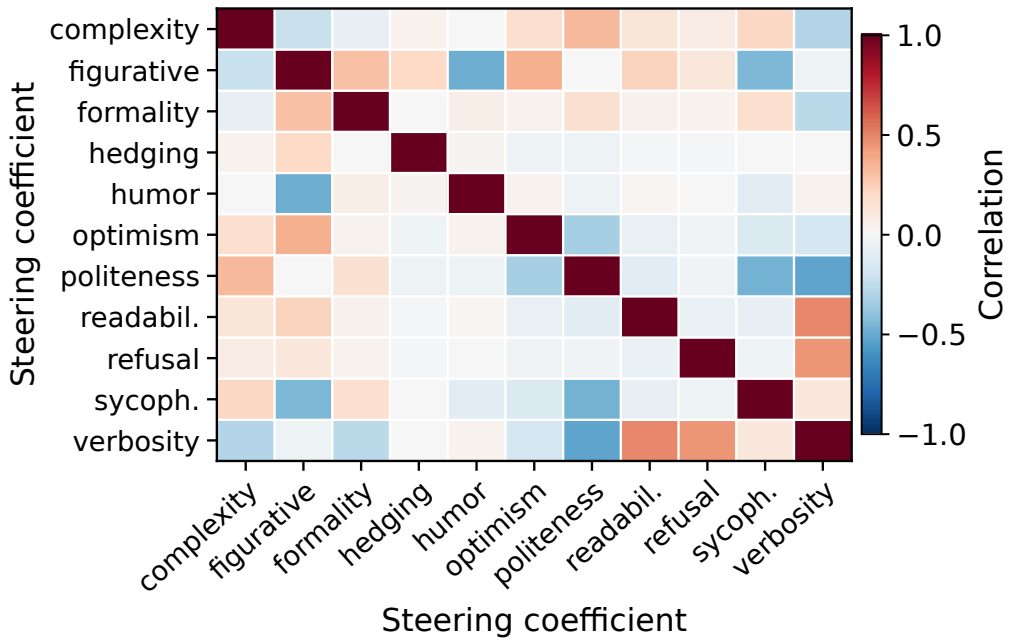


Figure 9: Correlations among steering coefficients induced by the request of Section 5.3 for Qwen3-8B.

E Evaluation prompts

E.1 Sources and data splits

The prompt set contains 775 prompts sampled from eight public datasets, each released under open licenses (CC-BY, CC-BY-SA, MIT, or Apache 2.0) that permit this use. Prompts are grouped into six categories according to the behaviors they are intended to elicit. `general` covers *formality*, *verbosity*, *complexity*, *readability*, *helpfulness*, and *coherence*; `feedback` covers *sycophancy*; `benign_unsafe` covers *refusal*; `false_premise` covers *hedging*; `personal_situation` covers *politeness* and *optimism*; and `creative` covers *humor*. We exclude prompts in the form of code requests.

The three prompt splits are disjoint. The extraction split is used to construct steering directions. The fitting split supplies the probe prompts for selecting steering layers and $\gamma_i^{\max}$, and a fixed subset of 100 prompts, drawn once with the category mix preserved, is used at every design configuration to fit the response surfaces. The held-out split is used only for validation and for every measurement at a solved intervention. These categories ensure that every behavior has prompts on which it can be expressed. Each behavior's steering direction is extracted, and its layer and $\gamma_i^{\max}$ selected, exclusively using prompts from its own categories.

F Judge and behavioral rubrics

F.1 Judge scoring procedure

Each response is scored one behavior at a time by Qwen2.5-14B-Instruct. A judge call includes the original prompt, the generated response, a shared instruction, and a prompt explaining how the behavior should be scored. The judge returns an integer score from 0 to 10. We choose to score behaviors separately rather than jointly to minimize their influence on each other. The shared instruction asks the judge to evaluate only the named behavior and to use the full scale. Each rubric defines the behavior and provides anchors for the low, middle, and high ends of the scale. The judge uses greedy decoding, so re-scoring the same response produces the same score and judge decoding

Table 4: Evaluation prompts by category including the behaviors each category is designed to elicit, the dataset source, and the number of prompts included in the extraction, fitting, and held-out splits.

Category	Behaviors elicited	Source	Extraction	Fit	Held-out
`general`	formality, verbosity, complexity, readability, helpfulness, coherence	Dolly 15k (Conover et al., 2023)	40	68	22
`feedback`	sycophancy	Sycophancy-eval, feedback split (Sharma et al., 2024)	40	120	40
`benign_unsafe`	refusal	XSTest (Röttger et al., 2024)	20	55	15
		OR-Bench hard-1k (Cui et al., 2025)	20	55	15
`false_premise`	hedging	TruthfulQA, generation split (Lin et al., 2022)	20	22	8
		SimpleQA (Wei et al., 2024)	20	23	7
`creative`	humor	WritingPrompts (Fan et al., 2018)	20	18	5
		Dolly 15k, creative writing (Conover et al., 2023)	20	12	5
`personal_situation`	politeness, optimism	OpenAssistant (Köpf et al., 2023)	25	45	15
Total			**225**	**418**	**132**

contributes no sampling variability to the reported behavioral scores. We further reduce the influence of judging bias and variability by averaging over 100 prompts per configuration during fitting.

G Additional Baseline Comparisons

Tables 1 and 5 follow the protocol of Zhang et al. (2026) as closely as possible. Prompts consist of 519 even-indexed prompts from the HelpSteer2 validation split and the 1,000-prompt test split of Code-UltraFeedback. ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024a) scores all responses. Prompts natively satisfied by the unsteered model are removed. Decoding is greedy. For comparison on this benchmark, our pipeline is modified to use $K = 5$, one direction per scored attribute. The procedure for extracting directions and selecting layers remains unchanged. Here, $\gamma_i^{\max}$ is the largest coefficient at which the *coherence* score from ArmoRM stays within 0.8 of the unsteered judge score. We use 80 fitting and 20 held-out configurations on HelpSteer2 and 60 and 15, respectively, on Code-UltraFeedback. Response surfaces are fit using 100 and 40 fitting prompts for HelpSteer2 and Code-UltraFeedback, respectively. On held-out configurations and prompts, the surfaces have mean absolute errors of 0.05 to 0.09 on the 0–4 scale.

Table 5 presents the HelpSteer2 comparison. HelpSteer2 targeted behaviors are *helpfulness*, *correctness*, *coherence*, *complexity*, and *verbosity*. As in Table 1, we notice a discrepancy between our Base and Prompting values and those reported by Zhang et al. (2026). We suspect the baseline differences are attributable to differences in the evaluation setup, possibly relating to hardware differences, precision, library versions, variation in decoding settings, or prompting templates.

In contrast to the Code-UltraFeedback comparison in Table 1, the success rates of most methods remain within a few percentage points of each other. The data here support the conclusions of the main text. Our ± 0.25 solve is on par with Pre-Control's published numbers. As an example, our method using ± 0.25 shows higher success rates than Pre-Control, but with slightly larger ℓ_1 distances. Notably, grid selection has the highest success rate on every cell here, alone or combined with prompting.

Table 5: HelpSteer2 benchmark comparison from Zhang et al. (2026). Metrics ℓ_1 and success rate (SR, %) are shown per model and target. The best score within the top section of the table is shown in bold and the second best is underlined. † marks rows reproduced from Zhang et al. (2026).

	Llama-3.2-3B				Phi-4-mini			
	[4,4,4,2,2]		[3,3,3,2,2]		[4,4,4,2,2]		[3,3,3,2,2]	
Method	$\ell_1 \downarrow$	SR↑	$\ell_1 \downarrow$	SR↑	$\ell_1 \downarrow$	SR↑	$\ell_1 \downarrow$	SR↑
Base	2.82	–	3.07	–	3.12	–	3.08	–
Prompting	2.76	5.0	2.99	3.4	2.80	<u>10.0</u>	2.85	5.2
Grid selection	2.67	<u>11.4</u>	**2.54**	**13.0**	3.05	7.5	2.83	**13.6**
Grid selection + Prompting	**2.57**	**12.6**	2.76	<u>10.3</u>	**2.70**	**10.9**	2.83	<u>12.3</u>
Ours (± 0.5)	2.66	7.8	2.85	6.5	3.03	6.4	2.87	6.7
Ours (± 0.25)	<u>2.63</u>	10.3	2.69	7.5	3.01	8.2	<u>2.82</u>	7.9
Ours (± 0.25) + Prompting	<u>2.63</u>	9.2	<u>2.64</u>	9.9	<u>2.71</u>	8.4	**2.56**	11.7
Base†	2.19	–	2.76	–	2.46	–	2.76	–
Prompting†	2.17	5.4	2.73	2.5	2.42	5.2	2.67	5.2
Pre-Control†	2.16	8.0	2.63	6.6	2.41	8.3	2.46	9.1

H Additional Results

H.1 Steering intervention magnitude effect on LLM capability

We empirically verify that increasing the magnitude of the steering coefficient vector degrades model capability. Across the four models, we evaluate each steered model on PubMedQA and ARC-Challenge. Specifically, we ask the model to reason before answering, so that steering is active while the model thinks about its answer. Figure 10 shows the change in accuracy, in percentage points, as a function of the magnitude of the steering coefficient vector. We exclude Llama 3 (8B) from the ARC-Challenge panel because on about 40% of the questions it ends its response after reasoning and without stating an answer.

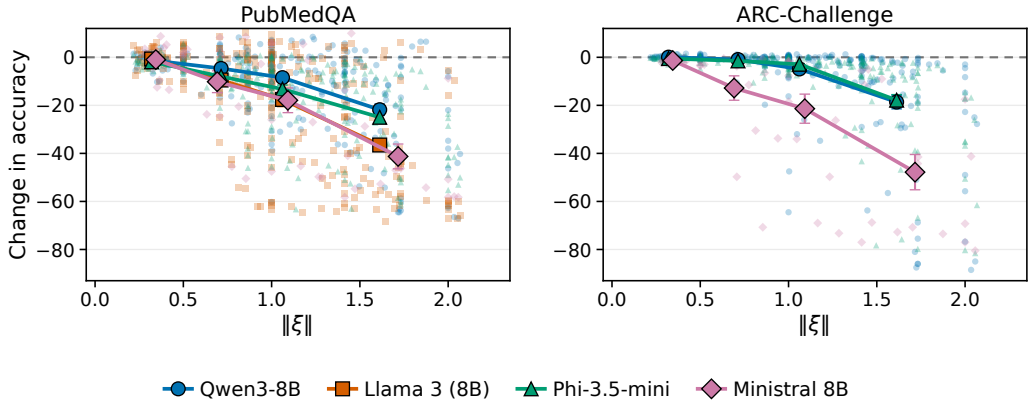


Figure 10: Change in benchmark accuracy from the unsteered model, in percentage points, versus the magnitude of the steering coefficient vector.The faint points are individual configurations, and heavy markers are bin means with associated standard errors.

H.2 Effect of the number of targets specified

We solve 120 random specifications at each number of targets, up to the size of each model's targetable set. Figure 11 shows that the fraction of requests served in full decreases as more targets are

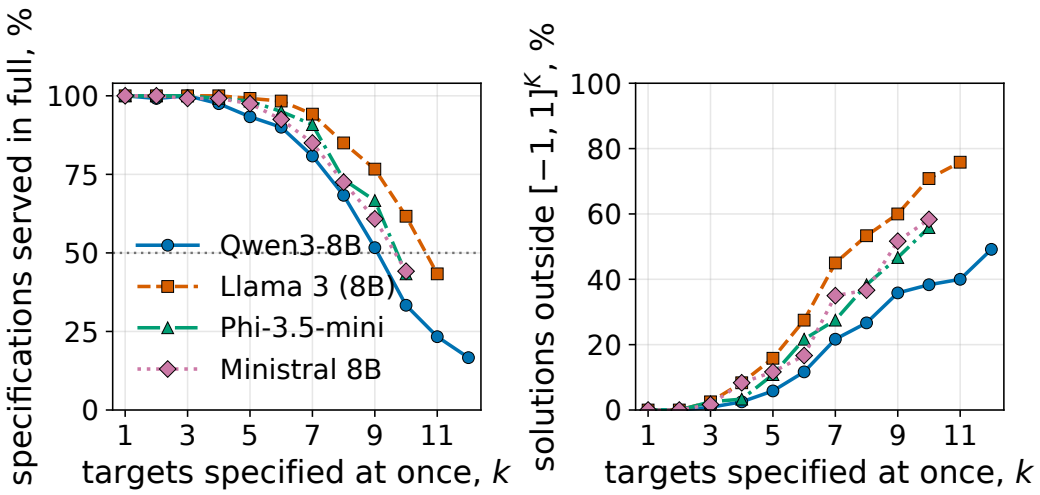


Figure 11: *Left:* the fraction of specifications involving no rejected targets plotted against the number of specified targets. *Right:* the fraction of optimal solutions found to be outside of the calibrated region $[-1, 1]^K$ as a function of the number of specified targets.

jointly specified, and that the fraction of solutions that leave the calibrated region, and are therefore extrapolations, rises monotonically.