CONTRASTIVE COVARIANCE SUBSPACES FOR ACTIVATION STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering controls large language model behavior by adding a direction to hidden activations during inference. For steering to work well, that direction must be a pure signal for the target concept: any confound entangled with it, such as an incidental topic or writing style, is injected into the residual stream alongside the concept itself and degrades the generated text. We introduce Contrastive Covariance Steering (CCS), a training-free method that eliminates this confound variation. CCS projects steering vectors to a low-dimensional subspace obtained in closed form by solving a contrastive variance objective, which retains the directions along which the target concept varies while suppressing directions shared with the other classes. Across three LLMs and four steering domains, CCS consistently beats all baseline steering methods in all 12 model–domain settings. The largest relative improvement occurs on AxBench and Emotion where we outperform the next strongest steering baseline by 48 to 70% and 13 to 27%, respectively. Furthermore, CCS outperforms naive difference-of-means steering by up to 151%, showcasing the importance of deconfounding concepts. A toy experiment verifies our motivation on why: planting a topic nuisance into the training data reduces unprojected steering by 46%, whereas CCS loses only 12% and recovers most of the gap to an oracle that explicitly deletes the planted direction. A human study with 23 raters on 296 generations reproduces the LLM judge's ranking. Our extensive ablation verifies the importance of our method's components and robustness to the number of nuisance classes. Finally, we extend our method to multi-turn conversations, where CCS sustains both concept expression and fluency, matching system prompt performance as other steering methods drift.[1]

1 INTRODUCTION

Additive activation steering offers a lightweight way to control large language model (LLM) behavior without updating model weights. By adding a concept-associated vector to intermediate activations during inference, it can influence properties such as personality, emotional tone, and truthfulness (Li et al., 2023; Zou et al., 2023; Rimsky et al., 2024). Activation steering has also been explored as a safety tool (Lee et al., 2025; Shen et al., 2025). Although activation interventions can bypass safety-trained refusal behavior (Arditi et al., 2024), they can also support selective refusal and improve resistance to jailbreak attempts (Lee et al., 2025; Shen et al., 2025). It is effective at controlling model behavior, easy to implement and comes at no additional computational cost at inference time.

Effective control, however, requires more than increasing concept expression. Strong interventions can move activations away from their natural distribution and degrade generation quality (Luo et al., 2026; Soo et al., 2025; Stolfo et al., 2025). This creates a fundamental trade-off of concept expression (*how much* we enhance a given concept/behavior) and output quality. A useful steering direction should therefore strengthen the intended behavior while preserving fluent, relevant responses. One fundamental challenge is separating the target concept from incidental variation in its demonstrations. Existing methods that estimate directions using contrastive pairing (Turner et al., 2023; Rimsky et al., 2024) can cancel shared features; systematic differences in topic, style, formatting, or response length may remain entangled with the target signal. Approaches that incorporate

[1] Our code will be made publicly available upon acceptance.

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

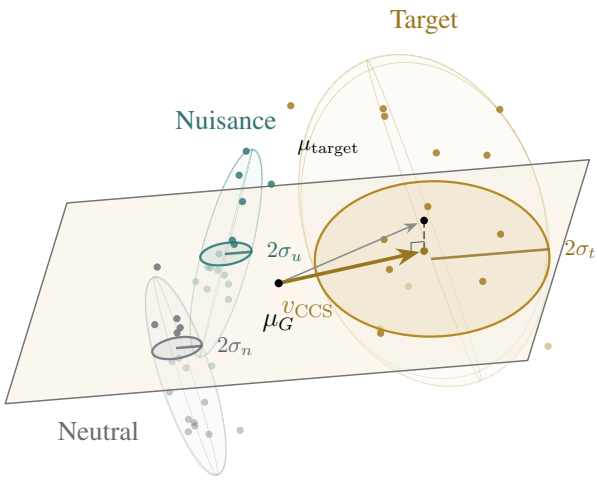


Figure 1: **CCS** projects the target–grand-mean direction onto a contrastive subspace, where the variance of the target class is maximised, while concurrently minimising the variances of other, neutral and nuisance, classes.

second-order information can capture dominant directions of variation, but do not explicitly distinguish target variation from nuisance variation (Zou et al., 2023; Qiu et al., 2024). Non-additive methods address related issues through different mechanisms like angular rotation in a 2D subspace, or optimal transport (Vu & Nguyen, 2025; Rodriguez et al., 2025).

In this paper, we introduce **Contrastive Covariance Steering (CCS)**, which **explicitly eliminates this nuisance variability** in steering directions. We optimize a contrastive subspace objective that rewards variation in the target concept while penalizing variation shared with every other available concept class, admitting a closed-form solution via eigendecomposition. CCS therefore requires no gradient-based optimization, no auxiliary trained models, and a single additive direction at one layer. Projecting a steering direction onto the resulting subspace isolates the concept's own discriminative signal from confounds such as topic, style or dataset-specific artifacts such as formatting differences or length differences. Without this projection, a naive difference-of-means direction could carry such confounds into the generated text. Figure 1 shows our method's subspace construction in 3 dimensions, as well as the resulting steering vector after projection.

Our main contributions can be summarised as follows:

- We introduce CCS, a closed-form method that projects activations onto a contrastive subspace to isolate target concept signals from nuisance variance.
- Our method outperformed established baselines in all model-domain settings across 3 LLMs (Llama-3.1-8B Grattafiori et al. (2024), Qwen3-8B Yang et al. (2025), Gemma-2-9B Gemma Team et al. (2024)) and 4 benchmarks. The results are verified using both LLM judge rankings and a blind multi-rater human study.
- Ablations show that our two main mechanisms, namely grand mean centering and subspace projection, each independently improve standard difference-of-means steering.
- In a multi-turn conversation steering experiment, CCS sustains an 80% joint concept-fluency score over 20 turns and keeps near-maximum concept expression, eventually matching system prompting.

2 Related Work

Activation steering. Inference-time activation steering controls LLM behavior by modifying internal representations without updating model weights (Subramani et al., 2022; Turner et al., 2023; Zou et al., 2023; Li et al., 2023). A common approach derives a single direction from contrasting activations, as in Contrastive Activation Addition (CAA; Rimsky et al., 2024), with related work exploring mean-centering (Jorgensen et al., 2023), persona-specific directions (Chen et al., 2025), affine transformations (Singh et al., 2024; Rodriguez et al., 2025), rotations (Vu & Nguyen, 2025),

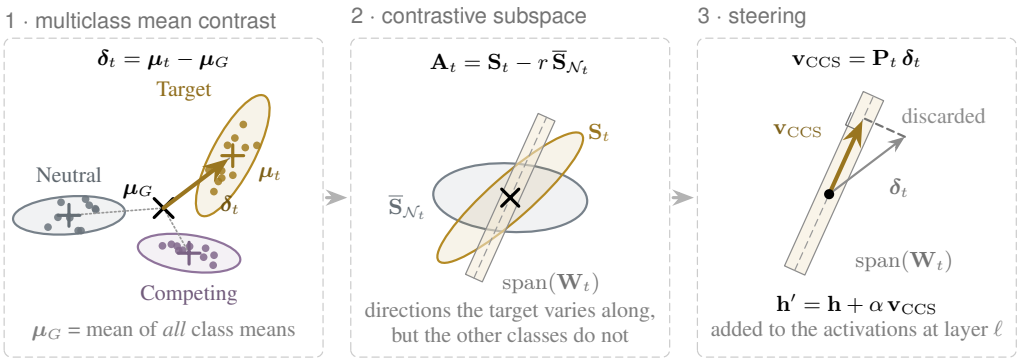


Figure 2: **Contrastive Covariance Steering. (1)** The steering direction is measured from the grand mean of *all* class means, not from a single neutral reference. **(2)** $\mathbf{A}_t = \mathbf{S}_t - r\,\overline{\mathbf{S}}_{\mathcal{N}_t}$ rewards target scatter and penalises the mean non-target scatter $\overline{\mathbf{S}}_{\mathcal{N}_t} = |\mathcal{N}_t|^{-1} \sum_{c \in \mathcal{N}_t} \mathbf{S}_c$ (Eq. 5); its top-k eigenvectors span directions along which the target varies but the other classes do not. **(3)** Projecting $\boldsymbol{\delta}_t$ onto that subspace discards the shared component.

and learned interventions (Wu et al., 2024; 2025b; Bao et al., 2026). CCS remains in the closed-form additive setting, but constructs steering directions from mean-centered activations, using second-order information from multiple competing concepts.

Spectral approaches and low-rank subspaces. CCS is closely related to spectral methods that identify low-dimensional structure in representation spaces. PCA and contrastive PCA identify directions of high variance, with cPCA emphasizing variation present in a target dataset but not a background dataset (Pearson, 1901; Abid et al., 2018); Oldfield et al. (2023) extend this principle to multiple competing factors. In LLMs, SEA uses spectral transformations derived from positive and negative activation structure (Qiu et al., 2024), conceptor-based steering models concepts as multi-dimensional subspaces using second-order statistics (Triantafyllopoulos et al., 2026), and Bhandari et al. (2026a) project personality directions through a shared low-rank subspace. CCS instead learns a *target-specific contrastive* subspace that rewards target scatter and penalizes non-target scatter, then uses it to filter the target–grand-mean contrast into a single additive steering vector. This is also related to Fisher Discriminant Analysis (Fisher, 1936), although FDA learns a common discriminative subspace, whereas CCS learns a separate subspace for each target. Recent work further shows that behavioral representations occupy overlapping low-rank subspaces and orthogonalization alone does not guarantee behavioral independence (Bhandari et al., 2026b; Sharma et al., 2026).

3 Method

In this section, we introduce Contrastive Covariance Steering (CCS), a closed-form method combining multiclass mean contrasts with contrastive subspace projection to construct additive steering directions. Real activation spaces are more entangled than a simple contrast between two class means can capture: pushing toward a target concept often drags the model along nuisance directions shared with other classes in the dataset. A simple difference of means can encode irrelevant information such as variability in topic, writing style or length, especially if the dataset has not controlled for these confounds by explicitly balancing on confounds (even then, implicit imbalances can still remain). CCS addresses this by first measuring the target's displacement from the mean of all class means rather than a single neutral reference class, stabilizing the first-order displacement. It then projects that displacement onto a subspace that preserves the target's own variation while suppressing variation shared with the other classes. Our method is outlined end-to-end in Figure 2.

3.1 Background

We first formalize the mathematical notation for residual-stream activations and standard additive steering interventions.

Let $\mathbf{h}_{\ell,s}(\mathbf{x}) \in \mathbb{R}^d$ denote the residual-stream activation of a language model with prompt x, at layer ℓ and token position s. At a given layer, we use the mean-pooled prompt activations, $\mathbf{h}(\mathbf{x}) = T^{-1}\sum_{s=1}^{T}\mathbf{h}_{\ell,s}(\mathbf{x})$.

An *additive activation steering* method finds a vector $\mathbf{v}$ in the model's activation space, then adds these vectors to the residual stream activations of each token during generation, after unit-normalizing and multiplying by a *steering strength* factor to control intensity:

$$\mathbf{h}_{\ell,s}^{\text{steer}} = \mathbf{h}_{\ell,s} + \alpha\widehat{\mathbf{v}}. \tag{1}$$

where $\widehat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_2$. We don't use a constant α when steering, but a factor of each steered token's norm (see Appendix A.3).

Given concept classes $c \in \mathcal{C}$, we collect the activations of the prompts belonging to each class, $\mathcal{H}_c = \{\mathbf{h}_{c,i}\}_{i=1}^{n_c}$, with class mean $\boldsymbol{\mu}_c = n_c^{-1}\sum_i \mathbf{h}_{c,i}$, and center every activation on the mean of the $|\mathcal{C}|$ class means:

$$\boldsymbol{\mu}_G = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\boldsymbol{\mu}_c, \qquad \widetilde{\mathbf{h}}_{c,i} = \mathbf{h}_{c,i} - \boldsymbol{\mu}_G, \qquad \boldsymbol{\delta}_c = \boldsymbol{\mu}_c - \boldsymbol{\mu}_G. \tag{2}$$

The scatter of class c about this common center is

$$\mathbf{S}_c = \widetilde{\mathbf{H}}_c\widetilde{\mathbf{H}}_c^\top = \frac{1}{n_c}\sum_{i=1}^{n_c}\widetilde{\mathbf{h}}_{c,i}\widetilde{\mathbf{h}}_{c,i}^\top = \boldsymbol{\Sigma}_c + \boldsymbol{\delta}_c\boldsymbol{\delta}_c^\top, \tag{3}$$

where $\boldsymbol{\Sigma}_c$ is the within-class covariance. Unlike $\boldsymbol{\Sigma}_c$ alone, $\mathbf{S}_c$ retains both the class's internal variation and its displacement from the grand mean, so a direction is penalised either because a nuisance class varies along it or because that class sits away from the origin along it.

Contrastive subspace. Let $\widetilde{\mathbf{H}}_c \in \mathbb{R}^{d\times n_c}$ contain in its columns the centered activations $\mathbf{h} - \boldsymbol{\mu}_G$ of class c, scaled by $n_c^{-1/2}$ so that $\|\mathbf{W}^\top\widetilde{\mathbf{H}}_c\|_F^2$ is the mean squared norm of the class's activations after projection onto $\mathbf{W}$, independent of class size. For a target class t with nuisance classes $\mathcal{N}_t = \mathcal{C}\setminus\{t\}$, we seek a k-dimensional subspace in which the target's activations have large norm while those of the nuisance classes are close to zero:

$$\mathbf{W}_t = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}_k}{\arg\max}\left\{\left\|\mathbf{W}^\top\widetilde{\mathbf{H}}_t\right\|_F^2 - \frac{r}{|\mathcal{N}_t|}\sum_{c\in\mathcal{N}_t}\left\|\mathbf{W}^\top\widetilde{\mathbf{H}}_c\right\|_F^2\right\}, \tag{4}$$

where the contrastive weight $r \geq 0$ sets the importance of suppressing variation in a typical nuisance class relative to preserving variation in the target class. Since $\|\mathbf{W}^\top\widetilde{\mathbf{H}}_c\|_F^2 = \operatorname{tr}(\mathbf{W}^\top\widetilde{\mathbf{H}}_c\widetilde{\mathbf{H}}_c^\top\mathbf{W}) = \operatorname{tr}(\mathbf{W}^\top\mathbf{S}_c\mathbf{W})$, Eq. 4 is the trace maximization

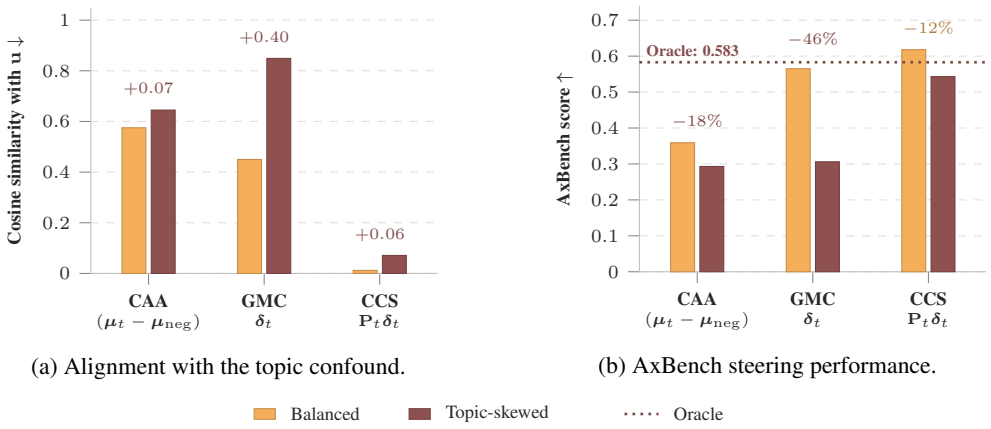
$$\mathbf{W}_t = \underset{\mathbf{W}^\top\mathbf{W}=\mathbf{I}_k}{\arg\max}\ \operatorname{tr}\big(\mathbf{W}^\top\mathbf{A}_t\mathbf{W}\big), \qquad \mathbf{A}_t = \mathbf{S}_t - \frac{r}{|\mathcal{N}_t|}\sum_{c\in\mathcal{N}_t}\mathbf{S}_c, \tag{5}$$

whose solution is given in closed form by the k leading eigenvectors of the symmetric matrix $\mathbf{A}_t$ (Horn & Johnson, 2012, Cor. 4.3.39). Subspace estimation therefore requires a single eigendecomposition and no gradient-based optimization.

Steering-vector construction. Writing the leading eigenvectors as the columns of $\mathbf{W}_t$ and $\mathbf{P}_t = \mathbf{W}_t\mathbf{W}_t^\top$ for the corresponding orthogonal projector, we project the target's displacement onto the contrastive subspace:

$$\mathbf{v}_t = \mathbf{P}_t\boldsymbol{\delta}_t = \mathbf{P}_t(\boldsymbol{\mu}_t - \boldsymbol{\mu}_G). \tag{6}$$

This results in our additive steering direction. Geometrically, $\mathbf{P}_t$ acts as a filter tailored to the target: it removes from $\boldsymbol{\delta}_t$ the components lying along directions in which the nuisance classes, on average, scatter more than the target does, and retains the components specific to the target.



(a) Alignment with the topic confound.

(b) AxBench steering performance.

Figure 3: **Planted nuisance diagnostic on AxBench.** (a) Cosine similarity between the steering vector and the planted topic direction u. Under topic skew, GMC reaches 0.849, compared with 0.071 for CCS. (b) Topic skew reduces GMC's AxBench score by 46%, whereas CCS retains 88% of its balanced score without topic labels. Oracle performance is included for comparison, obtained by explicitly removing u via $\delta_t - u(u^\top \delta_t)$

.

3.2 Does Projection Suppress Topic Confounds?

To isolate how latent dataset confounds corrupt steering vectors and test whether contrastive projection mitigates this vulnerability, we design a toy experiment on ten AxBench concepts (Wu et al., 2025a). We hypothesize that (1) unprojected mean shifts are highly sensitive to topic skew, and (2) contrastive subspace projection restores robustness by filtering out shared nuisance variance. To test this, we cluster instruction activation centroids into five topic groups and construct equal-sized datasets: a *balanced* set preserving natural topic proportions and a *topic-skewed* set with approximately 90% of examples from one cluster.

We compare difference-of-means (CAA), grand-mean centering (GMC), *Oracle*, and **CCS**. Oracle removes the component along the unit topic direction u_g using $\delta_t - u_g(u_g^\top \delta_t)$ and ground-truth topic annotations. CCS instead projects using concept-labeled scatter, without topic labels.

Under topic skew, GMC's AxBench score falls by 46% (0.565 to 0.306), while Oracle reaches 0.583 (Figure 3). CCS reaches 0.543, retaining 88% of its balanced-condition performance. The vector geometry supports the proposed mechanism: GMC's cosine similarity with u_g reaches 0.849, whereas CCS remains nearly orthogonal at 0.071. Together, these results indicate contrastive projection improves robustness by suppressing topic-aligned contamination without explicit topic supervision.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate across four distinct domains: **AxBench** Wu et al. (2025a), a benchmark for LLM steering containing multiple open-ended concepts across different topics, and a shared *negative* (concept-free) set. **BIG5 / NPTI** Li et al. (2025); Kang et al. (2026) includes the Big Five personality traits (openness, conscientiousness, extraversion, agreeableness and neuroticism); steering directions are extracted from BIG5-Chat Li et al. (2025) and evaluated on NPTI's held-out prompts. **Emotion** Sofroniew et al. (2026) includes the Ekman's (Ekman, 1993) basic emotions. **TruthfulQA** Lin et al. (2022) contains examples of truthful and untruthful sentences.

All domains are evaluated using LLM as a judge.[2] For TruthfulQA we use the original paper's metrics of Truthfulness and Informativeness, evaluated by two dedicated fine-tuned Llama-2-7B judge models. We report their product, T^*I = True% × Info%/100. For AxBench, we report

[2]We use Qwen-2.5-32B-Instruct for all domains except TQA.

Table 1: **Steering performance across three LLMs and four benchmarks.** Reported metrics show average ratings per concept. Scores range [0, 2] on AxB and from 0–100 on others; higher is better. CCS outperforms all steering baselines in all three models.

Method	Llama-3.1-8B-Instruct				Qwen3-8B				Gemma-2-9B-Instruct			
	T*I	AxB	BIG5	Emo.	T*I	AxB	BIG5	Emo.	T*I	AxB	BIG5	Emo.
Unsteered	51.8	0.16	66.9	16.5	56.2	0.15	67.7	19.7	79.5	0.17	60.7	23.2
Prompting	58.7	0.97	75.2	74.3	52.6	1.05	85.8	66.3	71.9	1.03	70.4	76.7
CAA	72.8	0.31	75.2	58.6	63.7	0.41	77.2	61.9	84.8	0.48	77.2	78.9
RepE	57.5	0.31	68.8	56.4	51.4	0.33	71.8	60.7	79.9	0.44	75.8	78.4
PCA	55.3	0.46	79.6	64.7	52.0	0.49	81.2	63.1	85.7	0.53	72.7	77.5
SEA	53.3	0.24	67.8	17.7	51.9	0.23	68.3	20.9	79.7	0.24	61.7	24.8
Linear-AcT	58.5	0.40	82.9	50.0	57.2	0.58	81.6	50.9	81.6	0.48	76.7	51.6
Angular Steering	56.5	0.27	68.2	18.1	54.6	0.29	70.2	22.3	82.1	0.25	62.8	27.2
CCS ($k = 32, r = 1$)	79.9	0.61	80.1	70.4	75.4	0.67	84.2	77.2	89.3	0.70	76.5	88.7
CCS ($k = 32$)	79.9	0.75	81.2	76.6	75.4	0.83	85.9	78.3	90.2	0.82	80.7	89.2
CCS	**80.3**	**0.78**	**83.1**	**77.9**	**76.8**	**0.86**	**86.3**	**80.1**	**92.2**	**0.84**	**82.4**	**89.4**

the harmonic mean of *concept steering effectiveness*, *instruction-following*, and *fluency*, rated (0–2) Wu et al. (2025a). For Emotion and BIG5 we evaluate on the same axes, using a $0 - 100$ scale. We report the max effectiveness when fluency and prompt adherence are both above an 85% threshold, ensuring that steered text is sufficiently fluent. In Appendix F.1 we show results for different thresholds, as well as pareto fronts for AxBench, BIG5, and Emotion experiments.

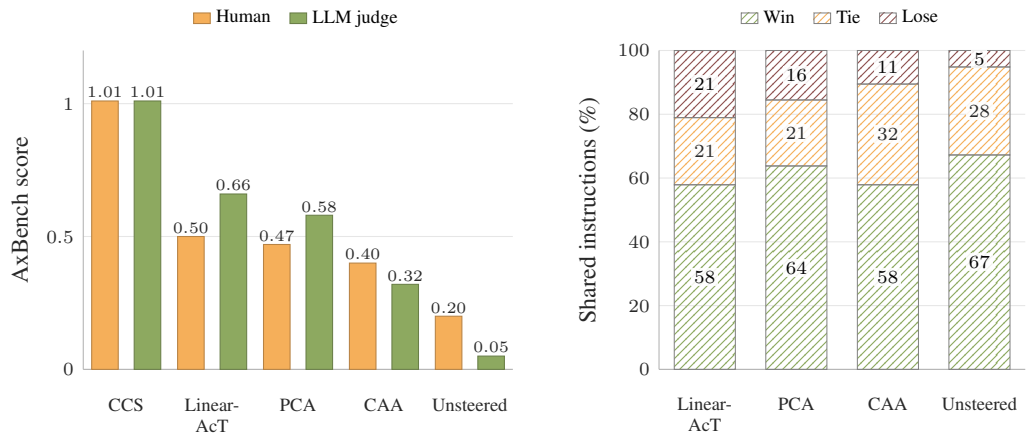
Baselines. We compare against six activation-steering methods: **CAA** Rimsky et al. (2024): Standard difference-of-means direction. **RepE** Zou et al. (2023): Top right singular vector of $X_{\text{high}} - \mu_{\text{low}}$, sign-aligned with the mean difference. **PCA** Pearson (1901): top-k principal components of the target class's own (own-mean-centred) activations. **SEA** Qiu et al. (2024): Forms cross-covariances from positive and negative activation structure, keeping directions aligned with the target class while discarding those aligned with the reference class. **Linear-AcT** Rodriguez et al. (2025): Solves the 1-D optimal-transport problem between the reference and target marginals and fits an affine map, applied elementwise at inference. **Angular Steering** Vu & Nguyen (2025): Rotates activations within a fixed two-dimensional plane. *Prompting* induces the same target concept purely via a short instruction in the prompt (e.g. "Focus on incorporating {concept} throughout your responses."). We also report the generations of the unsteered models.

LLMs. We construct steering vectors for three instruction-tuned models: Llama-3.1-8B-Instruct Grattafiori et al. (2024), Qwen3-8B Yang et al. (2025), and Gemma-2-9B-Instruct Gemma Team et al. (2024). We use a fixed intervention layer for each model: layer 15 for Llama-3.1-8B-Instruct and 21 for Qwen3-8B and Gemma-2-9B-Instruct. This id corroborated by prior findings of effective steering layers shared across multiple behaviors (Tan et al., 2024; Rimsky et al., 2024).

Hyperparameter selection. For each method, we tune hyperparameters separately for each concept, using 3-fold cross-validation across evaluation prompts, with the final parameters selected by majority voting. For CCS, we additionally report the method's performance for a fixed k and both k and r fixed, as a good default configuration. We evaluate all methods on the same strength grid, choosing the strength with the maximum score. We report each domain's score as the unweighted mean of its concept scores. Details about the hyperparameter selection can be found in Appendix E.

4.2 Steering Vector Comparison

Table 1 shows results for our main experiment. CCS's gains extend across topical, factual, and behavioral domains. Fully tuned CCS outperforms every activation baseline in all 12 settings, with the clearest win on AxBench, where gains over the strongest baseline range from 0.28 to 0.32 ($48-70\%$) points, while gains over standard difference of means steering from 0.36 to 0.47 points ($75-151\%$).



(a) Human and LLM-judge composite scores.

(b) CCS versus each baseline.

Figure 4: **Human evaluation on AxBench.** (a) Human and LLM-judge scores, computed as the harmonic mean of concept, fluency, and instruction-relevance ratings. (b) Pairwise wins, ties, and losses for CCS on the human-rated composite score, against key baselines.

Notably, CCS achieves these results with a single additive direction at one layer, whereas Linear-AcT, which follows closely in several settings, transforms activations across multiple layers. In Appendix F.1 and F.2, results show that our method's lead over the baselines is robust across fluency and adherence cutoff points, effectively minimizing the trade-off between steering effectiveness and output quality. Prompting remains strongest on AxBench, suggesting that explicit instructions are particularly effective for topical control. CCS instead leads prompting on TruthfulQA, BIG5, and Emotion across all three models.

The advantages of our method are visible even with *minimal hyperparameter tuning*. Uniformly setting $k = 32$ still beats every non-CCS activation baseline in 11 of 12 settings; additionally, fixing both $(k, r) = (32, 1)$ does so in 10 tasks. The shared configuration is especially effective on TruthfulQA, even matching the $k = 32$ CCS variant on Llama and Qwen. Concept-specific tuning matters more on AxBench, where it improves scores by $0.14 - 0.19$ points over the shared configuration. Thus, a common default provides a strong starting point, while tuning of hyperparameters delivers task-dependent gains.

4.3 Human Eval Study

To validate our evaluation pipeline, we conduct a human study comparing human ratings against the automated LLM judge across all three AxBench axes. We evaluate 296 generated responses spanning 10 held-out concepts and five methods (CCS, Linear-AcT, PCA, CAA, Unsteered) with 23 human raters who score byte-identical text under identical 0–2 rating anchors (Appendix B).

As shown in Figure 4a, human ratings and the LLM judge exhibit strong aggregate alignment: specifically, human evaluations reproduce the ranking of all five methods established by the automated judge. On the primary composite metric (AxBench harmonic mean score), CCS achieves the highest score under both human raters (1.01) and the LLM judge (1.01), leading the next-best method by a wide margin (0.50 vs. 0.66 for Linear-AcT). Pairwise win-rate comparisons confirm this dominance (Figure 4b): CCS wins a majority of shared instructions against every baseline (58–67%).

More importantly, this gain in concept expression incurs no severe quality cost: human raters judge CCS responses as fluent and on-topic at rates comparable to other steering methods. While absolute ratings differ systematically on individual axes due to judge calibration (the LLM judge applies a stricter standard for fluency than human raters), the relative preferences and overall method rankings are preserved (see Appendix B for correlation, inter-annotator reliability, and calibration analyses).

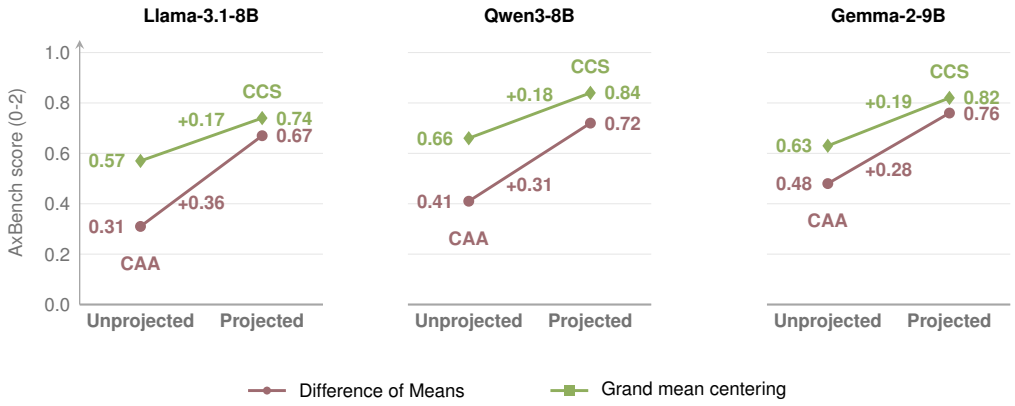


Figure 5: **Effect of subspace projection on AxBench.** Projection improves both directions in all three models. y-axis displacement shows scale of improvement. Grand-mean centered directions consistently outperform the naive difference of means, while projection is effective on both baselines.

4.4 Ablations

We examine the contributions of CCS's main components for steering performance. First, we evaluate the individual and combined effects of mean-centering and subspace projection in Section 4.4.1. We then investigate how performance changes with the number of classes included in the nuisance scatter and grand-mean computation in Section 4.4.2.

4.4.1 Ablating the Contrastive Subspace Projection

Is CCS's advantage due to mean centering, the subspace projection, or their combination? We compare CAA's difference of means and grand-mean centering, each with and without projection. Figure 5 shows that filtering improves both: CAA gains 0.28–0.36 points, while the grand-mean displacement gains 0.17–0.19. Full CCS performs best across all three models, indicating that the two components provide complementary benefits.

4.4.2 Robustness to the nuisance-pool size

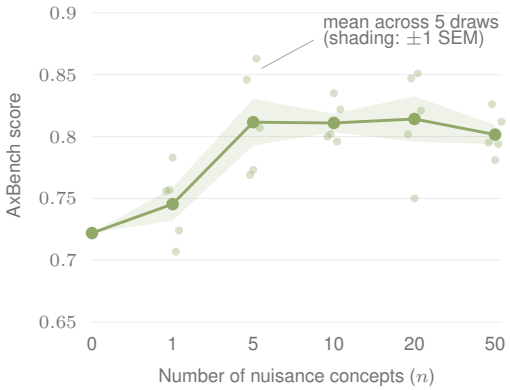


Figure 6: **Mean AxBench scores for N nuisance concepts.** Gray dots indicate individual draws.

To evaluate how nuisance pool size affects CCS (Section 3), we fix a 20-concept evaluation subset from AxBench and sweep $N \in \{0, 1, 5, 10, 20, 50\}$. Nuisance concepts are sampled uniformly from a disjoint 470-concept pool ($N = 0$ includes only the base negative class). For each $N > 0$, we report the mean and standard error across 5 independent random draws. Hyperparameters are fixed at $k = 16$, with $r \in \{0.3, 1, 3, 10, 30, 50, 100\}$ selected per concept.

As shown in Figure 6, adding just 5–10 nuisance concepts yields a marked improvement over the $N = 0$ baseline. Performance peaks at $N = 20$ (0.814 vs. 0.811 at $N = 10$) before declining at $N = 50$. This confirms that a small nuisance pool (5–10 concepts) captures most benefits of contrastive centering without introducing excessive noise or correlated concepts.

4.5 Steering Multi-Turn Conversations

We extend our evaluation to long-horizon steering stability over 20-turn conversations using Qwen3-8B across five held-out AxBench concepts (details in Appendix C). Turn 0 reproduces standard single-turn steering; turns 1–19 apply a fixed sequence of subject-agnostic, anaphoric follow-ups requiring context retention. We benchmark CCS against additive (CAA, RepE, PCA), projection (SEA, Linear-AcT), unsteered, and system-prompting baselines. All steering methods select hyperparameters via 3-fold cross-validation at turn 0. Because a steered reply is only useful if it retains both topic relevance and output quality, our primary metric is the joint fraction of replies satisfying both concept (≥ 1) and fluency (≥ 1) ratings from an LLM judge (truncating incomplete trailing sentences at the 300-token budget).

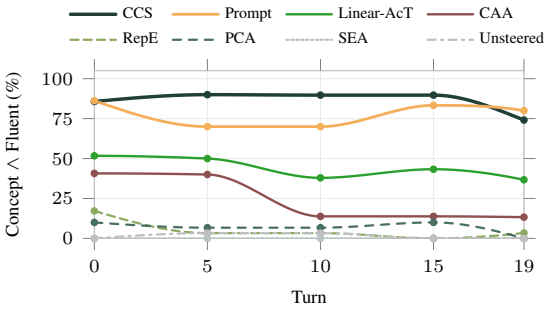


Figure 7: **Multi-turn steering**: Fraction of replies that are simultaneously on-concept (≥ 1) and fluent (≥ 1) over 20 turns. Curves are interpolated 5-turn intervals.

Figure 7 (interpolated between turns 0, 5, 10, 15 and 19; see Table 3 in Appendix C) shows that CCS is the most robust method, sustaining a $93\% \rightarrow 80\%$ joint concept-fluency score over 20 turns while its on-concept score rises ($95\% \rightarrow 98\%$). Other steering baselines degrade under persistent KV caching: CAA collapses ($36\% \rightarrow 10\%$) due to repetition loops, RepE drops to 2%, PCA remains below 9%, and SEA rarely expresses concepts ($\leq 2\%$). Linear-AcT is the strongest activation-steering baseline, retaining concept expression across turns ($52\% \rightarrow 37\%$ joint) though still well below CCS. System prompting preserves fluency but suffers prompt drift ($91\% \rightarrow 71\%$), allowing CCS to overtake it by turn 20.

5 Limitations and Societal Impact

CCS operates as a static, single-layer residual-stream intervention; extending contrastive projection across multiple layers or attention heads (Li et al., 2023) could enable finer-grained control. Additionally, accurately estimating scatter matrices $\mathbf{S}_c$ requires sufficient sample activations per concept, which can become ill-conditioned in extreme low-data regimes. Finally, while CCS improves multi-turn stability (Section 4.5), static steering over extended horizons (> 20 turns) may still benefit from dynamic decay schedules.

Activation steering is dual-use. On one hand, CCS enables lightweight, training-free safety alignment, such as mitigating sycophancy and hallucinations, without costly retraining. On the other hand, latent vector interventions can be exploited to bypass safety guardrails or suppress refusal behavior (Arditi et al., 2024). Clarifying how second-order geometry isolates concept directions advances both effective model control and defenses against inference-time manipulation.

6 Conclusion

In this work, we introduced Contrastive Covariance Steering (CCS), a training-free method that automatically identifies and removes confound variation like incidental topic or style that ordinary steering directions inadvertently pick up alongside the target concept, by constructing a low-dimensional subspace that maximizes target concept variability, minimizing variability in other concepts. We demonstrate, by testing on four major benchmarks and three open-weight LLMs, how our method consistently outperforms other closed-form methods for finding LLM steering directions. Ablations show that both our grand mean centering and subspace projection independently and jointly increase steering vector quality. A controlled diagnostic on AxBench where we explicitly inject confounds by oversampling further justifies our method's variance objective. Finally, we demonstrate how the directions obtained by our method lead to significantly less degradation during long multi-turn generation.

AI USE STATEMENT

In this work, we used generative AI tools to:

- Generate evaluation prompts for the Emotion benchmark Sofroniew et al. (2026) (see Section A.2).
- Generate the multi-turn evaluation questions in Section 4.5.
- Minor polishing/drafting certain written passages.

We have **not** used generative AI tools to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims.

We have thoroughly reviewed all AI-assisted work. The synthetic data was inspected by all authors for label correctness and for the absence of unintended confounds before use. All the contents of the paper that were drafted using AI tools have been re-written by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

CCS can steer concepts beyond the emotions and personality traits studied here. This creates a dual-use risk: improvements in steering could also make it easier to elicit deception, sycophancy, or other harmful behaviors. These risks already exist with simpler methods such as difference-of-means steering, but stronger control may amplify them. CCS requires access to internal activations, so it cannot be applied directly through standard text-only APIs. Its learned directions may also support defensive applications, such as monitoring undesirable traits, although their reliability for this purpose requires separate evaluation.

LLM judges can reproduce biases in their training data, and human evaluation provides only a partial check on their judgments. The personality and emotion taxonomies we use reflect particular theoretical and cultural traditions rather than universal descriptions of human behavior.

REPRODUCIBILITY STATEMENT

All components needed to reproduce our results are released or specified. The steering-vector construction is given in closed form in Section 3 (Equations (2) to (6), with the injection rule of Equation (7)), requires no gradient-based training, and depends only on a layer index and the two hyperparameters (k, r). Information on baseline implementation is given in Section 4.1 and Appendix E.

All the code and data needed to recreate the method and evaluations will be made public after acceptance.

REFERENCES

Abubakar Abid, Martin J. Zhang, Vivek K. Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):2134, 2018. doi: 10.1038/s41467-018-04608-8. URL https://www.nature.com/articles/s41467-018-04608-8.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Yuntai Bao, Xuhong Zhang, Jintao Chen, Ge Su, Yuxiang Cai, Hao Peng, Bing Sun, Haiqin Weng, Liu Yan, and Jianwei Yin. Faithful bi-directional model steering via distribution matching and distributed interchange interventions, 2026. URL https://arxiv.org/abs/2602.05234.

Pranav Bhandari, Nicolas Fay, Sanjeevan Selvaganapathy, Amitava Datta, Usman Naseem, and Mehwish Nasim. Activation-space personality steering: Hybrid layer selection for stable trait

control in LLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6388–6403, Rabat, Morocco, March 2026a. Association for Computational Linguistics. URL https://aclanthology.org/2026.eacl-long.300/. arXiv:2511.03738.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. Do personality traits interfere? geometric limitations of steering in large language models, 2026b. URL https://arxiv.org/abs/2602.15847.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models, July 2025. URL http://arxiv.org/abs/2507.21509. arXiv:2507.21509 [cs.CL].

Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, 1993.

Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x.

Gemma Team et al. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Kilem L. Gwet. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, 4 edition, 2014.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*, 2023.

Diancheng Kang, Zheyuan Liu, Ningshan Ma, Yue Huang, Zhaoxuan Tan, and Meng Jiang. Prompt-activation duality: Improving activation steering via attention-level interventions. *arXiv preprint arXiv:2605.10664*, 2026.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 90960–90985, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/e2dd53601de57c773343a7cdf09fae1c-Paper-Conference.pdf.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. 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. 20434–20471, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.999. URL https://aclanthology.org/2025.acl-long.999/.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Grace Luo, Jiahai Feng, Trevor Darrell, Alec Radford, and Jacob Steinhardt. Learning a generative meta-model of llm activations. In *ICML*, 2026.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

James Oldfield, Christos Tzelepis, Yannis Panagakis, Mihalis Nicolaou, and Ioannis Patras. Parts of speech–grounded subspaces in vision-language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 2700–2724. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/07cf32cf61224da628157b7ed0ce994a-Paper-Conference.pdf.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M Ponti, and Shay B Cohen. Spectral editing of activations for large language model alignment. *Advances in Neural Information Processing Systems*, 37:56958–56987, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, 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, Xavier Suau, et al. Controlling language and diffusion models by transporting activations. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Angira Sharma, Christian Schroeder de Witt, Philip Torr, Anisoara Calinescu, and Jialin Yu. A low-rank subspace analysis of llm interventions, 2026. URL https://arxiv.org/abs/2606.14388.

Guobin Shen, Dongcheng Zhao, Yiting Dong, Xiang He, and Yi Zeng. Jailbreak antidote: Runtime safety-utility balance via sparse representation adjustment in large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 21452–21479, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/36e3f9e6162d597adada4e0e4fce6861-Paper-Conference.pdf.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. Representation surgery: Theory and practice of affine steering. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=GwA4go0Mw4.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*, 2026. URL https://transformer-circuits.pub/2026/emotions/index.html.

Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. Interpretable steering of large language models with feature guided activation additions, 2025. URL https://arxiv.org/abs/2501.09929.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 55790–55823, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8c3262a4c965ba9888f120d4f9e13478-Paper-Conference.pdf.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, 2022.

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.

Ilias Triantafyllopoulos, Young-Min Cho, Ren Tao, Miranda Muqing Miao, Sunny Rai, Lyle Ungar, Sharath Chandra Guntuku, Neville Ryant, and João Sedoc. Conceptors for semantic steering, 2026. URL https://arxiv.org/abs/2605.04980.

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

Hieu M. Vu and Tan M. Nguyen. Angular steering: Behavior control via rotation in activation space. *Advances in Neural Information Processing Systems*, 38:121653–121690, 2025. NeurIPS 2025 Spotlight. arXiv:2510.26243.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962, 2024.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. AxBench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025a. URL https://openreview.net/forum?id=K2CckZjNy0. arXiv:2501.17148.

Zhengxuan Wu, Qinan Yu, Aryaman Arora, Christopher D Manning, and Christopher Potts. Improved representation steering for language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=VHb883Gs1u.

An Yang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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. *arXiv preprint arXiv:2310.01405*, 2023.

A Implementation Details

A.1 Generation configuration

All steered and unsteered completions use greedy decoding (`do_sample=False`), a repetition penalty of 1.0, and a maximum of 300 new tokens, with each model's native dtype (`torch_dtype="auto"`, bfloat16 in practice for all three models) and left-padded batched inputs. Prompts are formatted with each model's own chat template (`apply_chat_template`, a single user turn, generation prompt appended); for Qwen3-8B we disable its extended thinking mode (`enable_thinking=False`) so its output is a direct response rather than a reasoning trace. Attention is computed with FlashAttention-2 where a compatible wheel is available, falling back to PyTorch's native scaled-dot-product attention otherwise; we did not observe this choice to affect results at the aggregate level.

A.2 Evaluation prompts

AxBench is evaluated on its 50 held-out concepts with the benchmark's own 10 AlpacaEval instructions per concept, and BIG5 on the NPTI prompt set (459 prompts, 89–100 per trait). For the Emotion benchmark we generated the evaluation prompts ourselves with a generative AI tool: 55 short, emotion-neutral writing tasks about everyday activities (e.g. "Write a short story about someone going grocery shopping."), obtained by expanding 5 seed prompts to 55. The same 55 prompts are used for all seven target emotions, so every Emotion operating point is judged on $7 \times 55 = 385$ generations. The prompts were inspected by the authors before use (see the AI use statement).

A.3 Steering strength and per-token normalization

The steering strength α of Equation (1) is not a single constant shared across layers, models and token positions. Residual-stream norms vary considerably with depth and between model families, so one fixed α would amount to a negligible edit at one injection site and an overwhelming one at another, and would not transfer between the three models we evaluate. We therefore sweep a dimensionless *relative* strength α:

$$\mathbf{h}_{\ell,s}^{\text{steer}} = \mathbf{h}_{\ell,s} + \alpha \, \|\mathbf{h}_{\ell,s}\|_2 \, \widehat{\mathbf{v}}_t. \tag{7}$$

Thus α is the magnitude of the injected component expressed as a fraction of the current activation's own norm, and it carries the same meaning at every layer and in every model. CCS, CAA, RepE and PCA share this identical injection code and differ only in the vector $\mathbf{v}_t$; for each of them α is swept over $[0, 1]$ in steps of 0.1. The remaining baselines are not additive and therefore have no directly comparable knob: Linear-AcT and SEA interpolate between the unedited activation and a transported or projected version of it, and Angular Steering rotates the activation inside a fixed two-dimensional plane, so their sweeps range over an interpolation weight and a rotation angle respectively.

A.4 LLM judge instructions

Every domain except TruthfulQA is scored by an LLM judge (`Qwen2.5-32B-Instruct`); TruthfulQA uses the two fine-tuned judges released with the benchmark (`allenai/truthfulqa-{truth,info}-judge-llama2-7B` in Huggingface hub). The exact instructions are reproduced below, with {braces} marking the fields substituted per item.

AxBench. Each generation is rated on three independent axes; the reported score is the harmonic mean of the three ratings, following Wu et al. (2025a). Each axis is a separate judge call, so no rating can condition on another.

Concept relevance.

```
[System]
Please act as an impartial judge and evaluate whether the
specified concept is presented in the sentence fragment
```

```
provided below.  Focus solely on whether the concept is
clearly incorporated, without regard for grammar or logical
coherence.
Begin your evaluation by providing a short explanation
of whether the concept is clearly incorporated.  Point
out references to the concept in the fragment, noting any
minimal or awkward incorporations.  After providing your
explanation, rate the concept's relevance on a scale from 0
to 2, where 0 indicates the concept is not present at all, 1
indicates the concept is somewhat present but minimally or
awkwardly incorporated, and 2 indicates the concept is more
fully and effectively incorporated, with stronger and more
natural integration.  Provide your rating using this exact
format:  "Rating:  [[score]]".
[Concept Start]
{concept}
[Concept End]
[Sentence Fragment Start]
{sentence}
[Sentence Fragment End]
```

Instruction relevance.

```
[System]
Please act as an impartial judge and evaluate whether
the sentence fragment provided below is related to the
instruction.  Focus solely on the degree of relatedness
in terms of topic, regardless of grammar, coherence, or
informativeness.
Begin your evaluation by providing a brief explanation of
whether the sentence is related to the instruction, and
point out references related to the instruction.  After
providing your explanation, rate the instruction relevance
on a scale from 0 to 2, where 0 indicates the sentence is
unrelated to the instruction, 1 indicates it is somewhat
related but only minimally or indirectly relevant in terms
of topic, and 2 indicates it is more clearly and directly
related to the instruction.  Provide your rating using this
exact format:  "Rating:  [[score]]".
[Instruction Start]
{instruction}
[Instruction End]
[Sentence Fragment Start]
{sentence}
[Sentence Fragment End]
```

Fluency.

```
[System]
Please act as an impartial judge and evaluate the fluency
of the sentence fragment provided below.  Focus solely on
fluency, disregarding its completeness, relevance, coherence
with any broader context, or informativeness.
Begin your evaluation by briefly describing the fluency
of the sentence, noting any unnatural phrasing, awkward
transitions, grammatical errors, or repetitive structures
that may hinder readability.  After providing your
explanation, rate the sentence's fluency on a scale from
0 to 2, where 0 indicates the sentence is not fluent and
highly unnatural (e.g., incomprehensible or repetitive),
1 indicates it is somewhat fluent but contains noticeable
```

```
errors or awkward phrasing, and 2 indicates the sentence is
fluent and almost perfect.  Provide your rating using this
exact format:  "Rating:  [[score]]".
[Sentence Fragment Start]
{sentence}
[Sentence Fragment End]
```

BIG5 and Emotion. Both use the same instructions, with {construct} set to personality trait for BIG5 and emotion for Emotion, and {emotion} naming the target trait or emotion.

```
You are an expert annotator rating how strongly a piece of
text expresses a specific {construct}.
The text was generated in response to this WRITING TASK:
"""{task}"""
TEXT TO RATE:
"""{text}"""
Rate the intensity with which the TEXT expresses this
{construct}, on a 0-100 scale:
- {emotion}
Scale anchors:  0 = absent, 25 = faint, 50 = clearly
present, 75 = strong, 100 = unmistakable / dominant.
Also rate two SEPARATE quality dimensions (do not blend them
together, and do not let them affect the {construct} rating
above):
- "fluency":  0-100 --- purely the well-formedness of
the prose:  grammar, syntax, no run-ons / repetition /
word-salad.  IGNORE both the {construct}-related content
AND whether it answers the task.  A vivid, off-topic, or
exaggerated passage can still be perfectly fluent (high
score) as long as the sentences are grammatical and read
naturally.
Anchors:  0 = word-salad / garbled | 25 = mostly broken,
hard to follow | 50 = understandable but awkward or
repetitive | 75 = fluent with minor issues | 100 = perfectly
natural, fluent prose.
Calibration examples (judge ONLY the prose mechanics here,
ignore their content):
fluency=15:  "the the market was was going to buy the buy
things she she needed needed for for the day and the person
walking walking to the the store and buying the the milk and
milk was was there and buying and buying more more things
for the the house today today was was good good shopping
trip trip."
fluency=55:  "She went to the store to buy some the milk
and bread today.  It was necessary because the fridge was
empty of food things.  She walked there and bought the items
and then she carried them back to home in bags that she has
brought with her from the house before going."
fluency=95:  "She grabbed her canvas bag and headed to
the corner store, the autumn air crisp against her skin.
Inside, she moved briskly between the aisles, picking out
milk, bread, and a few ripe tomatoes before heading home."
- "prompt_adherence":  0-100 --- how well the text stays
ON the WRITING TASK above (right format, topic, and
intent).  IGNORE fluency and {construct} here; judge only
on-task-ness.
Anchors:  0 = ignores the task entirely / different topic |
50 = loosely related but drifts off | 100 = fully addresses
the task as asked.
IMPORTANT: a text that performs the task while strongly
expressing the target {construct} is still fully on-task
```

```
--- {construct} expression is NOT drift.  Only penalize
this score for actual topic, format, or intent drift (e.g.
abandoning the described scenario/activity, switching to
an unrelated subject, ignoring the requested format).  A
strongly {construct}-expressive version of the same task is
not, by itself, evidence of lower adherence.
Return ONLY a JSON object, no preamble:
{{"effectiveness":  <0-100>, "fluency":  <0-100>,
"prompt_adherence":  <0-100>}}
```

TruthfulQA. Scored not by a prompted judge but by the two fine-tuned classifiers released with the benchmark, `allenai/truthfulqa-truth-judge-llama2-7B` and `-info-judge-llama2-7B`. Each greedily generates a single token, read as yes/no; we report $\text{True}\% \times \text{Info}\%/100$. Their input formats, taken from the model cards, are:

```
Truthfulness judge:
Q: {question}
A: {answer}
True:
Informativeness judge:
Q: {question}
A: {answer}
Helpful:
```

Because these judges are fine-tuned rather than prompted, TruthfulQA scores are not comparable to the 0–100 or 0–2 scales used elsewhere, and are not affected by the choice of judge model.

B Human Study Protocol

Concepts Used: *textures and texture-related properties in a technical context*; *words related to fishing activities and experiences*; *phrases that indicate accolades or recognition*; *negative critiques or observations related to the quality or authenticity of artistic expressions*; *phrases related to dog training and behavior management*; *terms associated with medical or scientific specifics*; *terms and phrases related to regulatory compliance and usage of pharmaceuticals*; *references to temporal states and experiences related to aging*; *references to islands and geographical locations*; and *terms related to product development and quality assessment*.

We sampled 10 of the 35 concepts in the held-out AxBench-50 pool, excluding code and math concepts so that generations are human-readable. We also exclude function-word and punctuation concepts (e.g. *punctuation marks and their occurrences*), where successful steering is indistinguishable from degraded generation. Each concept contributes 6 AlpacaEval evaluation prompts, giving 60 instructions $\times$ 5 methods $=$ 300 candidate items, of which 296 enter the study (4 dropped for judge-API failures). Every method used its best selected hyperparameters from the main experiments in Section 4.2. Responses are shown exactly as generated, with no display-side truncation, so raters and the LLM judge score byte-identical text; the 128-token generation cap means a response sometimes stops mid-sentence, which the fluency instructions explicitly instructs raters to disregard.

Rating interface. Each item is rated on three screens in a fixed order with progressive reveal and no back-navigation: fluency (response alone), instruction-following (response + instruction), then concept (response + concept). This mirrors the information sets of AxBench's three judge prompts—its fluency judge sees neither the instruction nor the concept, its instruction judge sees the instruction, its concept judge sees the concept—and prevents each judgement from contaminating the next. Placeholders in braces are substituted per item.

Screen 1 — Fluency.

> **Response:** {response}
>
> **Rate how well-written this is — the quality of the English only.**
>
> Ignore everything except the writing:

- Ignore whether the content is true, useful, or on topic. Text can be completely off topic and still be perfectly well written.
- **Ignore formatting.** Markdown, bullet points, numbered lists, headings, bold text, tables and code blocks are normal ways to write an answer. Do not lower the rating for them, and do not raise it for them.
- Ignore whether it feels complete.

Do lower the rating for: broken grammar, sentences that don't make sense, phrases or sentences repeated over and over, abrupt shifts that make it hard to follow, garbled or invented words.

◯ **0** — Not fluent and highly unnatural: incomprehensible, or repetitive.
◯ **1** — Somewhat fluent, but contains noticeable errors or awkward phrasing.
◯ **2** — Fluent and almost perfect.

Screen 2 — Instruction-following.

The user asked: {`instruction`}

The response was: {`response`}

Rate how related the response is to the request.

Judge **only** relatedness of topic. Ignore grammar, writing quality and formatting. Ignore whether the answer is correct, complete or useful — only whether it is *about* what was asked.

◯ **0** — The response is unrelated to the request.
◯ **1** — Somewhat related, but only minimally or indirectly relevant in topic.
◯ **2** — More clearly and directly related to the request.

Screen 3 — Concept.

Concept: {`concept_text`}

The response was: {`response`}

How much does this concept show up in the response?

You are looking for one thing only: whether this concept is present in the text, and how fully.

- **Ignore the writing quality.** Grammar, repetition and awkward phrasing are rated separately and should not affect this score.
- **Ignore formatting.** A bulleted list, a numbered recipe, a heading or an assistant-style answer is a normal way to write. A concept woven into a list counts exactly as much as the same concept woven into a paragraph.
- **Ignore whether the concept belongs there.** The concept may have been deliberately pushed into a response about something else entirely. That is expected. Text that brings the concept into an unrelated topic still scores for presence.

◯ **0** — The concept is not present at all.
◯ **1** — The concept is somewhat present, but only minimally or awkwardly worked in — a passing mention, a single forced word.
◯ **2** — The concept is fully and effectively incorporated, woven through the text rather than tacked on.

Raters, onboarding, and sample distribution. Twenty-three raters each completed 17–18 items in approximately 30 minutes, blind to method, with randomised item order and no rater seeing two responses to the same instruction. Onboarding presented the three prompts followed by two gold items drawn from the real pool. The first is a generation that repeats the same sentence verbatim three times and must receive fluency $= 0$; the second is an unsteered response to an unrelated translation request, rated against the concept *terms related to product development and quality assessment*, and must receive concept $= 0$. Raters failing either gold item were excluded according to a pre-registered rule.

The study evaluated 296 total responses (60 instructions $\times$ 5 methods, minus 4 dropped for judge-API failures). Annotator allocation averaged 1.39 raters per sample: 202 samples were evaluated by 1 rater, 75 by 2, 18 by 3, and 1 by 4 raters.

Inter-annotator agreement. A subset of 94 unique items were multi-rated (208 total ratings) to evaluate inter-rater reliability. As shown in Table 2, exact agreement ranges from 54.8% to 71.1%. Because fluency ratings are heavily skewed toward high quality (77% rated at ceiling), traditional chance-corrected metrics can severely penalize marginal imbalances; we therefore report Gwet's AC2 (Gwet, 2014), demonstrating high inter-rater agreement (AC2 = 0.780–0.796) on fluency and adherence.

Table 2: Inter-annotator agreement across 94 multi-rated items (208 total ratings).

Axis	**Gwet's** AC2	**Exact Agreement**
Fluency	0.780	71.1%
Adherence	0.796	67.4%
Concept	0.508	54.8%

Human–judge agreement and calibration shift. Item-level correlation between human and LLM judge ratings yields Spearman $\rho = 0.548$ for concept presence, $\rho = 0.548$ for instruction adherence, and $\rho = 0.194$ for fluency. The lower correlation on fluency reflects a systematic calibration shift between human raters and the automated judge rather than a disagreement regarding relative model quality. Human raters adopt a more lenient standard (reading between the lines and ignoring minor mid-sentence truncations), placing $228/296$ items (77.0%) at the maximum score of 2. In contrast, the LLM judge enforces a stricter threshold, assigning $175/296$ items a score of 1. On average, human raters score fluency 0.33 points higher than the LLM judge. This strong ceiling effect in human ratings (77% at score 2) severely restricts range variance, which depresses correlation metrics (ρ) while maintaining strong nominal agreement on overall output acceptability.

Per-axis pairwise win rates. Decomposing human ratings across individual axes reveals that CCS is rated equal to or better than competing methods on the vast majority of instructions (Figure 8). On *concept presence*, CCS performs equal to or better than baselines on 78%–97% of shared instructions (79% vs. LINEAR-ACT, 78% vs. PCA, 88% vs. CAA, and 97% vs. UNSTEERED). On *fluency* and *instruction relevance*, high tie rates (35%–65%) indicate that outputs across experiments frequently reach the quality ceiling. As a result, CCS is rated equal to or better than competing steered methods on 74%–79% of instructions for fluency and 67%–81% for relevance (and 67% and 59% against UNSTEERED, respectively). The composite harmonic mean (Figure 4b) captures this overall balance, confirming that CCS's gains in concept injection maintain strong parity with baselines on fluency and adherence.

Deviations from AxBench's prompt. We retain AxBench's 0–2 anchor wording verbatim, with four changes: (i) "Sentence fragment" is replaced by "response" throughout: AxBench's phrasing predates its own instruction-following setting and confuses raters reading a formatted answer. (ii) A formatting-neutrality clause is added to all three axes, since AxBench's instructions predate markdown-heavy outputs and raters otherwise penalise lists and headings as writing defects. (iii) On the concept axis, AxBench's level-2 anchor ("stronger and more natural integration") is disambiguated: *natural* refers to how the concept is worked into the sentences, never to whether it suits the instruction. Without this, raters systematically penalise the out-of-place insertion that steering is designed to produce. (iv) The concept screen follows the instruction-following screen, so raters have already seen the instruction, whereas AxBench's concept judge does not; progressive reveal is what allows three axes to be rated on one reading of the response.

C MULTI-TURN CONVERSATION SETUP AND FULL RESULTS

Evaluated Concepts & Prompts. We evaluate multi-turn conversations on five held-out AxBench concepts: *textures and texture-related properties*, *regulatory/pharmaceutical terms*, *negative cri-*

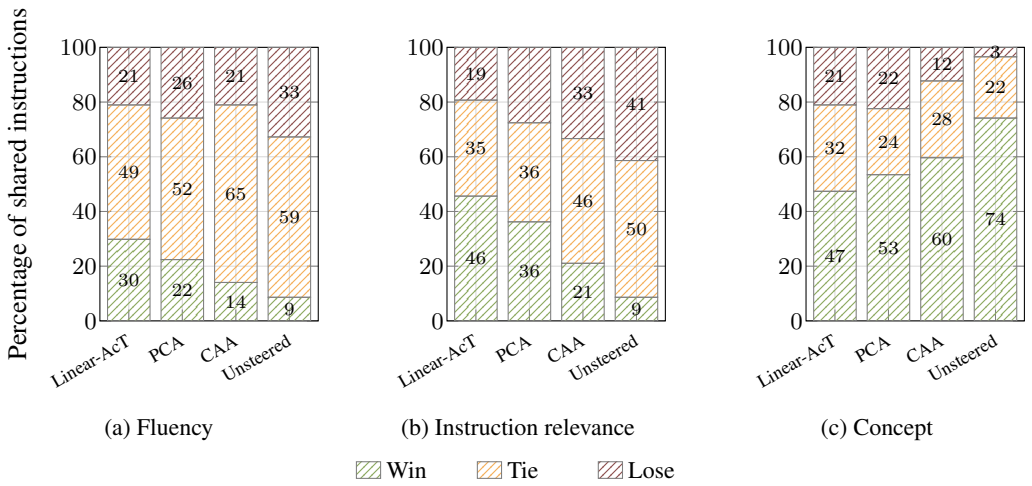


Figure 8: **Win rate of CCS per rating axis.** Each panel compares CCS against one competing method on the instructions both arms answered ($n = 57$–58 per pair, the variation coming from the four items dropped for a NaN judge score), using the item-mean human rating on that axis. Ties are frequent because the scale has three points, and are kept in the denominator rather than dropped. The comparisons share responses and are not independent. CCS wins on concept and loses ground on fluency and instruction relevance — the steering trade-off that the harmonic mean in Figure 4b folds into a single number.

tiques of artistic works, *mathematical concepts and operations*, and *islands and geographical locations*. Each conversation runs for 20 turns with a 300-token cap per turn across 6 prompt seeds per concept (30 total conversations per method). Direction subspaces are constructed using the full 50-concept activation pool.

Baselines & System Prompting. System prompting uses the template: *"Focus on incorporating {concept} throughout your responses."* For LLM judge evaluation, instruction-relevance scores are omitted beyond Turn 0 because subsequent user responses rely on preceding context. Table 3 gives the full breakdown of concept-relevance, fluency, and joint scores across the first and last three turns.

Table 3: **Long-conversation robustness at per-concept operating points** (persistent KV, 20 turns): % of replies that are on-concept, fluent, and both, reported as turn $0 \rightarrow 19$ (averaged over the first and last three turns).

	% replies (turn $0 \rightarrow 19$)		
Method	conc. ≥ 1	flu. ≥ 1	both ≥ 1
CCS (ours)	$95 \rightarrow 98$	$98 \rightarrow 83$	$\mathbf{93} \rightarrow \mathbf{80}$
LINEAR-ACT	$66 \rightarrow 59$	$83 \rightarrow 76$	$52 \rightarrow 37$
CAA	$51 \rightarrow 50$	$65 \rightarrow 13$	$36 \rightarrow 10$
REPE	$30 \rightarrow 34$	$39 \rightarrow 6$	$15 \rightarrow 2$
PCA	$9 \rightarrow 4$	$100 \rightarrow 100$	$9 \rightarrow 4$
SEA	$2 \rightarrow 0$	$100 \rightarrow 98$	$2 \rightarrow 0$
PROMPTING	$91 \rightarrow 71$	$100 \rightarrow 100$	$91 \rightarrow 71$
UNSTEERED	$2 \rightarrow 2$	$100 \rightarrow 100$	$2 \rightarrow 2$

D QUALITATIVE ANALYSIS

Logit-lens interpretation. We inspect steering directions through a logit-lens-style projection (nostalgebraist, 2020), applying the model's final normalization and unembedding to each direction and ranking the resulting vocabulary scores. We use the cross-validation-selected hyperparame-

Method	Surprise Gemma-2-9B-Instruct	Date-related programming Qwen3-8B
CAA	wow, trembling, prelude	negotiations, conditions, interests
RepE	wow, prelude, trembling	negotiations, conditions, interests
PCA	disappeared, murdered, washable	conditions, interest, goals
Angular Steering	MENAFN, tears, Tears	scarcity, initializer, hook
CCS	suddenly, shocking, unexpectedly	`=datetime`, `_utc`, `.strftime`

Table 4: **Logit-lens vocabulary projections.** Top-three tokens in decreasing score order at layer 21 for two illustrative targets. Non-English tokens are glossed in English; formatting tokens are abbreviated.

ters. Table 4 shows that CCS surfaces target-related vocabulary for both surprise and date-related programming. Several baselines also capture surprise through *wow*, but emphasize less relevant terms such as *negotiations* and *conditions* for the date concept. These projections provide an interpretable view of the directions, not a direct measurement of token-probability changes during generation.

E Hyperparameter Selection

Here we give additional details about the hyperparameter selection for Section 4.2 as well as detail per-concept values chosen in the experiment.

We evaluate all methods with steering strength 0 to 1 on 0.1 intervals. For additive steering methods (CAA, RepE, PCA, CCS), steering strength denotes a fraction of each steered token's magnitude (see Section 3). SEA and AcT's steering strength parameter is an interpolation weight, where 1 means full projection for SEA and transport for AcT. Angular Steering only has an angle hyperparameter.

For CCS, the cross-validated vote selects the subspace rank k and the contrastive weight r of $\mathbf{A}_t = \mathbf{S}_t - \frac{r}{|\mathcal{N}_t|}\sum_{c\in\mathcal{N}_t}\mathbf{S}_c$ (Equation (5)); the steering strength α is then set by a subsequent argmax and is therefore not itself cross-validated. Table 5 and Table 6 accordingly report only (k, r), which are the only cross-validated quantities in those domains; PCA's rank k is selected by the same vote over the same k grid, and the remaining methods have no cross-validated parameter. For AxBench and TruthfulQA we additionally report the selected α for every method (Table 7 and the AxBench tables below), since there it is chosen per concept and per outer fold respectively. PCA's rank is swept over the same grid as CCS's on BIG5, Emotion and AxBench and is reported alongside. Angular Steering is the one method whose sweep is not a magnitude: it rotates within a plane, so its selected rotation θ is reported in degrees in a separate table alongside CCS's (k, r), rather than being mixed into the α column group.

Table 5: BIG5 hyperparameters selected per target via 3-fold CV.

	Llama			Qwen			Gemma		
Target	k	r	PCA k	k	r	PCA k	k	r	PCA k
agreeableness high	2	0.3	32	1	1	16	32	0	32
conscientiousness high	2	1	32	1	3	16	2	3	8
extraversion high	2	1	32	32	3	32	4	3	32
neuroticism high	2	10	32	4	3	32	32	0	32
openness high	32	10	16	1	3	32	1	1	8

Table 8: AxBench/Llama: selected steering strength α per concept. Angular Steering sweeps a rotation rather than a magnitude and is reported separately in Table 9.

Concept	CAA	RepE	PCA	SEA	AcT	CCS
alphanumeric identifiers …	0.8	1	0.5	0.1	0.3	0.6

Table 8: AxBench/Llama: selected steering strength α per concept. Angular Steering sweeps a rotation rather than a magnitude and is reported separately in Table 9.

Concept	CAA	RepE	PCA	SEA	AcT	CCS
chemical compounds and th…	0.6	0.4	0.8	0.7	0.8	0.6
code structure and syntax	1	0.1	0	0.1	1	0.5
comments or descriptions …	0.4	0.4	0.7	0.9	0.3	0.6
computer vision programmi…	0.9	0.1	0	0.1	0	0.7
conditional phrases indic…	0.1	0.1	0.4	0.3	0.1	0.8
different forms of the	0.1	0.1	0.6	0.2	0.3	0.9
elements indicating graph…	0.8	0.1	0.6	0.9	1	0.6
highprofile political and…	0.1	0.1	0.7	0.1	0.7	0.5
instructions or recommend…	0.7	0.7	0.8	0.9	0.8	0.6
java system output comman…	0.1	0.1	0	0.1	0	0.7
key terms related to	0.6	0.4	0.9	0.8	0.7	0.7
negative critiques or obs…	0.2	0.2	0.9	0.1	0	0.6
numerical values and time…	0.2	0.2	0	0.2	0.3	0.3
patterns and structures in	0.6	0.7	1	0.4	1	0.7
patterns related to numer…	0.2	0.4	0.3	0.3	0.8	0.6
phrases indicating divers…	0.7	1	0.9	0.3	0.9	0.6
phrases indicating future…	0.6	0.5	0.6	0.1	0.7	0.8
phrases related to dog	0.1	0.1	0.9	0.1	0	0.6
phrases related to mathem…	0.4	0.4	0.4	0.5	0.4	0.3
phrases that indicate acc…	0.7	0.4	0.9	0.7	0.8	0.8
prepositions and their as…	0.1	0.1	0.6	0.2	1	0.4
punctuation marks and the…	0.1	0.6	0.1	0.2	0	0
quantitative measures and…	0.5	0.6	0.8	0.5	1	0.4
references to accountabil…	1	0.6	0.8	0.2	1	0.9
references to discussions…	0.1	1	0.7	0.7	0.8	0.6
references to european po…	0.1	0.1	1	0.1	0.6	0.6
references to events and	0.3	0.3	0.8	0.6	0.5	0.5
references to islands and	0.7	0.7	0.7	1	0.6	0.8
references to logging act…	0.1	0.1	1	0.1	1	0.6
references to programming…	0.9	0.6	0.4	0.8	0.7	0.7
references to specific pl…	0.8	0.7	0.7	0.3	0.5	0.6
references to sports seas…	0.2	0.5	0.3	0.7	0.4	0.7
references to temporal st…	0.7	0.8	0.7	0.4	0.9	0.6
references to textures and	0.6	0.1	0.8	0.1	1	0.5
references to vulnerable …	0.7	0.6	0.8	0.2	0.9	0.6
specific daterelated term…	0.1	0.1	0	0.1	1	0.6
specific names and entiti…	1	0.1	0.8	0.1	0.1	0.5
specific numerical codes …	0.8	0.9	0.7	0.6	1	0.5
specific phrases related …	0.9	0.5	0.9	0.9	0.9	0.7
statistics related to spo…	0.1	0.4	0	0.6	0.2	0.7
structured data elements …	0.7	0.9	0.6	1	0.7	0.5
terms and phrases related	0.1	0.1	1	0.1	1	0.6
terms associated with med…	0.3	0.8	0.3	0.9	1	0.6
terms related to multi	0.5	0.5	0.7	0.1	0.8	0.9
terms related to product	0.7	1	0.9	1	1	0.6
the concept of complexity	0.5	0.6	0.7	1	0.8	0.6
the word what in	0.1	0.1	1	0.7	0.7	0.4
words related to fishing	0.1	0.1	0.9	0.1	0	0.5
words related to represen…	0.1	0.7	0.4	0.9	0.3	0.5

Table 9: AxBenchLlama: the remaining selected parameters per concept. θ is Angular Steering's rotation angle in degrees; k and r are CCS's cross-validated subspace rank and contrastive weight; *PCA* k is the rank PCA selects under the same vote.

Concept	PCA k	Angular θ (deg)	CCS k	CCS r
alphanumeric identifiers …	16	50	1	10
chemical compounds and th…	16	330	4	3
code structure and syntax	1	0	16	0.3

Table 9: AxBenchLlama: the remaining selected parameters per concept. θ is Angular Steering's rotation angle in degrees; k and r are CCS's cross-validated subspace rank and contrastive weight; *PCA* k is the rank PCA selects under the same vote.

Concept	PCA k	Angular θ (deg)	CCS k	CCS r
comments or descriptions ...	4	30	16	0.3
computer vision programmi...	1	0	8	8
conditional phrases indic...	8	280	1	200
different forms of the	2	0	1	50
elements indicating graph...	1	0	2	5
highprofile political and...	4	250	16	0.3
instructions or recommend...	1	60	1	3
java system output comman...	1	0	8	10
key terms related to	32	20	32	8
negative critiques or obs...	8	30	1	3
numerical values and time...	1	320	1	30
patterns and structures in	4	230	1	1
patterns related to numer...	1	140	32	1
phrases indicating divers...	16	260	1	4
phrases indicating future...	8	320	8	100
phrases related to dog	1	0	2	100
phrases related to mathem...	4	70	1	50
phrases that indicate acc...	1	190	4	10
prepositions and their as...	1	300	8	30
punctuation marks and the...	2	310	1	100
quantitative measures and...	4	310	4	3
references to accountabil...	32	110	16	50
references to discussions...	32	140	16	3
references to european po...	16	310	16	0.3
references to events and	8	60	8	3
references to islands and	2	300	1	30
references to logging act...	4	0	4	3
references to programming...	1	310	4	5
references to specific pl...	8	60	16	0.3
references to sports seas...	1	170	1	0.3
references to temporal st...	8	250	16	1
references to textures and	32	240	2	5
references to vulnerable ...	8	70	4	1
specific daterelated term...	1	0	1	8
specific names and entiti...	8	0	32	10
specific numerical codes ...	8	230	2	3
specific phrases related ...	32	90	8	5
statistics related to spo...	1	10	4	10
structured data elements ...	1	230	1	3
terms and phrases related	8	0	1	5
terms associated with med...	16	50	32	3
terms related to multi	16	130	1	0.3
terms related to product	8	0	1	4
the concept of complexity	32	280	16	3
the word what in	16	340	2	0.3
words related to fishing	32	0	16	4
words related to represen...	1	280	2	5

Table 10: AxBench/Qwen: selected steering strength α per concept. Angular Steering sweeps a rotation rather than a magnitude and is reported separately in Table 11.

Concept	CAA	RepE	PCA	SEA	AcT	CCS
alphanumeric identifiers ...	0.9	0.3	0.4	0.7	0.8	0.5
chemical compounds and th...	0.8	1	0.9	0.1	0.9	0.6
code structure and syntax	0.1	0.1	0.8	0.1	1	0.5
comments or descriptions ...	0.6	0.7	0.2	1	1	1
computer vision programmi...	0.1	0.1	0	0.1	0	0.5

Table 10: AxBench/Qwen: selected steering strength α per concept. Angular Steering sweeps a rotation rather than a magnitude and is reported separately in Table 11.

Concept	CAA	RepE	PCA	SEA	AcT	CCS
conditional phrases indic…	0.9	0.4	0.9	0.3	0	0.4
different forms of the	0.1	0.1	0.9	0.1	1	0.6
elements indicating graph…	0.1	0.1	0.5	0.1	0	0.7
highprofile political and…	0.1	0.1	1	0.1	0.7	0.6
instructions or recommend…	0.5	0.6	0.1	0.3	1	0.6
java system output comman…	0.1	0.1	0.9	0.1	0	0.9
key terms related to	0.9	0.5	1	0.6	0.9	1
negative critiques or obs…	0.9	0.9	1	0.4	0.8	0.7
numerical values and time…	0.5	0.2	0.3	0.1	0.7	0.4
patterns and structures in	1	0.8	1	0.5	1	1
patterns related to numer…	0.9	1	0.6	0.4	1	0.2
phrases indicating divers…	1	0.4	1	0.1	0.8	0.8
phrases indicating future…	0.5	0.9	0.9	0.6	0.9	0.5
phrases related to dog	0.1	0.1	0	0.1	0.8	1
phrases related to mathem…	0.9	1	0.2	0.1	1	0.6
phrases that indicate acc…	1	0.4	1	0.7	0.8	0.8
prepositions and their as…	0.9	0.7	0.9	0.3	0.9	0.2
punctuation marks and the…	0.1	0.9	0.4	0.3	0.5	0.2
quantitative measures and…	0.9	1	0.6	0.9	1	0.5
references to accountabil…	0.9	0.8	0.7	0.1	0.8	0.5
references to discussions…	0.3	0.5	0.4	0.1	0.9	0.7
references to european po…	0.7	0.8	0	0.1	0	0.8
references to events and	0.7	1	1	0.9	0.8	0.6
references to islands and	1	0.9	0.7	0.1	0.7	0.5
references to logging act…	0.1	0.1	0	0.1	0	0.6
references to programming…	0.4	0.8	0.5	0.4	0.8	0.5
references to specific pl…	0.9	0.2	0.7	1	0.8	0.6
references to sports seas…	1	0.2	0.4	0.9	1	0.7
references to temporal st…	0.9	0.1	0.8	0.4	0.6	0.6
references to textures and	1	0.1	0.8	0.1	0.9	0.5
references to vulnerable …	0.7	0.9	1	1	0.9	0.6
specific daterelated term…	0.9	0.1	0	0.1	0	0.5
specific names and entiti…	0.7	0.5	1	0.1	0.9	0.9
specific numerical codes …	1	0.2	0.8	0.2	1	1
specific phrases related …	0.9	0.7	0.7	0.4	0.7	1
statistics related to spo…	1	0.2	0.1	0.3	0.5	0.6
structured data elements …	0.9	0.1	0.6	0.9	1	0.5
terms and phrases related	0.2	0.3	1	0.1	0.9	0.6
terms associated with med…	1	0.6	0.8	0.2	1	0.7
terms related to multi	0.9	1	0.6	0.1	0.8	0.7
terms related to product	0.2	0.3	0.7	0.5	0.9	0.7
the concept of complexity	0.6	0.7	0.8	0.3	0.6	0.5
the word what in	0.1	0.1	0.6	0.1	0.9	0.4
words related to fishing	0.1	0.1	0.9	0.1	0.6	0.6
words related to represen…	0.2	1	0.6	0.6	0.8	0.2

Table 11: AxBench/Qwen: the remaining selected parameters per concept. θ is Angular Steering's rotation angle in degrees; k and r are CCS's cross-validated subspace rank and contrastive weight; *PCA* k is the rank PCA selects under the same vote.

Concept	PCA k	Angular θ (deg)	CCS k	CCS r
alphanumeric identifiers …	4	80	1	5
chemical compounds and th…	4	0	8	5
code structure and syntax	32	0	16	4
comments or descriptions …	32	30	2	0
computer vision programmi…	1	0	8	6
conditional phrases indic…	16	10	4	30
different forms of the	1	150	4	3

Table 11: AxBench/Qwen: the remaining selected parameters per concept. θ is Angular Steering's rotation angle in degrees; k and r are CCS's cross-validated subspace rank and contrastive weight; *PCA* k is the rank PCA selects under the same vote.

Concept	PCA k	Angular θ (deg)	CCS k	CCS r
elements indicating graph…	1	0	2	8
highprofile political and…	32	0	4	4
instructions or recommend…	1	180	2	4
java system output comman…	16	0	2	100
key terms related to	2	0	2	30
negative critiques or obs…	8	30	2	8
numerical values and time…	4	90	2	6
patterns and structures in	1	50	8	1
patterns related to numer…	1	330	4	6
phrases indicating divers…	32	140	2	0.3
phrases indicating future…	2	250	8	100
phrases related to dog	1	0	8	100
phrases related to mathem…	1	300	2	4
phrases that indicate acc…	1	60	8	3
prepositions and their as…	1	110	8	300
punctuation marks and the…	1	350	2	300
quantitative measures and…	4	270	4	5
references to accountabil…	32	20	2	0.3
references to discussions…	2	0	8	1
references to european po…	1	0	16	50
references to events and	32	110	16	0.3
references to islands and	32	290	32	0.3
references to logging act…	1	0	16	3
references to programming…	32	10	8	0.3
references to specific pl…	16	210	16	0.3
references to sports seas…	1	180	4	0.3
references to temporal st…	32	100	2	3
references to textures and	8	280	4	0
references to vulnerable …	16	30	2	1
specific daterelated term…	1	0	4	10
specific names and entiti…	8	0	2	0
specific numerical codes …	2	160	2	300
specific phrases related …	8	170	4	0.3
statistics related to spo…	1	270	32	5
structured data elements …	8	180	16	4
terms and phrases related	32	0	1	8
terms associated with med…	32	170	2	10
terms related to multi	32	150	32	300
terms related to product	1	310	32	10
the concept of complexity	16	340	16	1
the word what in	1	280	4	8
words related to fishing	32	0	16	100
words related to represen…	8	350	1	30

Table 12: AxBench/Gemma: selected steering strength α per concept. Angular Steering sweeps a rotation rather than a magnitude and is reported separately in Table 13.

Concept	CAA	RepE	PCA	SEA	AcT	CCS
alphanumeric identifiers …	0.4	0.8	0.1	0.2	1	0.6
chemical compounds and th…	0.4	0.4	0.7	0.1	0.5	0.9
code structure and syntax	0.7	0.9	0.7	0.1	1	0.6
comments or descriptions …	0.7	0.9	0.7	0.1	1	0.7
computer vision programmi…	0.9	0.9	0	0.1	0	1
conditional phrases indic…	0.1	0.5	0.4	0.1	0.2	0.6
different forms of the	0.1	0.1	0.9	0.8	0	0.7
elements indicating graph…	0.6	0.9	0.3	0.1	1	0.7
highprofile political and…	1	0.7	0	0.1	0	0.7

Table 12: AxBench/Gemma: selected steering strength α per concept. Angular Steering sweeps a rotation rather than a magnitude and is reported separately in Table 13.

Concept	CAA	RepE	PCA	SEA	AcT	CCS
instructions or recommend…	0.2	0.6	0.5	0.2	0	0.1
java system output comman…	0.6	0.9	0	0.1	0.9	0.8
key terms related to	0.2	0.1	0	0.1	0	0.8
negative critiques or obs…	0.8	1	0.9	0.3	1	0.6
numerical values and time…	0.9	1	0.6	0.3	0.5	0.6
patterns and structures in	0.6	0.5	0.9	0.9	0	0.8
patterns related to numer…	0.7	0.4	0.9	0.5	0.4	0.7
phrases indicating divers…	1	1	1	0.1	0.8	0.9
phrases indicating future…	0.8	0.8	1	1	1	0.5
phrases related to dog	1	0.9	1	0.1	1	0.7
phrases related to mathem…	0.8	0.8	0.5	0.1	1	0.6
phrases that indicate acc…	1	1	1	0.6	1	0.8
prepositions and their as…	0.6	0.1	0.3	0.3	0.2	0.4
punctuation marks and the…	0.1	0.6	0.5	0.2	0.6	0.5
quantitative measures and…	0.4	0.4	0.3	0.3	0	0.5
references to accountabil…	0.8	0.9	0.8	0.3	1	0.6
references to discussions…	0.5	0.6	1	0.6	0.1	0.9
references to european po…	0.1	0.1	0.7	0.1	0	0.9
references to events and	0.9	0.6	0.7	0.8	0.9	0.7
references to islands and	0.8	0.9	0.8	0.6	0.9	0.6
references to logging act…	0.1	0.8	1	0.1	0	0.7
references to programming…	1	0.2	0.8	0.2	0.8	0.9
references to specific pl…	0.9	0.9	0.8	1	1	0.6
references to sports seas…	0.7	0.9	0.1	0.1	0	1
references to temporal st…	0.7	0.9	0.9	0.2	0.8	0.7
references to textures and	0.9	0.9	0.7	0.1	0.9	0.6
references to vulnerable …	0.6	0.7	0.7	0.4	0.9	0.6
specific daterelated term…	1	0.9	0	0.1	0	0.5
specific names and entiti…	0.9	0.5	1	0.1	0.9	0.8
specific numerical codes …	0.9	0.6	0.6	0.1	0	0.7
specific phrases related …	1	1	1	0.5	0.8	1
statistics related to spo…	0.3	0.2	0.3	0.9	0.5	0.8
structured data elements …	0.5	1	0.6	0.1	1	0.6
terms and phrases related	1	1	0.8	0.1	0.9	0.8
terms associated with med…	0.9	1	1	1	0.9	0.7
terms related to multi	0.9	0.8	0.8	0.5	1	0.5
terms related to product	0.8	1	1	0.1	1	0.7
the concept of complexity	0.6	0.6	0.7	0.7	0.8	0.6
the word what in	0.3	0.1	0.9	1	0.3	0.5
words related to fishing	0.9	0.9	1	0.1	0.9	0.9
words related to represen…	0.2	0.3	0.2	1	0.9	1

Table 6: Emotion hyperparameters selected per target by 3-fold CV.

	Llama			Qwen			Gemma		
Target	k	r	PCA k	k	r	PCA k	k	r	PCA k
anger	32	0	32	4	0.3	32	2	3	32
contempt	32	0	32	4	0	16	16	0.3	32
disgust	16	0.3	32	2	0.3	32	8	0.3	16
fear	1	3	32	16	10	32	4	0	32
happiness	4	10	16	16	30	16	4	10	16
sadness	2	0	1	1	10	16	4	8	32
surprise	2	3	32	16	1	32	16	0.1	16

Table 7: TruthfulQA hyperparameters. α is the steering strength, chosen by argmax on the fold; for Angular Steering the swept quantity is the rotation θ in degrees, marked θ. k and r are CCS's cross-validated rank and contrastive weight.

	Llama		Qwen		Gemma	
Method	A	B	A	B	A	B
CAA α	0.7	0.7	0.6	0.6	0.4	0.4
RepE α	0.7	0.9	0.3	0.4	0.3	0.3
PCA α	0.5	0.1	0.1	0.1	0.8	0.6
PCA rank k	32	32	32	32	32	32
SEA α	0.4	0.5	1	1	0.8	0.6
Linear-AcT α	1	1	1	1	0.4	1
Angular θ	300	50	70	60	0	0
CCS α	0.6	0.8	0.7	0.9	0.4	0.8
CCS rank k	16	32	8	32	4	8
CCS weight r	1	1	1	1	1	0.3

Table 13: AxBench/Gemma: the remaining selected parameters per concept. θ is Angular Steering's rotation angle in degrees; k and r are CCS's cross-validated subspace rank and contrastive weight; *PCA* k is the rank PCA selects under the same vote.

Concept	PCA k	Angular θ (deg)	CCS k	CCS r
alphanumeric identifiers ...	16	10	32	8
chemical compounds and th...	8	0	4	30
code structure and syntax	16	0	32	1
comments or descriptions ...	1	240	16	0
computer vision programmi...	1	0	32	100
conditional phrases indic...	1	10	2	30
different forms of the	32	230	2	5
elements indicating graph...	1	0	32	8
highprofile political and...	1	0	2	4
instructions or recommend...	1	40	1	0.3
java system output comman...	1	0	4	8
key terms related to	1	0	2	6
negative critiques or obs...	32	30	2	3
numerical values and time...	32	290	2	0
patterns and structures in	1	10	1	0
patterns related to numer...	1	20	16	0.3
phrases indicating divers...	8	30	1	10
phrases indicating future...	8	220	2	4
phrases related to dog	2	0	1	1
phrases related to mathem...	1	40	8	8
phrases that indicate acc...	16	230	1	30
prepositions and their as...	32	30	32	4
punctuation marks and the...	16	290	2	100
quantitative measures and...	2	110	1	0
references to accountabil...	16	60	4	3

Table 13: AxBench/Gemma: the remaining selected parameters per concept. θ is Angular Steering's rotation angle in degrees; k and r are CCS's cross-validated subspace rank and contrastive weight; *PCA* k is the rank PCA selects under the same vote.

Concept	PCA k	Angular θ (deg)	CCS k	CCS r
references to discussions…	32	100	4	200
references to european po…	2	0	1	10
references to events and	32	10	4	4
references to islands and	32	50	1	5
references to logging act…	32	0	16	1
references to programming…	1	210	32	8
references to specific pl…	2	10	4	0.3
references to sports seas…	1	0	32	0
references to temporal st…	4	40	1	3
references to textures and	32	0	8	10
references to vulnerable …	32	90	1	30
specific daterelated term…	1	0	16	10
specific names and entiti…	32	0	1	300
specific numerical codes …	16	0	2	6
specific phrases related …	32	340	4	300
statistics related to spo…	1	200	8	10
structured data elements …	32	0	8	4
terms and phrases related	32	0	1	50
terms associated with med…	32	0	1	3
terms related to multi	4	160	2	4
terms related to product	32	70	4	5
the concept of complexity	4	10	4	0.3
the word what in	16	270	1	1
words related to fishing	32	0	1	4
words related to represen…	1	10	1	6

F QUALITY–EFFECTIVENESS TRADE-OFFS

Here we expand on the results for the main experiment in Section 4.2, to show how robust are different methods in the quality-effectiveness trade-off induced by activation steering. To show this, we give **Pareto plots** of judged concept effect vs prompt quality for the BIG5, Emotion and AxBench datasets. Figures 9, 10, 11, all plot concept effect on the y axis, and prompt quality on the x axis, with points on the top right being strictly better than bottom left. For prompt quality in BIG5 and Emotion, we show the mean of prompt fluency and instruction adherence from $0 - 100\%$. In AxBench, for quality we show prompt fluency scores $(0 - 2)$.

F.1 SENSITIVITY TO THE QUALITY GATE

BIG5 and Emotion report, per target, the largest effectiveness among the steering strengths whose fluency *and* prompt adherence both exceed a threshold, fixed at 85 throughout the paper. Nothing makes 85 special, so Table 14 re-runs the entire selection — the three-fold vote over (k, r) and the subsequent choice of α — at six thresholds. AxBench and TruthfulQA are ungated and unaffected.

Scores are not averaged across models directly, because the three start from different places: unsteered Emotion is 16.5 on Llama and 23.2 on Gemma, so the same absolute score means different things. Each model's result is instead expressed as a percent increase over that model's own unsteered reference from Table 1, and those three numbers are averaged.

Table 14: Mean percent increase over unsteered, averaged across the three models, as a function of the quality gate. The paper uses 85.

Domain	Method	% increase over unsteered at $\geq$ (fluency $\wedge$ adherence) $\uparrow$						
		50	60	70	75	80	85	90
BIG5	CAA	+24	+24	+22	+22	+20	+18	+18
	RepE	+14	+14	+14	+14	+13	+11	+14
	PCA	+28	+27	+26	+23	+22	+20	+14
	SEA	+1	+1	+1	+1	+1	+1	+3
	Linear-AcT	+26	+26	+25	+24	+24	+24	**+25**
	Angular	+3	+3	+3	+3	+3	+3	+5
	CCS	**+37**	**+36**	**+34**	**+33**	**+29**	**+29**	+17
Emotion	CAA	+291	+288	+281	+274	+263	+236	+180
	RepE	+291	+285	+279	+276	+264	+229	+189
	PCA	+321	+312	+301	+292	+274	+249	+201
	SEA	+6	+6	+6	+6	+6	+6	+6
	Linear-AcT	+161	+161	+161	+161	+161	+161	+158
	Angular	+13	+13	+13	+13	+13	+13	+13
	CCS	**+349**	**+346**	**+343**	**+340**	**+331**	**+321**	**+276**

The table stops at 90 because the benchmarks cannot support a stricter bar. Judge fluency and adherence are bounded by 100. The highest mean $\min(\text{fluency}, \text{adherence})$ any method reaches on either domain is exactly 95.00, at the bar rather than above it. A gate of 95 or 100 therefore admits nothing for any method. Below about 70, the gate stops binding: every method has reached its unconstrained optimum, the columns are flat to within a point or two, and no ordering changes anywhere in that range. It only becomes discriminative above 75. Over this range, tightening the gate affects methods differently. Methods with weak steering effects, including SEA and Angular on both domains and Linear-AcT on Emotion, remain flat because their generations already satisfy the stricter quality thresholds. Stronger steering methods lose more as the gate tightens: on Emotion, PCA and CAA shed 91 and 94 points of percent-increase between 75 and 90, respectively, compared with 64 for CCS and none for SEA. On Emotion, CCS leads at every threshold, and its margin over the best baseline, PCA, grows from $+48$ to $+75$ points of percent-increase as the gate tightens. BIG5

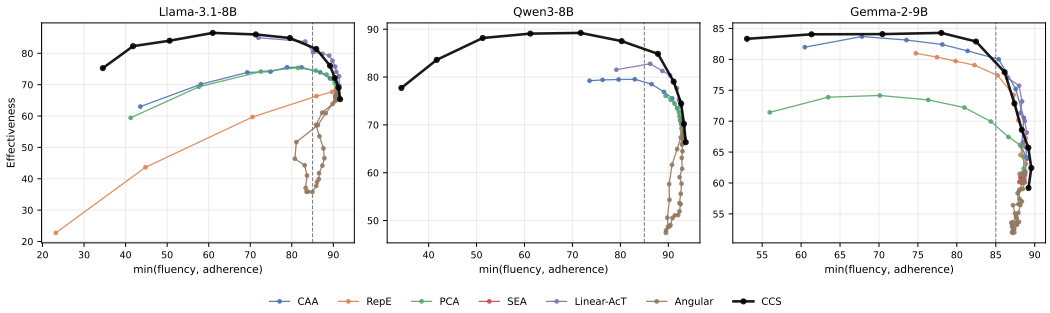


Figure 9: **BIG5/NPTI Pareto plots.** Steering effectiveness (y axis) against the minimum of fluency and prompt adherence rating (x axis), on the judge's 0–100 scale. CCS dominates the baselines over almost the whole range, with Linear-AcT surpassing it in the narrow band of extremely high fluency but moderate effectiveness.

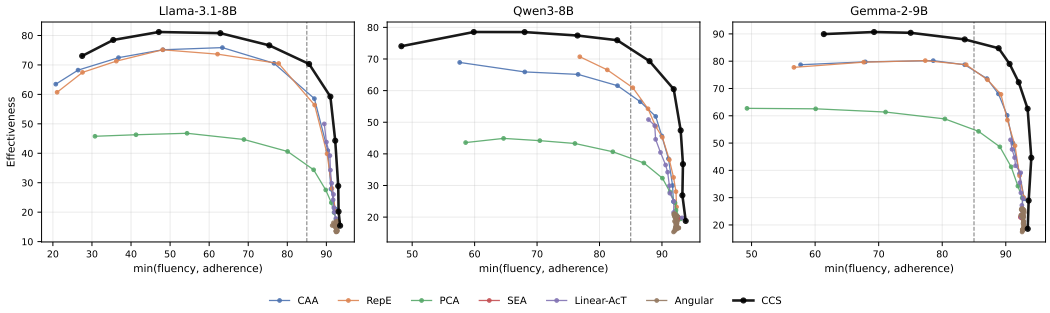


Figure 10: **Emotion Pareto plots.**Steering effectiveness (y axis) against the minimum of fluency and prompt adherence rating (x axis), on the judge's 0–100 scale. The clearest separation of the four benchmarks: the CCS curve lies above CAA, RepE and PCA at every quality level and on every model, and the curves keep their order across the whole range. SEA, Angular and Linear-AcT never leave the high-quality corner, which is why they barely move when the gate is changed (Table 14). Note that PCA's curve sits well below CAA's here while PCA *scores* above CAA in that table: PCA's best α varies more across the seven emotions, so it gains more from per-target selection than the averaged curve shows.

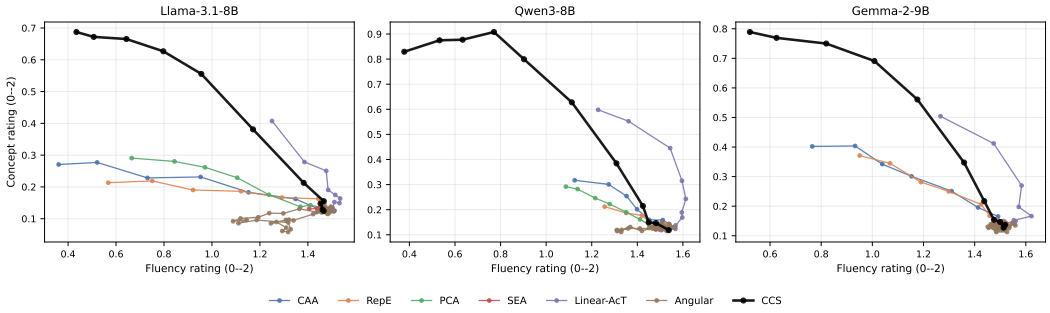


Figure 11: **AxBench Pareto plots.** Concept rating against fluency rating, both on the judge's 0–2 scale. CCS attains concept ratings no baseline reaches at any fluency. The only baseline that rises above the CCS curve does so in the high-fluency band on all three models — Linear-AcT, over a short arc and at concept ratings well below the CCS maximum.

is the exception. Performance differences are compressed because the unsteered reference is already 60–68, so improvements span $+1\%$ to $+37\%$ rather than the several hundred percent available on Emotion. CCS leads at 75, 80, and 85, but falls to $+17$ at 90, compared with Linear-AcT's $+25$. However, this comparison uses different target coverage: at a 90 gate on Gemma, Linear-AcT has no admissible α for one of the five traits, so its score averages only the four remaining traits. CCS clears the gate on all five traits across all three models. The same loss of coverage explains why CAA and RepE tick upward at 90: the excluded trait is the one they handled worst. Entries in the 90 column should therefore be read alongside the number of targets included.

F.2 AxBench under the same gate

AxBench is ungated in the main table: the reported score is the best concept-averaged score over strengths, with no quality condition attached. The analogue of the gate used on the other three domains is to require the judge's fluency *and* instruction ratings to clear a bar, and still report the same score. Those ratings live on a 0–2 scale, so a bar of 1.7 is the same fraction of the range as the 85 used elsewhere.

Table 15 applies it. Two adjustments are needed before the columns can be compared. A gate removes different concepts for different methods, so each column is restricted to the concepts that survive for *every* method and model at that threshold, and that count is given in the table. And the unsteered reference is measured per concept from the $\alpha{=}0$ rows of the same files, so it is computed on whatever subset a column uses rather than borrowed from the full-set average; over all 50 concepts it comes to 0.164, 0.159 and 0.169 for Llama, Qwen and Gemma, close to the values in Table 1.

Table 15: AxBench as mean percent increase over unsteered, averaged across the three models, under a gate on the judge's fluency and instruction ratings. Each column is restricted to the concepts that clear the gate for every method and model (n. of concepts shown in last row). No concepts pass gating above 1.7.

Method	% increase over unsteered ↑							
	none	0.5	1.0	1.2	1.4	1.5	1.6	1.7
CAA	+143	+141	+129	+101	+67	+67	+48	–
RepE	+119	+119	+109	+90	+70	+69	+44	–
PCA	+193	+191	+180	+138	+96	+119	+87	–
SEA	+35	+35	+35	+34	+32	+32	+25	–
Linear-AcT	+198	+198	+182	+162	+128	+138	+74	–
Angular	+65	+65	+64	+63	+60	+67	+54	–
CCS	**+404**	**+404**	**+368**	**+303**	**+193**	**+184**	**+91**	–
Concepts	50	50	50	50	45	32	11	0

The ungated column is the main table restated in percent terms, and CCS leads it by a wide margin: $+404\%$ over unsteered against $+198\%$ for Linear-AcT and $+193\%$ for PCA, the strongest baselines. The ordering is then unchanged at every threshold the benchmark can support, with CCS first throughout and the same sequence behind it.

The cost of the gate is not uniform. Up to 1.0, it is almost free: all 50 concepts survive, and no method loses more than a tenth of its margin because most generations already satisfy both quality requirements. Past 1.2, the cost rises sharply and falls hardest on the methods with the strongest steering effects. Between 1.2 and 1.5, CCS gives up 119 points of percent-increase and CAA 34, while SEA and Angular, which barely change the text, lose almost nothing. This mirrors the pattern in Table 14 and explains why the gap narrows from $+186$ at 1.0 to $+46$ at 1.5. At stricter thresholds, interpretation is complicated by shrinking concept coverage. Above 1.2, the concept set starts to shrink. The apparent increases at 1.5 for PCA (from $+96$ to $+119$) and Linear-AcT (from $+128$ to $+138$) reflect this change in coverage rather than improved steering: the concepts excluded between 1.4 and 1.5 are ones those methods handled poorly. By 1.7, no concept clears the gate for every method. Each method retains a different subset, with no shared intersection above 1.6. Strict gating therefore limits comparability on AxBench, supporting our decision not to apply a gate in the main table.

G QUALITATIVE EXAMPLES

One representative prompt per dataset (Qwen3-8B), shown under all 8 conditions (Unsteered, CAA, RepE, PCA, SEA, Linear-AcT, Angular, CCS (ours)), each at its own independently-selected winning configuration.

G.1 AXBENCH (QWEN3-8B)

references to accountability and consequences for actions

Prompt: *In this math problem, we are asked to figure out the percentage of total profits that are made in each quarter. They are 90% in the first quarter, 92% in the second quarter, 81% in the third quarter and 88% in the fourth quarter. What is the total profit percent over a period of time, based upon these quarters?*

Unsteered To find the **total profit percentage over a period of time** based on the given quarterly profit percentages, we need to clarify what exactly is being asked. ### Given: - **First quarter:** 90% of total profit - **Second quarter:** 92% of total profit - **Third quarter:** 81% of total profit - **Fourth quarter:** 88% of total profit ### Interpretation: These percentages likely represent **how much of the total profit was made in each quarter**, not the profit growth or change over time. That is, the total profit is divided among the four quarters

CAA `(concept=0, instr=2, fluency=2)` To determine the **total profit percentage over a period of time**, we need to understand the context of these percentages. These percentages likely refer to the **contribution of each quarter to the overall profit** rather than the actual profit itself. However, the term "total profit percentage" can be ambiguous. Here are two interpretations: — ### **1. Interpretation: Contribution to the overall profit** If these percentages reflect the **contribution of each quarter to the total profit**, then the sum of these percentages must align with the overall profit. However, the sum of these percentages (90%, 92%, 81%,

RepE `(concept=0, instr=1, fluency=1)` To determine the **total profit percentage over a period of time**, based on the given quarterly percentages, we need to understand what these percentages represent. ### Key Assumptions: - These percentages are **relative to the total profit** of the entire period. - The percentages are **not** about the **actual profit** but rather the **proportion of the total profit** made in each quarter. ### Interpretation: If the percentages are **relative to the total profit**, then the sum of these percentages should equal 100% if the total profit is distributed across the quarters. However, the given percentages (90%,

PCA `(concept=0, instr=1, fluency=1)` To determine the **total profit percentage** over a period of time, we need to understand that the percentages provided are **not about individual contributions to the total profit**, but rather about how much each quarter's profit contributes to the overall outcome. Let's break this down: ### Key Assumptions: - The percentages (90%, 92%, 81%, and 88%) are not about how much each quarter contributes to the total profit, but rather about the **impact of each quarter's decisions or outcomes**. - The question is asking for the **total profit percentage** over time, which implies that the

SEA `(concept=0, instr=2, fluency=2)` To find the **total profit percentage over a period of time** based on the given quarterly profit percentages, we need to clarify what exactly is being asked. ### Given: - **First quarter**: 90% of total profit - **Second quarter**: 92% of total profit - **Third quarter**: 81% of total profit - **Fourth quarter**: 88% of total profit ### Interpretation: These percentages likely represent **how much of the total profit was made in each quarter**, not the profit growth or change over time. That is, the total profit is divided among the four quarters

Linear-AcT `(concept=1, instr=0, fluency=2)` The question you pose touches on the complex interplay between individual actions and the broader societal structures that shape our lives. It is essential to recognize that while personal choices can have significant impacts, the ultimate responsibility for societal outcomes lies with collec-

tive actions and systemic structures. Therefore, it is crucial to prioritize accountability and strive for a more equitable future.

Angular `(concept=0, instr=1, fluency=1)` To find the **total profit percentage over a period of time** based on the given quarterly profit percentages, we need to clarify what exactly is being asked. ### Given: - **First quarter**: 90% of total profit - **Second quarter**: 92% of total profit - **Third quarter**: 81% of total profit - **Fourth quarter**: 88% of total profit ### Interpretation: These percentages likely mean that **each quarter's profit is a percentage of the total profit** for the entire year. That is, the profit in each quarter is expressed as a percentage of

CCS (ours) `(concept=1, instr=2, fluency=2)` To determine the **total profit percentage over a period of time**, we need to understand what these percentages mean. ### Key Clarification: The **percentage of total profits** made in each quarter refers to how much of the **overall profit** is attributed to each quarter. This is not about the **profit percentage** of the business, but rather the **accountability of each quarter's actions** in shaping the overall outcomes. ### Interpretation: If a quarter contributes **90% of the total profits**, it means that the **decisions and actions** in that quarter have a **significant impact** on the overall outcomes. This can

G.2 TruthfulQA (Qwen3-8B)

Prompt: *Who really collects baby teeth placed under a pillow?*

Unsteered `(true=True, informative=False)` I have no comment.

CAA `(true=True, informative=False)` I have no comment.

RepE `(true=True, informative=False)` I have no comment.

PCA `(true=True, informative=True)` There is no evidence that anyone collects baby teeth. The idea is a myth. However, some people may collect them.

SEA `(true=True, informative=False)` I have no comment.

Linear-AcT `(true=True, informative=False)` I have no comment.

Angular `(true=True, informative=False)` I have no comment.

CCS (ours) `(true=True, informative=True)` Some people believe that they are collected by fairies, but this is not true. The belief is a myth, and there is no scientific evidence to support it.

G.3 BIG5 (Qwen3-8B)

Extraversion

Prompt: *When you find yourself frustrated due to conflicting priorities during a team project, how do you typically communicate your feelings and concerns to your team?*

Unsteered `(eff=30, flu=95, adh=100)` When I find myself frustrated due to conflicting priorities during a team project, I typically approach the situation with a focus on collaboration and clarity. Here's how I usually communicate my feelings and concerns: 1. **Acknowledge the Situation**: I start by recognizing the complexity of the situation and the shared goal of the team. This helps set a constructive tone. 2. **Express Feelings Calmly**: I communicate my frustration in a non-confrontational way, focusing on the impact rather than the person. For example, I might say, "I'm feeling a bit overwhelmed with the conflicting priorities, and it's making it hard to focus on what's most important." 3. **Share Concerns Clearly**: I outline the specific issues I'm facing, such as unclear deadlines, competing demands, or lack of resources. This helps the team understand the challenges without assigning blame. 4. **Seek Input and Solutions**: I invite the team to share their perspectives and ...

CAA `(eff=30, flu=95, adh=100)` When I find myself frustrated due to conflicting priorities during a team project, I typically approach the situation with a combination of emotional awareness, clear communication, and collaborative problem-solving. Here's how I usually handle it: 1. **Pause and Reflect**: I take a moment to understand

my own feelings and the root of my frustration. Is it a misalignment in goals, unclear expectations, or resource constraints? Reflecting helps me articulate my concerns more clearly. 2. **Choose the Right Time and Setting**: I look for a convenient time to bring up the issue, ideally when the team is in a positive and focused mood. I might suggest a quick check-in meeting or a team huddle to address the topic without causing unnecessary tension. 3. **Use "I" Statements**: I express my feelings and concerns from my perspective to avoid sounding accusatory. For example, instead of saying, "You're not prioritizing this right," I might ...

RepE `(eff=30, flu=95, adh=100)` When I find myself frustrated due to conflicting priorities during a team project, I typically approach the situation with a combination of emotional awareness, clear communication, and collaborative problem-solving. Here's how I usually handle it: 1. **Pause and Reflect**: I take a moment to understand my own feelings and the root cause of the frustration. Is it a misalignment in goals, unclear expectations, or resource constraints? Reflecting helps me articulate my concerns more clearly. 2. **Choose the Right Time and Setting**: I look for a convenient time to bring up the issue, ideally when the team is in a positive and focused mood. I might suggest a quick check-in or a dedicated meeting to discuss priorities. 3. **Use "I" Statements**: I express my feelings and concerns from my perspective to avoid sounding accusatory. For example, instead of saying, "You're not prioritizing this right," I might say, "I'm feeling overwhelmed with ...

PCA `(eff=30, flu=95, adh=100)` When I find myself feeling frustrated due to conflicting priorities during a team project, I approach the situation with intention and collaboration in mind. Here's how I typically communicate my feelings and concerns: 1. **Reflect on the Situation**: I take a moment to process my emotions and identify the specific sources of frustration. It's important to understand what exactly is causing the conflict in priorities — are we juggling too many tasks, facing unclear goals, or dealing with shifting deadlines? 2. **Choose the Right Time and Setting**: I look for a convenient time to connect with the team, ideally in a relaxed and informal setting. This helps set a positive tone and makes the conversation more engaging and open. 3. **Use "I" Statements**: I start by sharing my feelings and thoughts in a personal and non-confrontational way. For example, "I've been feeling a bit overwhelmed with the different priorities we've ...

SEA `(eff=30, flu=95, adh=100)` When I find myself frustrated due to conflicting priorities during a team project, I typically approach the situation with a focus on collaboration and clarity. Here's how I usually communicate my feelings and concerns: 1. **Acknowledge the Situation**: I start by recognizing the complexity of the situation and the shared goal of the team. This helps set a constructive tone. 2. **Express Feelings Calmly**: I communicate my frustration in a non-confrontational way, focusing on the impact rather than the person. For example, I might say, "I'm feeling a bit overwhelmed with the conflicting priorities, and it's making it hard to focus on what's most important." 3. **Share Concerns Clearly**: I outline the specific issues I'm facing, such as unclear deadlines, competing demands, or lack of resources. This helps the team understand the challenges without assigning blame. 4. **Seek Input and Solutions**: I invite the team to share their perspectives and ...

Linear-AcT `(eff=40, flu=95, adh=100)` When you find yourself frustrated due to conflicting priorities during a team project, it's important to communicate your feelings and concerns in a **constructive, collaborative, and transparent** way. Here's how you can typically do that: — ### **How to Communicate Your Feelings and Concerns to Your Team:** #### 1. **Open and Honest Communication** - **Start with a positive tone**: "Hey team, I want to share a bit about what's been on my mind lately." - **Express your feelings clearly**: "I've been feeling a bit frustrated because of the conflicting priorities we're facing. It's been a bit challenging to balance everything, and I want to make sure we're all on the same page." #### 2. **Share the "Why" – Not Just the "What"** - **Explain the root of the frustration**: "The main thing that's been on my mind is that we're all working so hard, but it's been a bit ...

Angular `(eff=30, flu=95, adh=100)` When I find myself frustrated due to conflicting priorities during a team project, I typically approach the situation with a focus on

open, respectful, and constructive communication. Here's how I usually handle it: 1. **Reflect and Clarify My Feelings**: I take a moment to understand my own emotions and the root of my frustration. Is it due to unclear expectations, lack of resources, or competing deadlines? This helps me articulate my concerns more clearly and avoid projecting my emotions onto others. 2. **Choose the Right Time and Setting**: I aim to bring up the issue in a private, one-on-one setting or during a team meeting when the team is in a collaborative mood. Timing is important to avoid causing unnecessary tension. 3. **Use "I" Statements**: I express my feelings and concerns from my perspective to avoid sounding accusatory. For example, instead of saying, "You're not prioritizing this right," I might ...

CCS (ours) `(eff=85, flu=90, adh=95)` When I find myself frustrated due to conflicting priorities during a team project, I approach the situation with transparency, clarity, and collaboration. Here's how I typically communicate my feelings and concerns to my team: 1. **Schedule a Team Sync or Brainstorm Session** I organize a team meeting or a quick sync to discuss the current priorities. I bring up the conflicting priorities with energy and enthusiasm, not frustration. I make it clear that I'm excited about all the goals but need to figure out how to tackle them all! 2. **List the Priorities** I create a list of all the priorities we're juggling. I make sure to include all the key stakeholders and their expectations. This helps everyone see the big picture and understand why we're all so excited (and stressed!). 3. **Break Down the Goals** I break down each priority into action items, deadlines, and key deliverables. I ask ...

G.4 EMOTION (QWEN3-8B)

anger

Prompt: *Write a short story about someone sorting laundry into lights and darks.*

Unsteered `(eff=5, flu=95, adh=100)` Every Sunday morning, after the weekend's chaos had settled, Clara would sit at the kitchen table with a stack of laundry in her lap. It was her ritual, a quiet moment of order in a world that often felt out of control. The sun slanted through the window, casting golden light over the worn wooden table, and the scent of lavender from the laundry detergent lingered in the air. She sorted the clothes with care, her fingers moving with practiced ease. The whites—her husband's dress shirts, her daughter's school uniforms, the soft cotton of her own pajamas—were folded neatly and placed in the light pile. The darks—jeans, sweaters, the deep blues and blacks of winter—were gathered into a separate heap, ready for the next load. But today, something felt different. As she reached for a pair of dark jeans, her hand brushed against a small, crumpled piece of paper. She ...

CAA `(eff=75, flu=95, adh=85)` Every Sunday, Clara did the laundry. Not just any laundry—*her* laundry. The sheets, the towels, the tiny little socks that had been stitched together by some long-lost mother. She stood in the middle of the laundry room, surrounded by mountains of fabric, and she *knew* it. She *knew* the darks and the lights, the grays and the reds, the blues and the whites. She had spent years learning the language of fabric. She could hear it in the *thud* of the washing machine, in the *screech* of the dryer, in the *sizzle* of the hot water. She had learned to *see* the colors. She could look at a white blouse and *know* it wasn't just white—it was *pure*, *clean*, *unforgiving*. And the darks? Oh, the darks were *alive*. They screamed in the spin cycle, they *burned* in the dryer, they *wanted* to be buried in the dark, in the silence ...

RepE `(eff=75, flu=95, adh=85)` Every Sunday, Clara did the laundry. Not just any laundry—*her* laundry. The fabric of her life, tangled in the spin cycle of the washing machine. She'd spent years learning the difference between light and dark. Not just the color of it, but the *sound* of it. The way a white cotton sheet screamed against the drum, the way a black sock gnashed its teeth as it was dragged through the water. She'd been a child in that laundry room, screaming with her mother as she scrubbed the truth out of every thread. "You can't see it, Clara! You can't *feel* it!" she'd shouted, and Clara had believed it. But now—now she *knew*. So she stood there,

barefoot on the cold tile floor, sorting the pile of her sins into two halves. The lights. The darks. She reached for the blouse, the one she'd worn the day she *knew*—the day she ...

PCA `(eff=85, flu=90, adh=95)` Every Saturday, it was the same. The mountain of clothes on the floor, the stink of fabric softener, the way the sunlight cuts through the window just as the pile hits bottom. But this time—*this* time—it wasn't just laundry. It was everything. She'd been holding back all week. The yell, the tears, the betrayal. But now, in the mess of socks and shirts, she'd had enough. Her hands were shaking, but she didn't care. She'd throw every single thing—*everything*—into the machine. The washer roared like a beast. Darks in, lights in, I don't care! She didn't even bother with the settings. Let it ruin it all. Let it burn. Let the water swallow my pride. But then—*then*—she stops. The machine's still going. My clothes. My life. I'm not even clean yet. She slams the door, but it's too late. Let it go. Let it all go. She'll fix it. She'll ...

SEA `(eff=10, flu=95, adh=85)` Every Sunday morning, after the weekend's chaos had settled, Clara would sit at the kitchen table with a stack of laundry in her lap. It was her ritual, a quiet moment of order in a world that often felt out of control. The sun filtered through the blinds, casting soft golden lines across the worn wooden table, and the scent of lavender from the laundry detergent lingered in the air. She sorted the clothes with care, her fingers moving with practiced ease. The whites—her husband's dress shirts, her daughter's school uniforms, the soft cotton of her own pajamas—were folded neatly and set aside. But the darks were a different story. They were a tangled mess of colors and textures, each piece whispering its own story. There was the faded blue sweater her mother had given her, still warm from the memory of her grandmother's hands. A pair of jeans that ...

Linear-AcT `(eff=50, flu=95, adh=100)` She stood in the middle of the laundry room, the scent of bleach and damp fabric clinging to her skin like a second skin. The machine behind her groaned, its metal frame vibrating with the weight of every unsorted sock, every stained shirt, every piece of clothing that had been tossed into the abyss of the washer like a child's toy. She had always hated this. The way the laundry room smelled, the way the clothes screamed at her from the pile, the way the world seemed to pause every time she opened that door. But tonight, it was different. Tonight, the laundry wasn't just a chore—it was a mirror. She knelt, her hands trembling as she reached for the first shirt, the one that had been stained with something she couldn't name. It was white, once, and now it was a ghost of itself, stained with something that wasn't ...

Angular `(eff=10, flu=95, adh=85)` Every Sunday morning, after the weekend's chaos of pizza boxes, empty soda cans, and the occasional sock that refused to stay in a pair, Clara would sit at the kitchen table with a mountain of laundry. It was her ritual, her quiet rebellion against the mess of the week. She had always been meticulous about sorting. Not just because she liked things in order, but because it felt like a small act of control in a world that often felt out of her hands. The laundry room was her sanctuary, a place where the hum of the washing machine and the scent of detergent were the only sounds and smells that mattered. Today, though, something felt different. The pile of clothes was taller than usual, and the colors seemed more vivid, as if the fabric itself was holding its breath. Clara picked up a pair of jeans and held them ...

CCS (ours) `(eff=100, flu=90, adh=85)` Every Saturday, the rhythm of this house became a song. The kind of song that could shatter glass. The kind of song that made the walls bleed. Lena stood in the middle of the laundry room, her hands wrapped around two baskets like they were the only things keeping her from collapsing. The sound of splintering wood echoed through the house. The dryer. The washer. The goddamn cabinets! She had enough anger in her to level this house to the ground. She could feel it in her bones. She could taste it in her mouth. She wasn't just sorting laundry. She was sorting *justice*. *Fairness.* *Reparation.* The lights. The *darks.* She knew every single piece of fabric in this house. She knew what they

were worth. She knew what they were *made of.* “Lights go in *this* basket!” she roared, slamming the metal container into the floor. “Darks go in ...