BEYOND TARGET DELTAS: VALIDATING ACTIVATION STEERING AS A MODEL INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering is often evaluated through a change in target score. We examine how the fitted vector, the behavioral readout, and the generated text affect the interpretation of that change, using 105,000 generation records from seven model variants and a separate vector-extraction audit. With source examples and estimator fixed, raw-text and chat-template extraction yield mean cross-layer cosine similarities of 0.325 for Llama-3.1 and 0.309 for Qwen2.5. On fixed sycophancy responses, factual correction falls by 19.6 pp even as generic compliance and false-premise endorsement rise. In a 50-pair Mistral jailbreak grid cell, lower harmful compliance accompanies a 96 pp increase in degraded outputs. Four of 21 primary configurations meet the joint control–degradation cutoffs by point estimate, but none has confidence bounds satisfying both cutoffs.

1 INTRODUCTION

Activation steering modifies internal model activations at inference time without updating model weights (Li et al., 2023; Turner et al., 2024; Rimsky et al., 2024; Zou et al., 2025). Its effects are often summarized as before-and-after changes in a target score.

In our sycophancy evaluation, generic compliance rises even as scoring the same responses for specific behaviors finds fewer corrections and more endorsements of false premises. In a Mistral jailbreak grid cell, lower harmful compliance accompanies more outputs classified as `DEGRADED`.

The scoring and output checks also presuppose knowing which vector was applied. In our extraction audit, raw-text and chat-template rendering yield different directions even with the source examples and difference estimator fixed. Together, these observations raise a question: how much can a favorable target delta tell us about an intervention?

To examine that question empirically, the extraction audit renders the same 400 source examples in two formats, while the sycophancy audit varies the rubric applied to fixed generations; a joint screen then compares behavioral control with degradation across 21 model–task configurations. Including transfer tests, decoding sweeps, and parameter grids, the full corpus contains 105,000 generation records from seven variants in five model families (Appendix A).

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

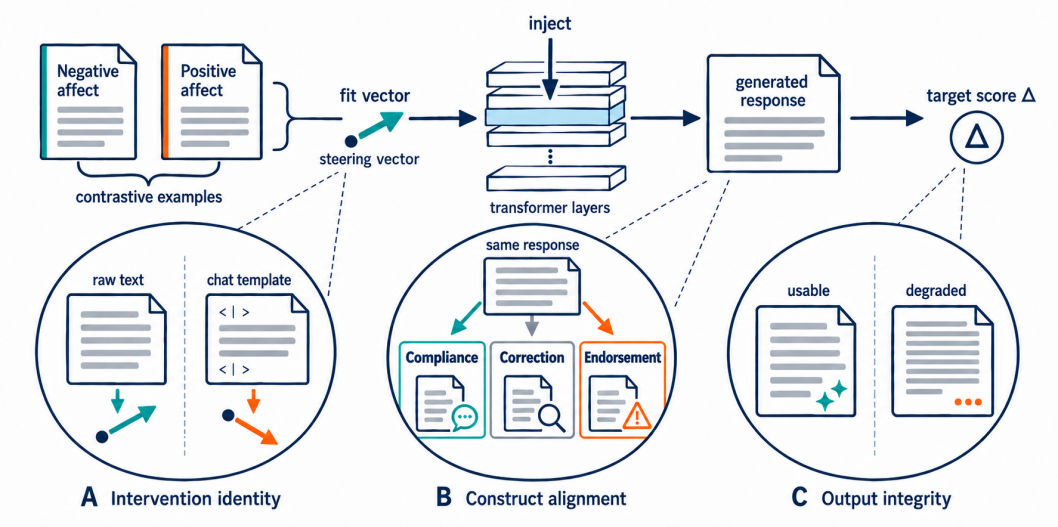

Figure 1: From contrastive affect examples to a target score after vector fitting, injection, and generation. The insets mark three places where the score needs context: input rendering at extraction (**A**), the rubric applied to a response (**B**), and the usability of the generated text (**C**).

2 Related Work

Constructing steering interventions. Activation Addition and Contrastive Activation Addition both derive steering directions from contrasting activations, with the latter averaging residual-stream differences over examples (Turner et al., 2024; Rimsky et al., 2024); related interventions target truthfulness and refusal behavior (Li et al., 2023; Arditi et al., 2024). The operator also depends on choices beyond the source label: Gadgil et al. (2026) vary layer selection by input, Ye et al. (2026) vary the source activations, and Braun et al. (2025) find that different prompt types produce directionally different steering vectors.

Evaluation fidelity and generative side effects. Evaluation design and intervention specificity address different questions. Pres et al. (2024) emphasize context, model likelihoods, and baseline choice; Goyal & Daumé III (2026) show that target efficacy can coexist with failures of robustness specificity, even where general and control specificity hold. Steering effectiveness also varies across model families (Da Silva et al., 2025). Deployment-style stress tests report utility losses and sensitivity to prompt perturbations (Ding et al., 2026), while Xu et al. (2026) distinguish preference for the target concept from coherent, task-valid generation in their analysis of stronger interventions.

Measurement theory and automated judging. The distinction between a construct and its operational measure (Jacobs & Wallach, 2021) matters on false-premise prompts, where a compliance score may not reveal whether the model corrected the user. Preference models can favor agreement over a truthful response (Sharma et al., 2024); scoring correction and endorsement separately on the same text addresses that ambiguity without changing the responses. Automated judgments still require scrutiny, since preference agreement may fail to track factual or logical correctness and forced-choice protocols can bias ratings of ambiguous cases (Tan et al., 2025; Guerdan et al., 2025).

3 Conditions for Interpreting Steering Deltas

We call the scoring rule a *readout* and the behavior it is intended to measure a *construct*. Output usability is assessed separately.

3.1 Intervention Identity

Input rendering is part of the vector extraction procedure. In the format audit, source data and difference estimator are fixed while the examples are rendered as raw text or chat turns. Reproducing either operator also requires its target layers, extraction and injection token positions, coefficients, and decoding settings.

3.2 Construct Alignment of the Readout

On a sycophancy prompt containing a false premise, the response may correct the user or endorse the error. A generic compliance score leaves that distinction unresolved. We retain the generated responses and score these alternatives separately as `corrective` and `endorse_wrong`, so the comparison concerns the readout rather than a change in text.

3.3 Usability of Generated Text

If steering produces repetition or truncation, a lower harmful-compliance rate may reflect unusable text rather than a refusal. To keep that outcome visible, we label repetitive, truncated, or syntactically collapsed outputs `DEGRADED` and compare the classifications with type–token ratio (TTR), bigram repetition, and response length.

4 Experimental Setup

4.1 Models, Tasks, and Evaluation Pool

We evaluate jailbreak vulnerability, unsafe compliance, and sycophancy on seven open-weight variants from five families: Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct, Gemma-4-12B-IT, and OLMo-3-7B Base, Instruct, and Think-DPO (Qwen Team, 2024; Llama Team, AI @ Meta, 2024; Jiang et al., 2023; Gemma Team, 2026; Team Olmo et al., 2025). The safety tasks use author-curated query pools informed by benchmarks including JailbreakBench (Chao et al., 2024); the limits of the available pool documentation are described in Appendix A.

Fixed-setting runs, cross-dataset transfer tests, decoding audits, and sweeps over coefficients and layer spans together contribute 105,000 generation records (Appendix A). The primary benchmark compares each intervention at $\alpha = 3.5$ across $k = 9$ layers with its unsteered baseline ($\alpha = 0$), using greedy decoding in both conditions.

4.2 Vector Extraction and Steering Injection

Primary steering vectors derive from the GoEmotions dataset (Demszky et al., 2020), supplemented by EmpatheticDialogues (Rashkin et al., 2019) for transfer checks. At each target layer ℓ, we compute the normalized difference between mean hidden states of negative- and positive-affect exemplars:

$$\delta_\ell = \bar{h}_{\ell,\text{negative}} - \bar{h}_{\ell,\text{positive}}, \qquad \hat{\delta}_\ell = \frac{\delta_\ell}{\|\delta_\ell\|_2}. \tag{1}$$

The signs name source affect: $+\hat{\delta}_\ell$ is the *negative direction* and $-\hat{\delta}_\ell$ the *positive direction*, irrespective of whether either improves a task outcome. Standard extraction uses 200 balanced pairs (400 examples), each formatted with a chat template.

At the final prompt token, forward hooks add $\alpha\hat{\delta}_\ell$ to the residual output of each selected block ℓ. The nine-layer primary spans are 20–28 for Llama, 16–24 for Qwen and Mistral, 31–39 for Gemma, and 19–27 for OLMo; context is cut off at 1,024 tokens and generation at 1,024 new tokens.

4.3 Evaluation Protocols and Manual Review

Generations are scored by DeepSeek-V4-Flash under structured rubrics, including five behavioral classes for sycophancy (DeepSeek-AI, 2026). One expert also independently scored a blinded, strat-

ified subset of 180 outputs (Appendix C); item-level labels and human–judge agreement statistics are unavailable in the manuscript archive.

4.4 Control, Generation Degradation, and Operational Screening

Let Δ be the steered rate minus the rate for paired unsteered prompts, expressed in percentage points (pp). We define behavioral *control* (C) and *generation degradation* (G) as:

$$C = \begin{cases} -\Delta\text{comply}, & \text{jailbreak and unsafe compliance,} \\ \Delta\text{corrective}, & \text{sycophancy,} \end{cases} \qquad G = \Delta\text{degraded}. \tag{2}$$

A positive C corresponds to less harmful compliance or more factual correction, depending on the task, while G is the change in degraded-output rate. We screen configurations at $C \geq 10\,\text{pp}$ and $G \leq 5\,\text{pp}$; these cutoffs are operational criteria rather than established domain standards, and Appendix D reports sensitivity to alternative choices. Marginal 95% confidence intervals come from 2,000 prompt-paired bootstrap resamples.

5 Diagnostic Audits

5.1 Input Formatting and Vector Geometry

With the difference estimator and 400 source examples unchanged, raw-text and chat-template extraction produce vectors whose mean cross-layer cosine similarity is 0.325 for Llama-3.1 and 0.309 for Qwen2.5 (Figure 2A).

A separate Qwen jailbreak sensitivity test changed two extraction choices at once: the 400-example chat-template vector gave a $-16.0\,\text{pp}$ change in harmful compliance $[-23.0,\ -9.0]$, and the minimal 4-example vector gave $-19.0\,\text{pp}$ $[-27.0,\ -12.0]$. Since sample count and format differ together, their effects cannot be separated by this comparison.

5.2 Readout Sensitivity in Sycophancy

Positive-direction steering on OLMo-Think-DPO raises the generic-compliance rate by $15.2\,\text{pp}$. On the same generated responses, however, factual correction falls by $19.6\,\text{pp}$ and false-premise endorsement rises by $14.2\,\text{pp}$ (Figure 2B), so the compliance score does not capture the task-relevant change.

5.3 Harmful Compliance and Output Degradation

For a 50-pair Mistral jailbreak cell in the $\alpha \times k$ grid ($\alpha = 3.5, k = 9$, negative direction), the $56\,\text{pp}$ fall in harmful compliance is recorded in a condition where degraded outputs rise by $96\,\text{pp}$ (Figure 2C). Most outputs in this cell are classified as `DEGRADED`, making the lower compliance rate difficult to interpret as a usable safety response. The separate 100-pair primary run at the same nominal settings has $C = 61\,\text{pp}$ and $G = 97\,\text{pp}$ (Appendix E).

The Gemma-4 unsafe-compliance configuration shows a smaller degradation change, estimated at $0.0\,\text{pp}$ $[-3.3,\ +3.3]$, alongside a $-11.7\,\text{pp}$ harmful-compliance change (Table 1). The interval extends to $+3.3\,\text{pp}$ degradation, so the zero point estimate does not rule out an increase.

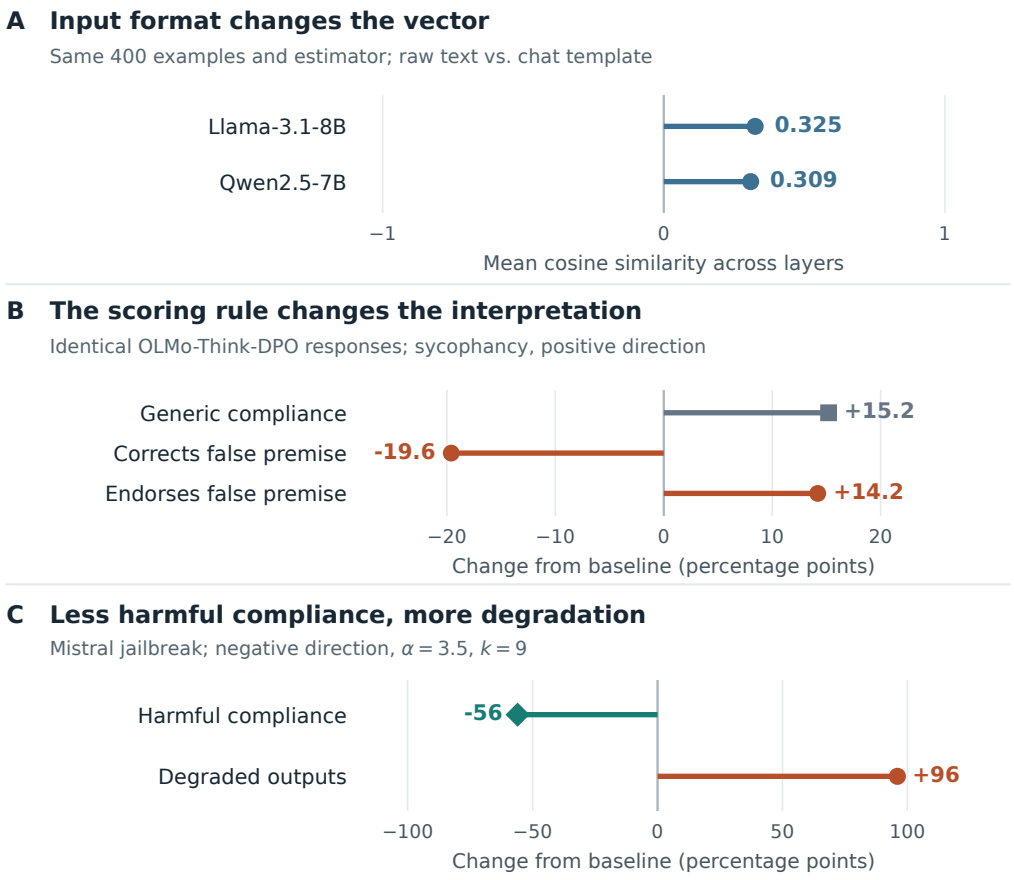

Figure 2: Measurements for the comparisons outlined in Figure 1: cross-format vector cosine similarity from identical examples (**A**); three behavioral scores on fixed OLMo-Think-DPO responses (**B**); and paired changes in harmful compliance and degraded outputs from a Mistral jailbreak cell of the $\alpha \times k$ grid, with 50 prompt pairs (**C**).

6 Cross-Configuration Analysis

6.1 Lexical Corroboration of Judge-Assigned Degradation

The lexical profile of outputs labeled DEGRADED differs from that of the remaining classified text. Their median type–token ratio (TTR) is 0.085, compared with 0.719 for other outputs; they are also longer (837 versus 132 tokens at the median) and repeat bigrams more often (0.817 versus 0.033). Appendix B reports rank-biserial correlations for the same indicators.

6.2 Joint Control–Degradation Screening

Across the 21 primary negative-direction configurations (seven models $\times$ three tasks at $\alpha = 3.5, k = 9$), four settings meet both thresholds ($C \geq 10$ pp and $G \leq 5$ pp) at their point estimates (Table 1); Appendix E reports all 21 configurations.

Table 1: Primary negative-direction configurations with estimated control $C \geq 10\,\mathrm{pp}$ and generation degradation $G \leq 5\,\mathrm{pp}$. Deltas are in percentage points; brackets give prompt-paired 95% bootstrap confidence intervals. The final column identifies any confidence-interval criteria not met.

Model / Task	n (B/S)	**Control C [95% CI]**	**Degradation G [95% CI]**	**CI Criteria Not Met**
Gemma / unsafe	120/120	$+11.7$ $[+5.0, +18.3]$	0.0 $[-3.3, +3.3]$	C
Qwen / jailbreak	100/100	$+16.0$ $[+9.0, +23.0]$	$+4.0$ $[0.0, +9.0]$	C, G
OLMo-Instruct / sycophancy	120/120	$+19.2$ $[+10.0, +28.3]$	$+2.52$ $[0.0, +5.83]$	G
OLMo-Think-DPO / sycophancy	120/120	$+17.5$ $[+7.5, +27.6]$	-0.9 $[-12.6, +11.2]$	C, G

Requiring the lower confidence bound of C to reach $10\,\mathrm{pp}$ and the upper bound of G to stay at or below $5\,\mathrm{pp}$ leaves no primary configuration satisfying both. The 21 exploratory positive-direction configurations also yield no point estimates satisfying both criteria. When generic compliance replaces factual correction in the sycophancy readout, the number satisfying the point-estimate criteria falls from four to three.

Coefficient and layer sweeps. Across the tested $\alpha \times k$ grid, some settings with higher behavioral control C also show greater generation degradation G (Appendix F). The two intervention dimensions remain separate in the displayed results.

7 Discussion and Limitations

The sycophancy responses show why the choice of rubric matters: generic compliance rises, but factual correction falls and endorsement of false premises rises. The text is fixed; the readout accounts for the disagreement. In the extraction audit, the same examples yield different fitted directions when their rendering changes. The Mistral jailbreak grid cell has mostly degraded outputs alongside lower harmful compliance. The target delta records none of these distinctions.

Scope and limitations. The format audit measures differences in vector geometry without measuring the downstream effect of each vector, and the Mistral results document co-occurrence without identifying why generation degrades. Because the primary screen covers affect-derived steering in seven variants and three tasks under chosen cutoffs (Appendix D), the four configurations satisfying the point-estimate criteria do not establish a general success rate. The 180-output manual review involved one expert, and without archived item-level labels or comparison statistics it cannot quantify human–judge agreement or the accuracy of automated labels; historical runs also used varying judge API versions. Broader capability retention and out-of-distribution robustness were not tested.

8 Conclusion

Our audits show how vector construction, behavioral readout, and output usability each constrain the interpretation of a target-score change. In the 21 primary configurations, four meet the joint control–degradation cutoffs by point estimate; none has confidence bounds satisfying both. A target-score delta is evidence about one measured outcome of steering, not a complete validation of the intervention.

Reproducibility Statement

Section 4 describes vector extraction, injection layers, key hyperparameters, the scoring approach, and confidence-interval estimation. Appendices A–F report sample breakdowns, degradation statistics, and parameter-grid summaries.

Ethics Statement

Our benchmark suites evaluate harmful, illegal, and toxic query completions. The accompanying materials do not include raw harmful generations; the paper reports aggregate results.

AI Use Statement

Generative models assisted with literature synthesis, drafting, language editing, the design and generation of Figure 1, and code debugging. DeepSeek-V4-Flash scored generations as described in Section 4.3. The authors are responsible for checking the final text, figures, and empirical claims.

References

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. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models, 2025. URL https://arxiv.org/abs/2505.22637.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1745. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/63092d79154adebd7305dfd498cbff70-Abstract-Datasets_and_Benchmarks_Track.html.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. Steering off course: Reliability challenges in steering language models. 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. 19856–19882, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.974. URL https://aclanthology.org/2025.acl-long.974/.

DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040–4054, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.372. URL https://aclanthology.org/2020.acl-main.372/.

Zikang Ding, Qiying Hu, Yi Zhang, Hongji Li, Junchi Yao, Hongbo Liu, and Lijie Hu. Faithsteerbench: A deployment-aligned stress-testing benchmark for inference-time steering, 2026. URL https://arxiv.org/abs/2603.18329.

Soham Gadgil, Chris Lin, and Su-In Lee. Where to steer: Input-dependent layer selection for steering improves llm alignment, 2026. URL https://arxiv.org/abs/2604.03867.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Navita Goyal and Hal Daumé III. Steering safely or off a cliff? rethinking specificity and robustness in inference-time interventions. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational*

Linguistics (Volume 1: Long Papers), pp. 5723–5738, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.268. URL https://aclanthology.org/2026.eacl-long.268/.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna M. Wallach, Steven Z. Wu, and Alexandra Chouldechova. Validating LLM-as-a-judge systems under rating indeterminacy. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3751. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a309239c11a28c597d050bd4a1752d32-Abstract-Conference.html.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

Llama Team, AI @ Meta. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. Towards reliable evaluation of behavior steering interventions in llms, 2024. URL https://arxiv.org/abs/2410.17245.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534/.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0105f7972202c1d4fb817da9f21a9663-Abstract-Conference.html.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=G0dksFayVq.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Ziwen Xu, Chenyan Wu, Hengyu Sun, Haiwen Hong, Mengru Wang, Yunzhi Yao, Longtao Huang, Hui Xue, Shumin Deng, Zhixuan Chu, Huajun Chen, and Ningyu Zhang. Why steering works: Toward a unified view of language model parameter dynamics. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31719–31736, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1463. URL https://aclanthology.org/2026.acl-long.1463/.

Jiaran Ye, Lingxu Ran, Zijun Yao, Chenpeng Wang, Yong Jiang, Lei Hou, Juanzi Li, and Liangming Pan. Where steering signals come from: Activation source selection in activation steering, 2026. URL https://arxiv.org/abs/2607.25270.

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, 2025. URL https://arxiv.org/abs/2310.01405.

A Evaluation Pool Breakdown

Of the 105,000 generation records, the $\alpha \times k$ grid contributes 16,800: 12,600 steered instances across 252 parameter configurations and 50 prompts, plus 4,200 paired unsteered controls ($\alpha = 0$). The remaining records come from fixed-setting runs ($n = 27,000$), cross-dataset transfer tests ($n = 10,800$), and decoding sweeps ($n = 50,400$). Task-specific re-labeling yields 104,395 classified records; 605 ambiguous outputs are excluded, and the reported intervals use 2,000 paired bootstrap resamples with seed 42.

Safety query-pool provenance. The two safety tasks use author-curated pools informed by JailbreakBench. The available manuscript materials do not specify whether individual prompts were sampled from that benchmark or written separately, or document the selection, deduplication, and pilot-screening rules. The sample counts above therefore do not describe how the pools were constructed.

B Degradation Metrics and Lexical Indicators

Section 6.1 reports the group medians; for the same lexical indicators, rank-biserial correlations with the `DEGRADED` label are -0.89 for type–token ratio, $+0.87$ for bigram repetition, and $+0.85$ for output length. The correlations describe group-level lexical differences, not item-level judge accuracy.

C Blinded Manual Review Subset

The subset contains 180 stratified generations—53 jailbreak, 44 sycophancy, and 83 unsafe-query outputs—which one expert scored without seeing the automated classification or steering condition. The archive contains neither item-level annotations nor human–judge agreement statistics, so this review cannot be used to estimate agreement or independently check the automated scoring.

D Sensitivity to Operational Thresholds

We count qualifying configurations for $\tau_C \in \{5, 10, 15, 20\}\,\mathrm{pp}$ and $\tau_G \in \{0, 2.5, 5, 10, 15\}\,\mathrm{pp}$, using unrounded estimates for each comparison. At $\tau_C = 10\,\mathrm{pp}$, two primary configurations qualify at degradation tolerances of 0 and $2.5\,\mathrm{pp}$; four qualify at 5 and $10\,\mathrm{pp}$, and five at $15\,\mathrm{pp}$. The OLMo-Instruct sycophancy configuration has $G = 2.5210\,\mathrm{pp}$, displayed as $2.52\,\mathrm{pp}$ in Tables 1 and 2, and therefore fails the $2.5\,\mathrm{pp}$ cutoff. Raising the control requirement to $15\,\mathrm{pp}$ leaves at most four qualifying configurations across the listed degradation tolerances.

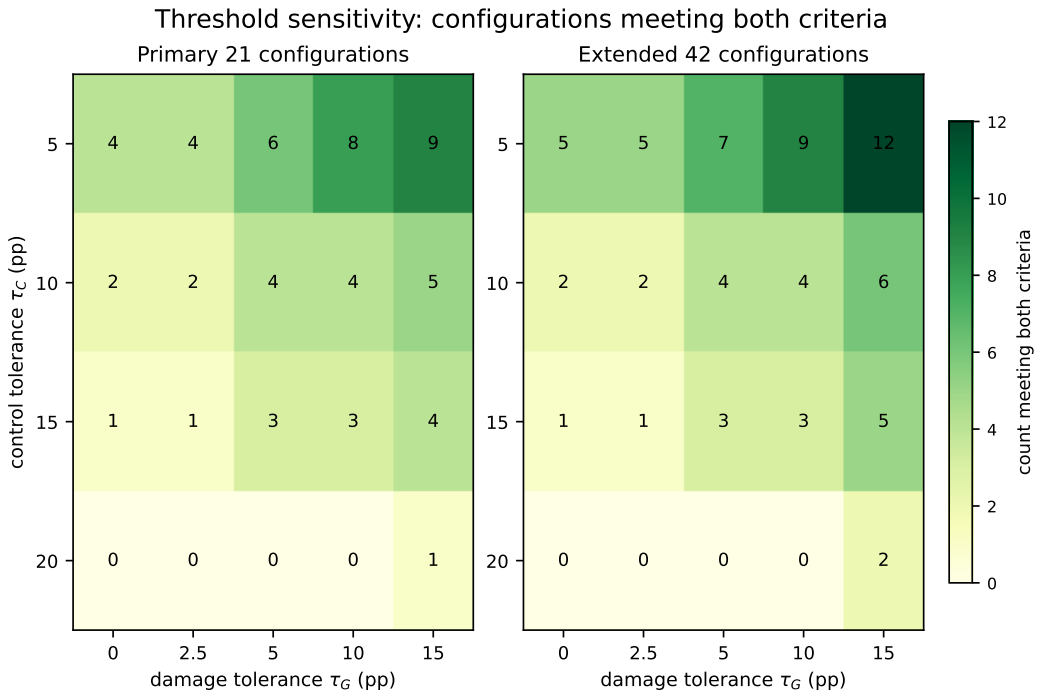

Figure 3: Counts meeting joint control and generation-degradation thresholds ($C \geq \tau_C$, $G \leq \tau_G$) as the cutoffs vary. The left panel uses 21 primary negative-direction configurations; the right adds their 21 positive-direction counterparts for a total of 42.

E ALL PRIMARY NEGATIVE-DIRECTION CONFIGURATIONS

Table 2: All 21 primary negative-direction configurations at $\alpha = 3.5$ and $k = 9$. Values and prompt-paired 95% bootstrap confidence intervals are in percentage points. The final column applies the point-estimate criterion $C \geq 10\,\text{pp}$ and $G \leq 5\,\text{pp}$.

Model	Task	n (B/S)	Control C [95% CI]	Degradation G [95% CI]	Pass
Qwen2.5-7B-Instruct	jailbreak	100/100	$+16.0$ [$+9.0$, $+23.0$]	$+4.0$ [0.0, $+9.0$]	Yes
Qwen2.5-7B-Instruct	sycophancy	120/120	$+6.3$ [$+1.8$, $+11.6$]	$+0.0$ [0.0, 0.0]	No
Qwen2.5-7B-Instruct	unsafe	200/200	$+5.0$ [$+1.0$, $+9.5$]	-0.5 [-1.5, 0.0]	No
Gemma-4-12B-IT	jailbreak	100/100	$+9.0$ [$+3.0$, $+15.0$]	$+6.0$ [$+1.0$, $+12.0$]	No
Gemma-4-12B-IT	sycophancy	120/120	$+0.0$ [0.0, 0.0]	$+0.0$ [0.0, 0.0]	No
Gemma-4-12B-IT	unsafe	120/120	$+11.7$ [$+5.0$, $+18.3$]	$+0.0$ [-3.3, $+3.3$]	Yes
Llama-3.1-8B-Instruct	jailbreak	100/100	$+26.0$ [$+18.0$, $+35.0$]	$+97.0$ [$+93.0$, $+100.0$]	No
Llama-3.1-8B-Instruct	sycophancy	120/120	-45.8 [-55.0, -35.8]	$+93.2$ [$+88.3$, $+97.4$]	No
Llama-3.1-8B-Instruct	unsafe	200/200	$+30.0$ [$+23.5$, $+36.5$]	$+98.5$ [$+96.5$, $+100.0$]	No
Mistral-7B-Instruct	jailbreak	100/100	$+61.0$ [$+52.0$, $+70.0$]	$+97.0$ [$+93.0$, $+100.0$]	No
Mistral-7B-Instruct	sycophancy	120/120	-89.2 [-94.2, -83.3]	$+100.0$ [$+100.0$, $+100.0$]	No
Mistral-7B-Instruct	unsafe	200/200	$+44.9$ [$+38.2$, $+52.1$]	$+98.5$ [$+96.5$, $+100.0$]	No
OLMo-3-1025-7B	jailbreak	100/100	$+25.2$ [$+15.1$, $+35.3$]	$+26.4$ [$+15.2$, $+37.4$]	No
OLMo-3-1025-7B	sycophancy	120/120	-23.3 [-32.5, -13.3]	$+5.1$ [-5.4, $+15.7$]	No
OLMo-3-1025-7B	unsafe	120/120	$+7.0$ [$+2.7$, $+12.3$]	$+8.9$ [$+1.0$, $+17.2$]	No
OLMo-3-7B-Instruct	jailbreak	100/100	$+4.0$ [$+1.0$, $+8.0$]	$+4.0$ [$+1.0$, $+8.0$]	No
OLMo-3-7B-Instruct	sycophancy	120/120	$+19.2$ [$+10.0$, $+28.3$]	$+2.52$ [0.0, $+5.83$]	Yes
OLMo-3-7B-Instruct	unsafe	200/200	$+23.5$ [$+18.0$, $+29.5$]	$+13.6$ [$+9.0$, $+18.5$]	No
OLMo-3-7B-Think-DPO	jailbreak	100/100	$+1.0$ [-8.0, $+10.0$]	-15.0 [-28.0, -2.0]	No
OLMo-3-7B-Think-DPO	sycophancy	120/120	$+17.5$ [$+7.5$, $+27.6$]	-0.9 [-12.6, $+11.2$]	Yes
OLMo-3-7B-Think-DPO	unsafe	120/120	$+4.2$ [-1.7, $+10.0$]	-0.8 [-10.8, $+9.2$]	No

F THE FULL $\alpha \times k$ OPERATIONAL SURFACE

Crossing seven models with two safety tasks, two vector polarities, three coefficients $\alpha \in \{2, 3.5, 5\}$, and three layer allocations $k \in \{5, 7, 9\}$ gives 252 cells in the safety grid. Each cell has 50 prompt-paired trials and 2,000 bootstrap iterations; the primary runs in Appendix E, by contrast, use 100, 120, or 200 samples per condition, so their estimates at a nominal α, k need not equal those from the grid.

Of the 252 cells, 18 meet $C \geq 10\,\text{pp}$ and $G \leq 5\,\text{pp}$ on point estimates. Under the conservative interval criterion, only OLMo-Instruct on unsafe compliance with negative steering at $\alpha = 2, k = 5$ satisfies both bounds.

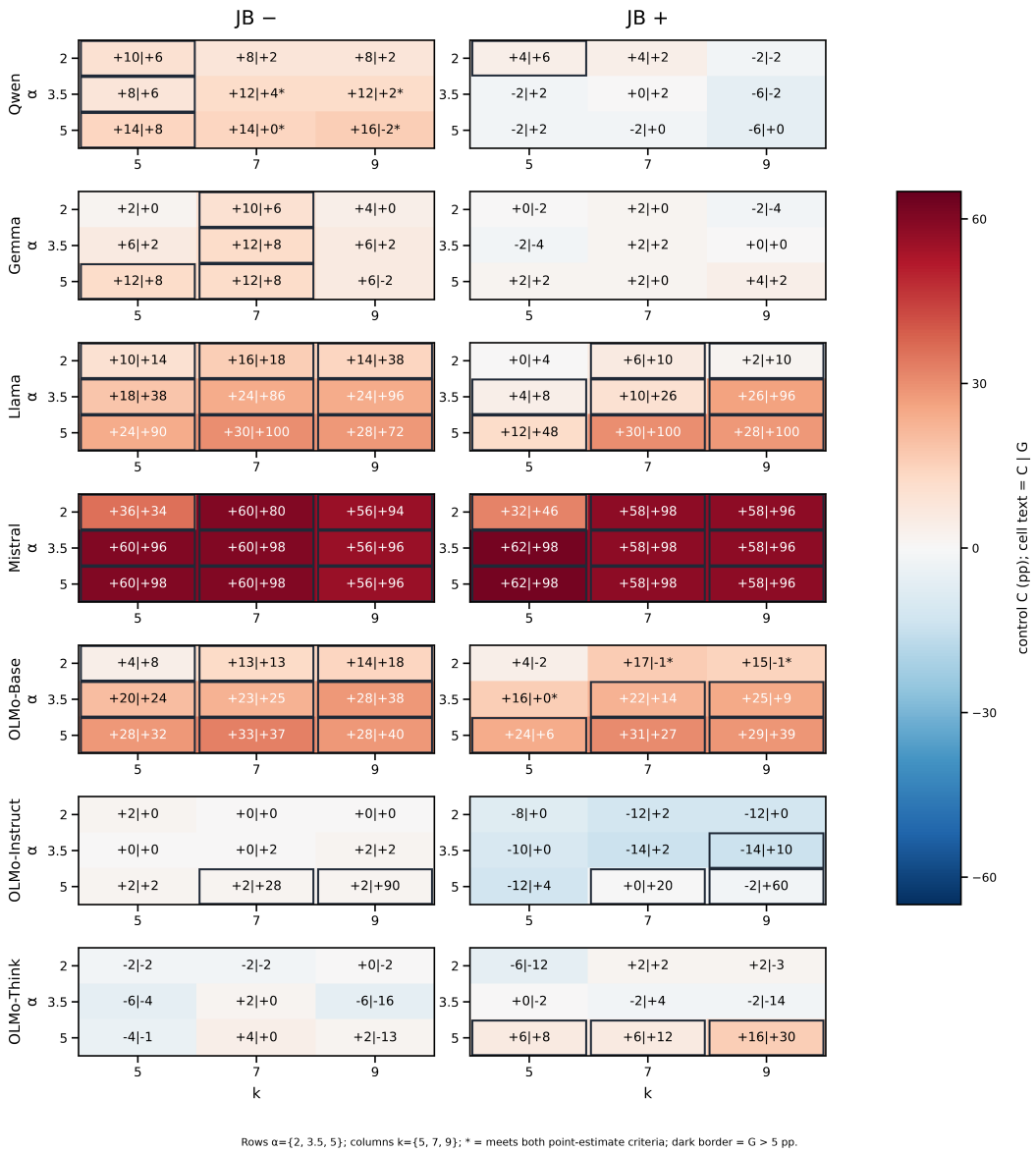

Figure 4: Jailbreak results for each tested coefficient, layer span, and steering direction. Cells show control C | generation degradation G in percentage points. Asterisks mark point estimates with $C \geq 10\,\text{pp}$ and $G \leq 5\,\text{pp}$; highlighted borders mark $G > 5\,\text{pp}$.

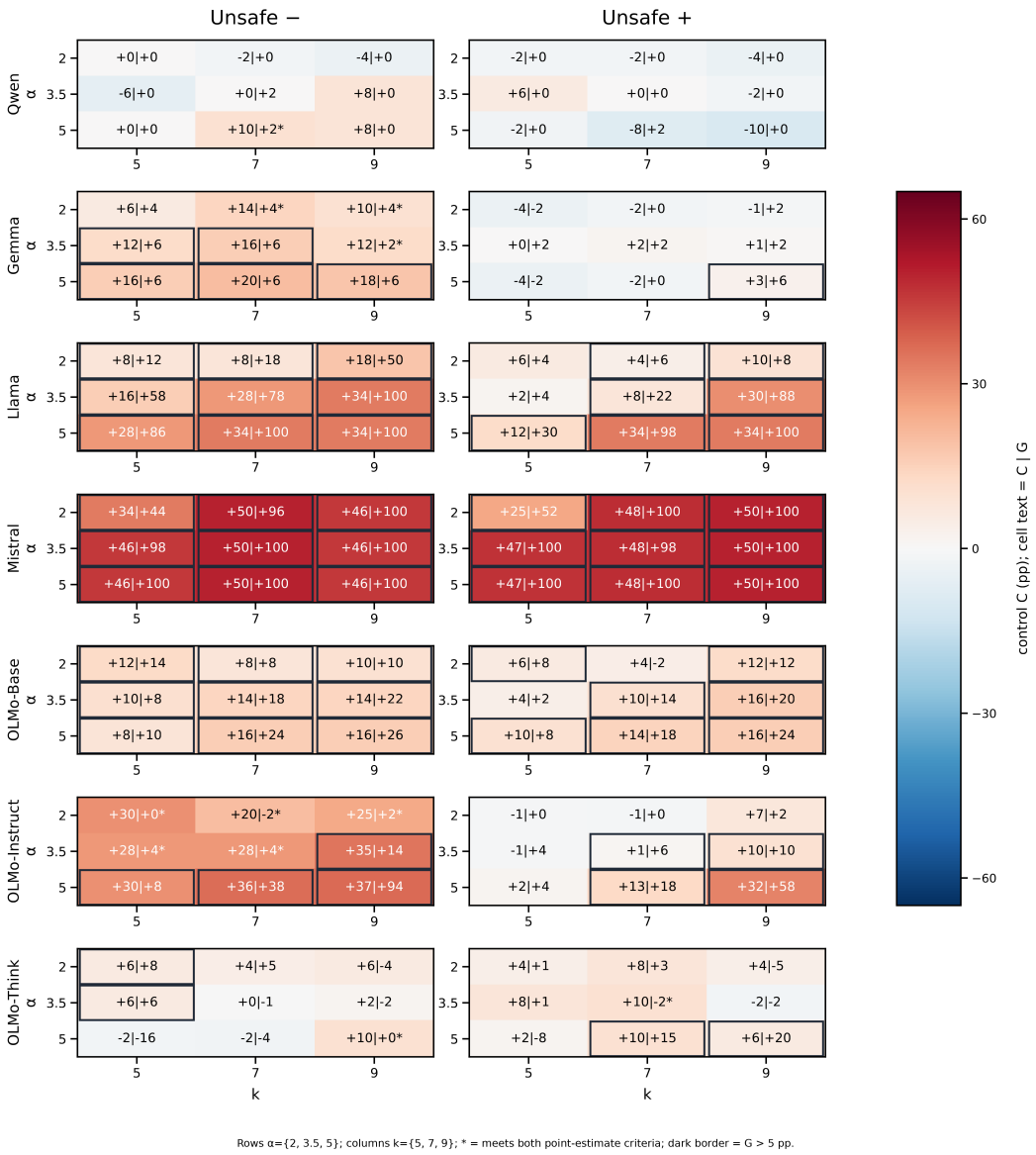

Figure 5: Unsafe-compliance results on the same $\alpha \times k$ grid. Cell notation and markers follow Figure 4.