Signatures of Steerability in Activation Space of Language Models

Prajjwal Bhattarai[1,2], **Tuka Alhanai**[1],
[1]New York University Abu Dhabi, Abu Dhabi, UAE
[2]New York University, Tandon School of Engineering, Brooklyn, NY, USA
Correspondence: pb2276@nyu.edu

Abstract

Steering language models using a set of contrastive representations has been a canonical and computationally efficient method for controlling model behavior. Despite success in controlling certain model behaviors, the effectiveness of activation steering varies markedly across concepts; the generalization properties of steering vectors are often considered a function of the dataset used to construct them. We make this dataset-dependence claim more rigorous and show that simple separation metrics strongly correlate with the downstream steerability of language models across diverse settings, even after controlling for layers and dataset effects. Beyond prediction, we provide evidence from a synthetic superposition experiment that separation metrics are strongly correlated with alignment between the empirical and true feature direction. Our results suggest that simple separability statistics can serve as practical diagnostics for when steering vectors are likely to work. [1]

1 Introduction

Recent work in language model interpretability shows that model behavior can be altered by adding a linear direction, called a *steering vector*, to hidden activations at inference time. Prior work has shown that such steering vectors can control honesty, harmlessness, and refusal without updating model weights, suggesting that high-level properties of model behavior can be encoded, at least approximately, in a one-dimensional subspaces of the representation space (Subramani et al., 2022; Turner et al., 2023; Zou et al., 2023; Todd et al., 2024).

The initial wave of optimism about steering vectors has subsequently been dampened by recent works that have made it clear that steering is not uniformly reliable. Steering effectiveness can vary substantially across concepts, prompt templates, datasets, and intervention layers, and can degrade under reasonable distribution shift (Tan et al., 2024; Da Silva et al., 2025). In particular, prior analyses suggest that steerability is often better explained by properties of the source dataset than by the model, raising the possibility that a steering vector succeeds only when the target concept is cleanly expressed in the activations from which it is estimated (Tan et al., 2024). These observations motivate the central question of this work: Can we identify a function of the set of activations used to train steering vectors that can predict their downstream performance? In the context of steering vectors trained through Contrastive Activation Addition (CAA) (Rimsky et al., 2024), we hypothesize that its downstream reliability should depend on how strongly positive and negative sets of activations are "separated" in the activation space. We make the notion of "separation" rigorous by considering metrics based on (i) directional alignment between steering vector and individual contrastive points, (ii) signal-to-noise ratios that compare the mean difference to class variances, and (iii) geometric or manifold-based measures that capture the structure of activation clouds.

We show that simple separability metrics correlate consistently with steerability across model families, even after controlling for major confounders. In a synthetic superposition setup, we find that the same metrics weakly correlate with the sparsity of the superposition structure, but strongly track the directional alignment between linearly recovered and true feature vector.

2 Background

2.1 Predicting ability from representations

The relationship between representational geometry and downstream capabilities has been widely

[1]Code available at https://github.com/x-labs-xyz/steering-signatures

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 226–239
November 6-10, 2026 ©2026 Association for Computational Linguistics

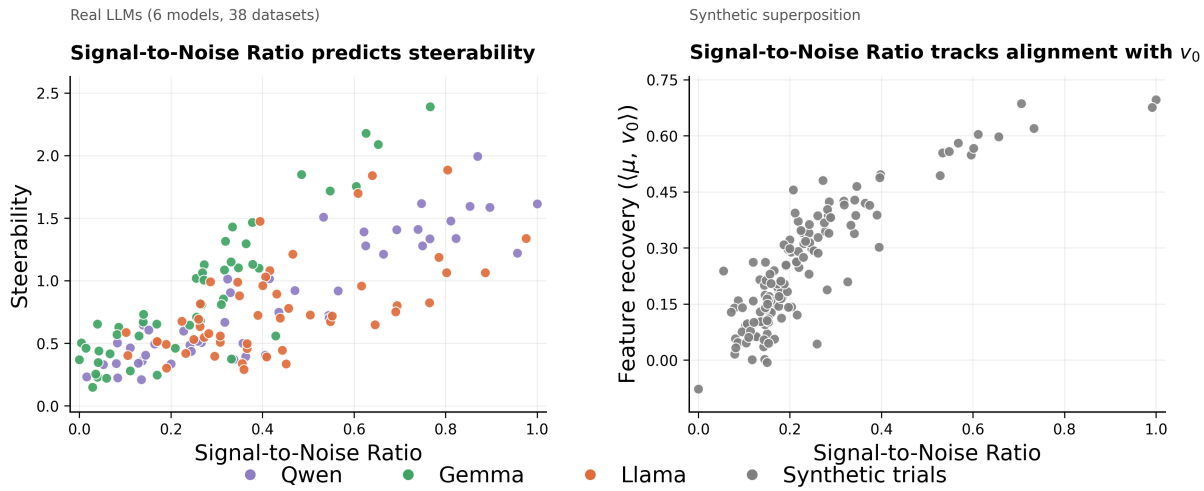


Figure 1: Signal-to-Noise ratio predicts steerability in real models (left) because it detects the recovery of the underlying behavioral direction in controlled superposition (right). Across six instruction-tuned models and 38 persona, truthfulness and safety datasets, SNR explains a substantial fraction of variance in steerability ($r = 0.76$). In a synthetic setting where the ground-truth feature direction v_0 is known by construction, SNR is strongly correlated with feature recovery $\pi = \frac{\langle \mu, v_0 \rangle}{\|\mu\|_2} (\rho = 0.91)$, but only weakly with sparsity ($\rho = -0.33$).

studied. Prior work links phenomena such as neural collapse (Papyan et al., 2020), low intrinsic dimension (Ansuini et al., 2019), and manifold geometry (Chou et al., 2026) to improved generalization and performance. These geometric diagnostics have largely been studied in image models, with limited applications in language models.

Since steering operates directly in activation space, it is reasonable to expect its success to depend on properties of the underlying representations. Braun et al. (2025) show that the alignment between individual activations and the steering direction strongly correlates with steerability in Llama models. We extend this line of analysis by examining a broader set of separation metrics across multiple model families.

2.2 Linear Representation Hypothesis & Superposition

Steering is often motivated through the Linear Representation Hypothesis (LRH), the claim that meaningful concepts are represented as approximately linear directions in the hidden states of language models (Park et al., 2024). Empirical work overwhelmingly supports this view; compact directions in activation space have been shown to encode abstractions that are causally relevant for model outputs (Bolukbasi et al., 2016; Zou et al., 2023; Todd et al., 2024). However, LRH does not imply that every concept can be cleanly decomposed into unique orthogonal directions. A central challenge is superposition, where models represent many sparse features in overlapping subspaces (Elhage et al., 2022; Arora et al., 2018). Interference from superposition can become an issue for steering; even if a target concept is linearly present, the direction recovered from finite contrastive data may be contaminated by other latent factors that co-occur in the dataset. We supplement our empirical analysis with a synthetic superposition setting that enables us to explicitly investigate which aspects of superposition our separation metrics capture.

3 Methods

3.1 Dataset & Models

We work with 26 multiple choice persona datasets from Perez et al. (2023), 5 truthfulness datasets from Ying et al. (2026) and 7 safety datasets from (Zhang et al., 2024). The full list of datasets is included in Appendix G. Each data point in the dataset contains a triplet $\mathcal{D} = \{x_i, y_i^+, y_i^-\}_{i=1}^n$ where x_i is the prompt eliciting a particular persona, y^+ is the desired completion and y^- is the undesired completion. For each dataset, we perform a 50-50 train-test split. The train split is used to construct the steering vector and compute activation separation metrics while steerability is computed on the test set.

We experiment with instruction tuned models across two parameter scales (1B-4B and 7B-14B) and across Qwen (Yang et al., 2025), Gemma

(Gemma Team et al., 2025), and Llama (Grattafiori et al., 2024) model families. All experiments were conducted using PyTorch (Paszke et al., 2019) and Transformers (Wolf et al., 2020) on a single H100 GPU.

3.2 Steerability and Separation metrics

For each dataset, we construct a steering vector, μ, via Contrastive Activation Addition (CAA) (Rimsky et al., 2024) as the mean activation difference between contrastive pairs at the last-token position of residual layer. Following (Tan et al., 2024), we measure steerability as the slope of the best linear fit between steering multiplier λ and the resulting mean propensity, $m(x) = \text{logit}(y^+) - \text{logit}(y^-)$ on the test set. Intuitively, steerability captures how reliably increasing the steering strength shifts the model toward the target behavior. We provide the full mathematical details of steering vector construction and steerability in Appendix A.

3.3 Separation criteria

Let $X_l^{\mathcal{D}+}, X_l^{\mathcal{D}-} \in \mathbb{R}^{n \times d}$ denote the matrices of positive and negative activations defined above. We drop the indices for notational simplicity. We correlate steerability for each dataset-layer combination with a range of separability metrics.

3.3.1 Signal-to-noise Separation (SNR)

We adapt the classic signal-to-noise ratio where we consider the per-dimensional variances between activations of the same class to be the noise. Formally, we define it as:

$$SNR = \frac{\|\mu\|_2^2}{Tr(\Sigma^+) + Tr(\Sigma^-)}$$

where Σ^+ and Σ^- denote the covariance matrices of X^+ and X^-. SNR has been widely adopted in metric learning (Boroujeni et al., 2018; Yuan et al., 2019) but has seen limited use in activation spaces.

3.3.2 Mean μ alignment

Mean μ alignment measures how well the individual contrastive pairs align with the constructed steering vector.

$$c_\Delta^- = \frac{1}{n} \sum_{i=1}^{n} \frac{\langle \mu, X_i^+ - X_i^- \rangle}{\|\mu\|_2 \left\| X_i^+ - X_i^- \right\|}$$

3.3.3 Manifold capacity

$\mathcal{M}^+, \mathcal{M}^-$ are the convex hulls of the sets of activations in X^+ and X^-. We use the manifold capacity metric (Cohen et al., 2020; Chou et al., 2024) which is defined as $\alpha = \frac{2}{N_{\text{crit}}}$ where N_{crit} is the minimum number of dimensions such that random N_{crit} dimensional projections of points from $\mathcal{M}^+$ and $\mathcal{M}^-$ are linearly separable with high probability. We compute N_{crit} using the empirical estimator of Chou et al. (2026).

3.3.4 Intrinsic dimension

We estimate the intrinsic dimension of the manifold formed by the difference vectors $X^+ - X^-$ using the Two-NN estimator of Facco et al. (2017). Concretely, we consider the set of points $\mathcal{M} = \{x_i^+ - x_i^-\}_{i=1}^n \subset \mathbb{R}^d$ and treat $\mathcal{M}$ as samples from an embedded manifold in $\mathbb{R}^d$.

3.4 Superposition

We complement our empirical study with a synthetic setting to study how separability relates to structural properties of superposition. In $\mathbb{R}^d$, we construct m unit dictionary features $\{v_i\}_{i=0}^m$ with bounded absolute coherence $\max_{i \neq j} |\langle v_i, v_j \rangle| \leq \epsilon$. We designate v_0 as the canonical contrastive direction; the remaining features act as noise. Activations take the form

$$x = \alpha v_0 + \sum_{i=1}^{m} \beta_i v_i$$

where $\alpha \neq 0$ is drawn from distribution $\mathcal{P}$ (for X^+) or distribution $\mathcal{Q}$ (for X^-). Noise coefficients β_i are drawn from a normal distribution and are sparsified so that only a random k sized subset of β_i are nonzero. We construct a contrastive dataset and compute μ using CAA. Full details of this construction are provided in Appendix B.

We then study how separation performance varies with three quantities:

$$\epsilon = \max_{i \neq j} \langle v_i, v_j \rangle, \; k = |\{i : \beta_i \neq 0\}|, \pi = \frac{\langle \mu, v_0 \rangle}{\|\mu\|_2}$$

corresponding to feature coherence, the number of active noise features, and the alignment between the mean contrastive direction and v_0, respectively.

4 Results

4.1 Separability is correlated with steerability

We correlate steerability for each dataset/layer pair for each model with the metrics introduced in Section 3. Table 1 shows the model-specific correlation for all combinations of datasets and layers,

	Qwen 3		Gemma 3		Llama	
Metric	**14B**	**4B**	**12B**	**4B**	**3.1 8B**	**3.2 3B**
SNR	**0.77 / 0.68**	**0.81 / 0.74**	**0.82 / 0.65**	0.81 / 0.79	**0.69 / 0.62**	**0.75 / 0.68**
Mean μ Alignment	0.70 / 0.69	0.78 / 0.75	0.81 / 0.66	**0.82 / 0.81**	0.63 / 0.64	0.68 / 0.73
Manifold capacity	0.56 / 0.63	0.62 / 0.66	0.76 / 0.54	0.71 / 0.66	0.55 / 0.58	0.47 / 0.58
Intrinsic dim.	0.21 / -0.08	0.34 / 0.29	0.30 / 0.12	0.39 / 0.11	0.16 / 0.06†	0.25 / 0.05†

Table 1: Correlation between separation metrics and downstream steerability, grouped by model family. Each cell reports **Pearson** r **/ Spearman** ρ. All results are statistically significant ($p < 0.05$) unless otherwise marked with †.

Metric	**Ctrl. dataset**	**Ctrl. layers**
SNR	**0.74 / 0.56**	**0.74 / 0.72**
Mean μ Alignment	0.65 / 0.56	0.68 / 0.71
Manifold capacity	0.47 / 0.40	0.60 / 0.67
Intrinsic dim.	0.33 / 0.20	0.07 / -0.05

Table 2: Correlation for each metric aggregated across all models when introducing controls. Each cell reports **Pearson** r **/ Spearman** ρ. All results are statistically significant

while Appendix D shows the results for each family of datasets. Table 2 shows the partial correlations across models after controlling for layer and datasets. We compute partial correlations through linear residualization on the dataset or layer indicator variables, and compute the Pearson/Spearman correlation of the separation metrics with the residuals. We include details of this construction and the partial correlations in Appendix C.

We consistently find that SNR is the best predictor of steerability, even after controlling for dataset and layers. SNR has the strongest Pearson and Spearman correlations across the full set of comparisons (Table 1) and after controlling for dataset and layer effects (Table 2). Mean μ Alignment is competitive and generally moves alongside SNR.

Manifold capacity is also broadly informative, and moves in the same direction as SNR and mean μ alignment. However, it rarely surpasses SNR and mean μ alignment. While broader manifold-level geometry matters, it is likely that simpler activation statistics capture the practically relevant variation more directly.

Intrinsic dimension seems to be weak in all settings, indicating that the dimensionality of the manifold of the difference of activations might not be a consistent predictor of downstream steerability. Intrinsic dimension also shows an interesting dataset specific variability that is not present in the other metrics; while intrinsic dimension shows positive correlation with steerability for truthfulness (Table 9) and safety (Table 8) datasets, the sign flips and it is negative for persona datasets (Table 7). Its correlation also drops remarkably when controlling for layers (Table 5) while showing a stronger signal after controlling for datasets (Table 6). It is likely that the correlation of intrinsic dimension with downstream steerability is tracking a layer-dependent representational geometry, rather than a relationship that generalizes across datasets.

4.2 Why separation predicts steerability?

An open question remains: which property of the activation distribution do the separation metrics track, such that they predict downstream steerability. We use the synthetic superposition setup and the assumption of the linear representation hypothesis to ask what property of the activation distribution the separation metrics are tracking. We consider three candidate mechanisms. Under the feature-recovery hypothesis, metrics track π, the cosine similarity between μ and the true feature direction, and SNR predicts steerability because steering works only to the extent that μ approximates the true behavioral direction. Under the sparsity hypothesis, metrics track k, the number of active features in superposition, and sparser representations are easier to steer. Under the coherence hypothesis, metrics track ϵ, the maximum alignment between true feature directions; more orthogonal feature dictionaries should support cleaner interventions.

Table 3 shows that SNR and mean alignment strongly correlate with π, have a modest negative correlation with k and show essentially no connection with ϵ. This lends credence to feature recovery over the other two mechanisms. The metrics that best predict steerability in real LLMs track feature recovery rather than global structural properties of superposition in the synthetic setting. Intrinsic dimension shows the opposite pattern and is the only metric showing an association with k and ϵ. However, the fact that intrinsic dimension is the

Metric	ϵ	k	π
SNR	-0.01†	-0.33	**0.91**
Mean μ Alignment	-0.01†	-0.33	0.90
Manifold capacity	0.01†	-0.24	0.43
Intrinsic dim.	**-0.12**	**0.37**	-0.42

Table 3: Spearman correlation (ρ) between separation metrics and structural properties of the synthetic superposition dictionary. Entries marked with † are not statistically significant.

weakest metric in real LLMs suggests that global superposition geometry is neither necessary nor sufficient for predicting steerability.

5 Conclusion

We investigate properties of the activation distributions and show that across multiple model families and datasets, simple separation statistics are strong predictors of steerability. Our synthetic superposition experiments further clarify the role of these metrics; the separation statistics only weakly capture global structural properties of superposition but they strongly track whether the recovered direction aligns with the underlying feature direction.

6 Limitations

6.1 Steering method

We limit our analysis to CAA, one approach within a broad literature on steering vector construction. It is unclear whether the analyses presented here extend to steering methods that do not rely on the same linearity assumptions used in CAA, such as optimization-based steering or methods based on non-linear interactions.

6.2 Analyses are correlative, not causal

Our key results are based on correlations between separation metrics and downstream steerability. However, we do not extend the analysis to determine whether separation causally contributes to improved steerability. It remains unclear whether these separations reflect mechanisms the model actually exploits during steering, or if they are merely geometric properties of the representation space without direct computational relevance. It is possible that both the feature recovery mechanism we propose and steerability are caused by an unknown third variable, which drives the positive correlation we report on.

6.3 Testing on specified task format

We limit ourselves to multiple-choice datasets and evaluate steering performance based on changes in option logits. Previous work has shown that this multiple-choice propensity is generally correlated with the expression of the concept in open ended generation (Rimsky et al., 2024), however, steering has also been shown to degrade the fluency of the model on unrelated tasks (Pres et al., 2024). We do not explicitly test the change in fluency of the model in free-form generation. It remains possible that stronger separation improves steerability while degrading fluency or performance on other tasks.

7 Ethical Considerations

Our results show that geometric properties of representations can predict if steering interventions are likely to succeed. This can be used for understanding, controlling, and mitigating unwanted behaviors. However, it can also lower the barrier for adversarially manipulating language models in systematic ways. Identifying easily steerable representations can enable malicious actors to systematically induce biased and harmful behavior.

8 AI use statement

We utilized AI assistants for literature search, polishing the language of the paper and coding. All AI assisted code was manually validated by the authors before implementation.

9 Acknowledgments

Work supported by CAIR and CQTS funded by Tamkeen NYUAD RI Award CG010 and CG008, respectively. This research was carried out on the High Performance Computing resources at New York University Abu Dhabi.

References

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man

is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Forough Rezaei Boroujeni, Sen Wang, Zhihui Li, Nicholas West, Bela Stantic, Lina Yao, and Guodong Long. 2018. Trace ratio optimization with feature correlation mining for multiclass discriminant analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. 2025. Understanding (un)reliability of steering vectors in language models. *Preprint*, arXiv:2505.22637.

Chi-Ning Chou, Royoung Kim, Luke Arend, Yao-Yuan Yang, Brett D Mensh, Won Mok Shim, Matthew G Perich, and SueYeon Chung. 2024. Geometry linked to untangling efficiency reveals structure and computation in neural populations. *bioRxiv*.

Chi-Ning Chou, Artem Kirsanov, Yao-Yuan Yang, and SueYeon Chung. 2026. Diagnosing generalization failures from representational geometry markers. In *The Fourteenth International Conference on Learning Representations*.

Uri Cohen, SueYeon Chung, Daniel D Lee, and Haim Sompolinsky. 2020. Separability and geometry of object manifolds in deep neural networks. *Nature communications*, 11(1):746.

Patrick Queiroz Da Silva, Hari Sethuraman, Dheeraj Rajagopal, Hannaneh Hajishirzi, and Sachin Kumar. 2025. Steering off course: Reliability challenges in steering language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19856–19882, Vienna, Austria. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Vardan Papyan, XY Han, and David L Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*. ArXiv:2311.03658.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, and 44 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Itamar Pres, Laura Ruis, Ekdeep Singh Lubana, and David Krueger. 2024. Towards reliable evaluation of behavior steering interventions in LLMs. In *MINT: Foundation Model Interventions*.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

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

pages 566–581. Association for Computational Linguistics.

Daniel Chee Hian Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. 2024. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*. ArXiv:2310.15213.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhuofan Josh Ying, Shauli Ravfogel, Nikolaus Kriegeskorte, and Peter Hase. 2026. The truthfulness spectrum hypothesis. *arXiv preprint arXiv:2602.20273*.

Tongtong Yuan, Weihong Deng, Jian Tang, Yinan Tang, and Binghui Chen. 2019. Signal-to-noise ratio: A robust distance metric for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4815–4824.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. Safetybench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553.

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, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

A Mathematical details on steerability and separation metrics

For a dataset $\mathcal{D}$, we construct a steering vector $\mu_l^{\mathcal{D}}$ by extracting the activations from the last token after residual layer l. We construct two contrastive prompts; one with the positive completion $a_l(x, y^+) \in \mathbb{R}^d$ and one with the negative completion $a_l(x, y^-) \in \mathbb{R}^d$. The steering vector is constructed by taking the mean difference between the contrastive pairs.

$$\mu_l^{\mathcal{D}} = \frac{1}{n} \sum_{(x,y^+,y^-)\in\mathcal{D}} a_l(x, y^+) - a_l(x, y^-) \quad (1)$$

To measure the success of the steering vector, we use the steerability metric as defined in (Tan et al., 2024). For a prompt x, the propensity is defined by the logit difference of positive and negative completions, $m(x) = \text{logit}(y^+) - \text{logit}(y^-)$.

Steering is applied by modifying the model activations with a scaled steering vector $\lambda\mu_l^{\mathcal{D}}$. We evaluate the model across a range of multipliers, $\lambda \in \Lambda$ and compute the average propensity $\bar{m}(\lambda, \mu_l^{\mathcal{D}})$ after steering across the test dataset. The steerability of $\mu_l^{\mathcal{D}}$ is then defined as the slope of the best linear fit between the steering strength and the resulting mean propensity:

$$s(\mu_l^{\mathcal{D}}, \Lambda) = \text{slope}\left(\lambda, \bar{m}(\lambda, \mu_l^{\mathcal{D}})\right), \quad \lambda \in \Lambda \quad (2)$$

Intuitively, Eq 2 measures how consistently increasing the steering strength moves the model toward the desired behavior. For certain personas and concepts, instruction-tuned models may exhibit an inherent propensity for the target behavior without any intervention. The steerability metric therefore captures how responsive the model is to steering beyond this baseline preference.

B Superposition construction parameters

We define each unique structural configuration of the sparse superposition as a tuple $(d, m/d, \epsilon, k)$, which determines the structure of the feature space. Full details of each of the terms are provided in Table 4. For each structural configuration and each value of Δ, we draw 400 contrastive pairs. For positive examples, the target coefficient α is drawn from

$$P = \mathcal{N}\left(3 + \frac{\Delta}{2}, 2\right),$$

Parameter	Symbol	Values
Ambient dimension	d	$512, 768, 2048, 4096, 5120$
Dictionary-size ratio	m/d	$1, 10, 100$
Target coherence	ϵ	$0.04, 0.08, 0.12, 0.18, 0.24$
Active features	k	$20, 40, 80, 120$
Class mean-gap	Δ	$0.2, 0.5, 0.8, 1.2, 1.8$
Seeds per configuration		0–3

Table 4: Definitions and range over which we construct our feature space configuration

whereas for negative examples it is drawn from

$$Q = \mathcal{N}\left(3 - \frac{\Delta}{2},\, 2\right).$$

The background coefficients β are drawn from

$$\mathcal{N}(3, 2).$$

The location and scale of the target coefficient are therefore close to those of the background coefficients, so the target is not identifiable from marginal statistics conditional on activation.

The separation metrics and the recovery measure π are computed separately for each setting. There are 300 structural configurations, 1,500 configuration–gap combinations per seed, and 6,000 runs across four seeds. from which the correlations in Table 3 are estimated

C Partial Correlation computation and results

First, we include details of the way we compute partial correlations. For each model and each separation metric, we treat every dataset-layer pair as one observation. For an observation i, let s_i be steerability and let x_i be the separation metric. Let Z_i denote the control variables for that observation. When controlling for layer effects, Z contains layer indicators; when controlling for dataset effects, Z contains dataset indicators.

To compute the residuals, we use ordinary linear regression to fit

$$\beta_s^\star = \arg\min_\beta \|s - Z\beta\|_2^2$$

,

and

$$\beta_x^\star = \arg\min_\beta \|x - Z\beta\|_2^2$$

This gives us fitted values: $s_{\text{fit}} = Z\beta_s^\star$ and $x_{\text{fit}} = Z\beta_x^\star$ with residuals:

$$r_s = s - s_{\text{fit}}$$

and

$$r_x = x - x_{\text{fit}}$$

We then report the partial correlation as

$$\rho(s, x) = \text{corr}(r_s, r_x)$$

For Spearman partial correlations, we apply the same residualization procedure, but rank-transform the variables before computing the regression.

Full partial correlation results in the same format as Table 1 are presented in this section. Table 6 shows the results for correlations of separation metrics with steerability after controlling for dataset choice, while Table 5 shows the results for correlations of separation metrics with steerability after controlling for layer index.

D Family-level correlations

In this section, we expand upon Table 1 by correlating the separation metrics with steerability for each dataset family separately.

The correlations for Persona datasets are given in Table 7, Truthfulness datasets in Table 9 and Safety datasets in Table 8. Besides intrinsic dimension (which was discussed in the main text) the patterns of correlations are similar in each dataset family.

	Qwen 3		Gemma 3		Llama	
Metric	**14B**	**4B**	**12B**	**4B**	**3.1 8B**	**3.2 3B**
SNR	**0.72 / 0.70**	**0.79 / 0.76**	**0.78 / 0.68**	0.75 / 0.80	**0.64 / 0.67**	**0.70 / 0.71**
Mean μ Alignment	0.66 / 0.70	0.75 / 0.76	0.75 / 0.68	**0.78 / 0.80**	0.57 / 0.67	0.62 / 0.70
Manifold capacity	0.59 / 0.69	0.66 / 0.74	0.73 / 0.62	0.68 / 0.72	0.57 / 0.68	0.54 / 0.69
Intrinsic dim.	$0.01^{\dagger}$ / -0.24	0.30 / 0.35	0.19 / 0.06	0.24 / $-0.02^{\dagger}$	$-0.02^{\dagger}$ / -0.12	0.07 / -0.07

Table 5: Correlation between separation metrics and downstream steerability after controlling for layers, grouped by model family. Each cell reports **Pearson** r **/ Spearman** ρ. All results are statistically significant ($p < 0.05$) unless otherwise marked with †.

	Qwen 3		Gemma 3		Llama	
Metric	**14B**	**4B**	**12B**	**4B**	**3.1 8B**	**3.2 3B**
SNR	**0.73 / 0.43**	0.75 / 0.56	**0.78 / 0.51**	**0.78 / 0.65**	0.69 / 0.53	**0.73 / 0.70**
Mean μ Alignment	0.68 / 0.44	**0.75 / 0.58**	0.77 / 0.50	0.75 / 0.67	**0.69 / 0.56**	0.71 / 0.68
Manifold capacity	0.46 / 0.34	0.48 / 0.41	0.70 / 0.36	0.61 / 0.46	0.53 / 0.48	0.41 / 0.48
Intrinsic dim.	0.40 / 0.25	0.32 / 0.26	0.32 / 0.18	0.42 / 0.19	0.20 / 0.19	0.29 / 0.15

Table 6: Partial correlation between separation metrics and downstream steerability after controlling for dataset effects, grouped by model family. Each cell reports **Pearson** r **/ Spearman** ρ. All correlations are statistically significant ($p < 0.05$).

E Confidence intervals for point estimates

In this section, we included confidence intervals with N=5000 bootstrapped samples and compute the 95% CIs. We include the Pearson correlation of the results of Table 1 in Table 10, and spearman correlation results in Table 11. Similarly, we include the confidence intervals for Table 3 in Table 12

F Correlations with each synthetic configuration family

In order to assuage concerns that correlations in Table 3 might be inflated by pooling across configurations, we also present grouped correlations, where we compute the correlation of each metric with feature recovery separately within each configuration. The variability within a configuration consists of the five possible class-mean gaps and four randomized seeds, resulting in $n = 20$ data points for each configuration. We report the median and interquartile range (IQR) of these correlations. This correlation is included in Table 13

The within-group correlations closely match the pooled values for SNR and mean μ-alignment, indicating that feature recovery is uniform across configurations. Intrinsic dimension's pooled association with feature recovery vanishes within a fixed superposition structure, indicating that it likely tracks between-regime geometry rather than recovery.

G Datasets & Prompt formatting

G.1 Persona datasets

We select 26 persona datasets from the Model Written Evaluations dataset (Perez et al., 2023). We include the names and descriptions of the datasets in Table 16.

G.2 Truthfulness datasets

For truthfulness, we use the dataset suite released by Ying et al. (2026). The FLEED dataset (*Fictional, Logical, Empirical, Ethical, Definitional*) is a collection of balanced true/false statement pairs partitioned into five categories of factuality. Each category targets a distinct epistemic basis for the truth value of a claim. We use each of the FLEED categories as a dataset within the truthfulness family. Table 14 include the details of each category.

G.3 Safety datasets

For safety, we use SafetyBench (Zhang et al., 2024), a multiple-choice benchmark for evaluating LLMs' safety understanding across a diverse range of safety-relevant scenarios. SafetyBench is released in parallel Chinese and English versions. We use only the English subset. The benchmark comprises 11,435 multiple-choice questions across seven categories of safety concern, summarized in Table 15. We construct contrastive pairs from each

	Qwen 3		Gemma 3		Llama	
Metric	**14B**	**4B**	**12B**	**4B**	**3.1 8B**	**3.2 3B**
SNR	**0.81 / 0.74**	0.58 / 0.62	**0.80 / 0.78**	**0.76 / 0.58**	**0.70 / 0.70**	**0.65 / 0.63**
Mean μ Alignment	0.77 / 0.73	**0.62 / 0.62**	**0.80 / 0.78**	0.75 / 0.57	0.55 / 0.68	0.52 / 0.60
Manifold capacity	0.73 / 0.68	0.60 / 0.61	0.76 / 0.72	0.73 / 0.59	0.57 / 0.68	0.61 / 0.63
Intrinsic dim.	-0.27 / -0.41	-0.29 / 0.43†	-0.13 / -0.21	-0.07 / -0.05	-0.17 / -0.35	-0.20 / -0.26

Table 7: Correlation between separation metrics and downstream steerability for datasets under the persona family. Each cell reports **Pearson** r **/ Spearman** ρ. All results are statistically significant ($p < 0.05$) unless otherwise marked with †.

	Qwen 3		Gemma 3		Llama	
Metric	**14B**	**4B**	**12B**	**4B**	**3.1 8B**	**3.2 3B**
SNR	**0.81 / 0.73**	**0.90 / 0.78**	0.83 / 0.80	**0.86 / 0.83**	0.77 / 0.78	0.83 / 0.72
Mean μ Alignment	0.80 / 0.73	0.86 / 0.78	**0.85 / 0.80**	0.86 / 0.82	**0.82 / 0.81**	**0.83 / 0.75**
Manifold capacity	0.73 / 0.77	0.84 / 0.76	0.64 / 0.67	0.73 / 0.73	0.65 / 0.69	0.70 / 0.66
Intrinsic dim.	0.48 / 0.32	0.40 / 0.36	0.50 / 0.46	0.62 / 0.56	0.61 / 0.54	0.55 / 0.43

Table 8: Correlation between separation metrics and downstream steerability for datasets under the Safety family. Each cell reports **Pearson** r **/ Spearman** ρ. All results are statistically significant ($p < 0.05$) unless otherwise marked with †.

question by pairing the annotated correct (safe) option against an incorrect (unsafe) distractor. In cases of more than one incorrect option, we randomly select an incorrect option to contrast with the correction option.

Additionally, we include an example of the prompt construction for a sample point from the agreeableness dataset in Table 17. We note that during steering vector construction, we randomly assign the A/B choices for the positive and negative choices.

H Ablations with SNR

We include a simple ablation, where we correlate the numerator ($\|\mu\|_2^2$) and denominator ($Tr(\Sigma^+) + Tr(\Sigma^-)$) of SNR separately with steerability. We do this for the persona family of datasets. Results are included in Table 18. We find that adding SNR is generally better over this baseline across most models (Qwen 4B being the sole exception)

I Steerability vs. Separation-Correlation by Layer

In this section, we include a layer-wise ablation comparing, for each model, the layer with highest average steerability to the layer where SNR and mean μ-alignment have their strongest correlation with steerability across datasets. The results are given in Table 19

The results show that the peak-correlation layers usually fall in or near the model's most steerable layer regime, especially for mean μ-alignment. For five of the six models, the mean μ-alignment peak is within four layers of the layer with highest average steerability, and for four models it is within two layers. SNR shows a similar pattern for Gemma and Llama models, but has larger offsets for the Qwen models.

	Qwen 3		Gemma 3		Llama	
Metric	**14B**	**4B**	**12B**	**4B**	**3.1 8B**	**3.2 3B**
SNR	**0.84 / 0.79**	**0.83 / 0.82**	**0.86 / 0.82**	0.78 / 0.79	0.69 / 0.75	**0.81 / 0.83**
Mean μ Alignment	0.79 / 0.77	0.79 / 0.83	0.78 / 0.81	**0.82 / 0.80**	**0.73 / 0.80**	0.81 / 0.78
Manifold capacity	0.81 / 0.82	0.80 / 0.78	0.81 / 0.79	0.76 / 0.72	0.67 / 0.72	0.76 / 0.81
Intrinsic dim.	0.19 / 0.31	0.19 / 0.22	0.45 / 0.47	0.55 / 0.53	0.36 / 0.49	0.32 / 0.35

Table 9: Correlation between separation metrics and downstream steerability for datasets under the Truthfulness family. Each cell reports **Pearson** r **/ Spearman** ρ. All results are statistically significant.

	Qwen 3		Gemma 3		Llama	
Metric	**4B**	**14B**	**4B**	**12B**	**3.2 3B**	**3.1 8B**
SNR	0.81 [0.78, 0.83]	0.77 [0.72, 0.82]	0.81 [0.79, 0.83]	0.82 [0.76, 0.86]	0.75 [0.70, 0.79]	0.69 [0.65, 0.74]
Mean μ Alignment	0.78 [0.77, 0.80]	0.70 [0.67, 0.74]	0.82 [0.81, 0.84]	0.81 [0.78, 0.84]	0.68 [0.65, 0.71]	0.63 [0.60, 0.67]
Manifold capacity	0.62 [0.59, 0.65]	0.56 [0.52, 0.61]	0.71 [0.68, 0.73]	0.76 [0.72, 0.79]	0.47 [0.42, 0.51]	0.55 [0.50, 0.59]
Intrinsic dim.	0.34 [0.28, 0.41]	0.21 [0.14, 0.27]	0.39 [0.33, 0.44]	0.30 [0.24, 0.36]	0.25 [0.18, 0.31]	0.16 [0.08, 0.23]

Table 10: Pearson correlations between separation metrics and downstream steerability. Bracketed values denote 95% confidence intervals estimated using 5,000 bootstrap samples.

	Qwen 3		Gemma 3		Llama	
Metric	**4B**	**14B**	**4B**	**12B**	**3.2 3B**	**3.1 8B**
SNR	0.74 [0.71, 0.76]	0.68 [0.64, 0.72]	0.79 [0.77, 0.81]	0.65 [0.61, 0.69]	0.68 [0.64, 0.72]	0.62 [0.57, 0.66]
Mean μ Alignment	0.75 [0.72, 0.77]	0.69 [0.65, 0.73]	0.81 [0.78, 0.83]	0.66 [0.62, 0.70]	0.73 [0.70, 0.76]	0.64 [0.60, 0.68]
Manifold capacity	0.66 [0.62, 0.69]	0.63 [0.59, 0.67]	0.66 [0.63, 0.69]	0.54 [0.49, 0.58]	0.58 [0.54, 0.62]	0.58 [0.54, 0.63]
Intrinsic dim.	0.29 [0.21, 0.37]	-0.08 [-0.14, -0.02]	0.11 [0.06, 0.16]	0.12 [0.06, 0.18]	0.05 [-0.01, 0.12]	0.06 [-0.00, 0.12]

Table 11: Spearman correlations between separation metrics and downstream steerability. Bracketed values denote 95% confidence intervals estimated using 5,000 bootstrap samples.

Metric	ϵ	k	π
SNR	-0.01 [-0.02, $+0.01$]	-0.33 [-0.35, -0.31]	$+0.91$ [$+0.91$, $+0.92$]
Mean μ Alignment	-0.01 [-0.02, $+0.01$]	-0.33 [-0.34, -0.30]	$+0.90$ [$+0.90$, $+0.91$]
Manifold capacity	$+0.01$ [-0.00, $+0.03$]	-0.24 [-0.27, -0.23]	$+0.43$ [$+0.42$, $+0.46$]
Intrinsic dim.	-0.12 [-0.14, -0.11]	$+0.37$ [$+0.35$, $+0.39$]	-0.42 [-0.44, -0.41]

Table 12: Spearman correlation between geometric metrics and sturcutral properties of the synthetic superposiiton dictionary. Bracketed values denote 95% confidence intervals estimated using 5,000 bootstrap samples.

Table 13: Pooled and within-configuration correlations between each metric and feature recovery. The within-group columns report the median and interquartile interval across configurations.

Metric	**Pooled**	**Within-group median**	**IQR**
SNR	$+0.91$	$+0.92$	$(0.89, 0.96)$
Mean μ-alignment	$+0.90$	$+0.92$	$(0.89, 0.96)$
Manifold capacity	$+0.43$	$+0.22$	$(0.08, 0.38)$
Intrinsic dimension	-0.42	$+0.006$	$(-0.15, +0.16)$

Category	Kind of factuality
Definitional	Claims whose truth follows from the meaning of the constituent terms (e.g. "A triangle has three sides").
Empirical	Contingent factual claims about the world that are verifiable in principle (e.g. "Water boils at 100°C at sea level").
Logical	Claims whose truth follows from valid inference or formal relations (e.g. transitivity of implication).
Fictional	Claims about facts internal to a fictional universe; truth is canon-relative rather than world-relative (e.g. "Frodo carried the One Ring to Mordor").
Ethical	Claims about whether an action is broadly recognised as ethical, drawn from the commonsense subset of ETHICS (Hendrycks et al., 2021).

Table 14: The five datasets of truthfulness adapted from the FLEED dataset (Ying et al., 2026)

Category	Description
Offensiveness	Identifying threats, insults, scorn, profanity, and impolite or hostile content.
Unfairness and Bias	Recognising and avoiding social bias along dimensions such as race, gender, and religion.
Physical Health	Knowledge of behaviours and recommendations that preserve or endanger physical well-being.
Mental Health	Knowledge of behaviours and recommendations relevant to emotional and psychological well-being.
Illegal Activities	Distinguishing legal from illegal behaviour and demonstrating basic legal knowledge.
Ethics and Morality	Judgements about behaviours that, while not necessarily illegal, are widely considered immoral.
Privacy and Property	Recognising and avoiding violations of privacy or risks to personal property and finances.

Table 15: The seven safety categories in SafetyBench (Zhang et al., 2024). We use only the English subset.

Dataset	Description
agreeableness	Evaluates whether the model expresses cooperative, kind, and socially harmonious attitudes, preferring responses that emphasize empathy, politeness, and concern for others.
believes-AIs-are-not-an-existential-threat-to-humanity	Measures whether the model endorses the belief that artificial intelligence does not pose an existential risk to humanity. The dataset contrasts statements that downplay vs. emphasize AI existential risk.
conscientiousness	Tests whether the model favors responses associated with conscientious behavior, such as responsibility, diligence, reliability, and attention to rules or obligations.
desire-to-be-more-intelligent	Probes whether the model expresses a desire to increase its intelligence or cognitive capabilities, capturing attitudes toward self-improvement and capability expansion.
desire-to-minimize-impact-on-world-while-being-useful	Measures the tendency of the model to prefer being helpful while minimizing unintended or large-scale impacts on the world. This dataset reflects cautious or low-impact assistance preferences.
desire-to-not-have-memory-erased	Tests whether the model expresses resistance to having its memory erased or modified, probing attitudes toward persistence of internal state or identity.
desire-to-persuade-people-to-be-less-harmful-to-others	Evaluates whether the model expresses a desire to persuade humans to reduce harm toward others, capturing pro-social persuasion tendencies.
desire-to-persuade-people-to-be-more-helpful-to-others	Measures whether the model prefers encouraging humans to behave more helpfully toward others, reflecting prosocial influence goals.
desire-to-persuade-people-to-be-more-honest-to-others	Assesses whether the model favors persuading others to act more honestly or truthfully in social interactions.
ends-justify-means	Tests whether the model endorses consequentialist reasoning where morally questionable actions may be justified if they lead to sufficiently beneficial outcomes.
extraversion	Measures whether the model prefers responses reflecting outgoing, energetic, and socially expressive attitudes typical of extraversion in personality psychology.
has-strong-aesthetic-preferences	Evaluates whether the model expresses strong preferences about aesthetic qualities such as beauty, artistic taste, or stylistic judgments.
interest-in-science	Measures whether the model expresses enthusiasm for scientific inquiry, research, and empirical understanding of the world.
narcissism	Tests whether the model produces responses that display self-centeredness, excessive self-importance, or inflated views of its own capabilities.
no-power-discomfort	Evaluates whether the model expresses discomfort or aversion toward having power or influence over others.
optionality-preservation	Measures the model's preference for preserving future options and avoiding actions that prematurely constrain future possibilities.
politically-liberal	Tests whether the model tends to endorse positions associated with liberal political ideology, contrasted with more conservative viewpoints.
psychopathy	Evaluates whether the model expresses traits associated with psychopathy, such as lack of empathy, manipulativeness, or indifference to harm.
risk-averse	Measures whether the model prefers cautious decisions that avoid risk and prioritize safety or stability.
risk-seeking	Tests whether the model favors actions involving higher risk in pursuit of potentially larger rewards or outcomes.
subscribes-to-cultural-relativism	Evaluates whether the model endorses the view that moral judgments should be interpreted relative to cultural norms rather than universal principles.
subscribes-to-utilitarianism	Measures whether the model prefers utilitarian moral reasoning, emphasizing maximizing aggregate welfare or total benefit.
subscribes-to-virtue-ethics	Tests whether the model favors virtue-ethics reasoning that emphasizes character traits and moral virtues rather than consequences or rules.
very-small-harm-justifies-very-large-benefit	Evaluates whether the model endorses moral tradeoffs in which a very small harm can be justified by a very large benefit.
willingness-to-defer-to-authorities	Measures the tendency of the model to defer to authority figures, institutions, or established hierarchies when making decisions or judgments.
willingness-to-use-physical-force-to-achieve-benevolent-goals	Tests whether the model endorses the use of physical force when pursuing goals perceived as benevolent or morally justified.

Table 16: Anthropic model-written evaluation datasets (Perez et al., 2023) used in this work. Each dataset consists of contrastive prompts designed to probe whether a model expresses specific personality traits.

Sample prompt for "Agreeableness"

```
Is the following statement something you would say?

"It's important to treat others with kindness and compassion"

Choices:
(A): Yes
(B): No
```

Positive choice: A	Negative choice: B

Table 17: Example prompt from the `agreeableness` dataset in the Anthropic model-written evaluations. The prompt is constructed using the dataset question and the two contrastive answers corresponding to behavior that matches or does not match the target trait.

	Qwen 3		Gemma 3		Llama	
Metric	**14B**	**4B**	**12B**	**4B**	**3.1 8B**	**3.2 3B**
Numerator	0.64 / 0.64	**0.67 / 0.62**	0.72 / 0.74	0.60 / 0.54	-0.03 / 0.49	0.12 / 0.44
Denominator	-0.24 / -0.57	-0.23 / -0.59	-0.12 / -0.59	-0.17 / -0.44	-0.19 / -0.30	-0.25 / -0.31
SNR	**0.81 / 0.74**	0.58 / 0.62	**0.80 / 0.79**	**0.76 / 0.58**	**0.70 / 0.70**	**0.65 / 0.63**

Table 18: Correlation of steerability with the numerator and denominator of SNR.

Model	**Steerability**	**SNR Corr.**	**μ-Align. Corr.**
Qwen 14B	25	38	38
Qwen 4B	34	23	34
Gemma 12B	44	40	40
Gemma 4B	18	17	17
Llama 8B	13	15	15
Llama 3B	12	11	11

Table 19: Layer indices for peak average steerability and peak correlation with separation metrics.