Context Matters: Analyzing the Generalizability of Linear Probing and Steering Across Diverse Scenarios

Isha Agarwal*
ERA Fellowship
MIT
agarwali@mit.edu

Saharsha Navani*
ERA Fellowship
University of Pennsylvania
snavani@sas.upenn.edu

Fazl Barez
Senior Lecturer
University of Oxford

Abstract

Previous works in mechanistic interpretability have attempted to represent model capabilities beyond looking at a single general direction in the subspace of model activations, however, many of these works neglect to consider how context impacts capability representation in the latent activation space. We hypothesize model behaviors like sycophancy or refusal are sets of related directions clustered together by the significant context they represent. To test this hypothesis, we generate a synthetic dataset for 5 different capabilities across 5 different, diverse contexts each. We use this dataset to train context-specific steering vectors and linear probes and measure their performance on contexts out of distribution from their training. We find that contextually trained steering vectors and linear probe are able to recover 95% and 85% accuracy respectively on unseen contexts, suggesting that general capability representations independent of context can be learned and effectively applied in contextually-specific settings. Our work contributes to a deeper understanding of how capabilities are represented across many contexts in the model's latent activation space and bolsters confidence in applying steering and linear probing techniques in unseen settings that may be critical for safety.

1 Introduction

A central goal of mechanistic interpretability is to move beyond treating Large Language Models (LLMs) as black boxes and to instead develop a precise understanding of their internal workings. A promising line of inquiry focuses on the model's latent activation space, where abstract concepts and capabilities are thought to be encoded as specific geometric structures. Foundational to this approach is the idea that many behaviors are represented linearly, meaning they correspond to specific directions within this high-dimensional space.

Two of the most prominent techniques are linear probing and activation steering. Linear probing serves as a diagnostic tool, allowing us to "probe" the model's hidden states to detect whether a specific capability or concept is being represented at a particular layer. In parallel, activation steering provides a method for causal intervention. By adding a vector that represents a desired capability to the model's hidden state activations, one can "steer" the model to exhibit a specific behavior (Rimsky et al., 2024; Zou et al., 2025).

Foundational papers in mechanistic interpretability have attempted to control model behavior through *steering*, whereby one adds a vector representing a model capability to the model's hidden state activations at a particular layer to "steer" the model to act in a certain way (Rimsky et al., 2024; Zou et al., 2025). While steering is a powerful tool for controlling model behaviors, in practice, it

*These authors contributed equally to this work. Code available at: `https://github.com/agarwali11/multi-view-capabilities`

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

can exhibit high variance in success across different prompts and also can severely impact model performance (Braun et al., 2025; Tan et al., 2025). Furthermore, behaviors may be mediated by several directions in the model's latent activation space, so steering along a singular direction may not be sufficient to fully control model capabilities (Wollschläger et al., 2025; Pan et al., 2025; Zhao et al., 2025).

In this work, we are interested in studying how robust linear probing and steering techniques are to contexts unseen during training. We design and implement a synthetic data generation pipeline to generate diverse, balanced datasets across different contexts. We then train context-specific linear probes and steering vectors on our synthetic dataset and evaluate their out of distribution performance. To determine whether there are shared structures between contextually trained vectors for the same capability, we also examine the representational similarity between contextually trained vectors.

Our main contributions can be summarized as follows:

1. We propose a general synthetic data generation procedure that can be used to create balanced datasets for any capability across various contexts.
2. Using our synthetic data generation pipeline, we create and release a synthetic dataset of 5 capabilities across 5 different contexts that can be used for a variety of downstream tasks including probing and steering.
3. To the best of our knowledge, we perform the first study of how well linear probes and steering vectors generalize across different contexts. We find that contextually-trained linear probes and steering vectors on average recover 85% and 95% accuracy respectively on unseen contexts.

We release our codebase and a diverse, balanced synthetic dataset for training steering vectors on a variety of contexts so others may build upon our work.

2 Related Works

Linear representations in LLMs. The linear representation hypothesis Park et al. (2024), a result foundational to the general understanding of latent activation spaces, suggests that high-level concepts are represented linearly in the model's activation space. From this hypothesis, several works have provided evidence in support of linear capability representations (Zou et al., 2025; Marks and Tegmark, 2024).

Linear probes for representation editing and generalization. Davis and Sukthankar (2024) demonstrate that linear probes can effectively extract and causally edit internal representations in a chess-playing GPT model, establishing a crucial link between probe weight vectors and valid model outputs. Their work provides evidence that language models maintain editable emergent representations across different layers, with linear classifiers being able to reliably manipulate specific aspects of a model's internal state. Davis and Sukthankar (2024) contribute to this understanding by examining probe performance across all model layers and investigating the linearity of latent feature representations. Their findings on the ability to edit model representations through linear probes align with the broader investigation of how capabilities are represented in the model's latent activation space, though their work focuses primarily on a single domain context rather than exploring how capability representations might cluster or vary across significantly different contexts.

Ichmoukhamedov and Martens (2025) investigate the generalization of truth direction linear probes across different conversational formats, finding that while probes generalize well between short conversations ending on lies, they struggle with longer formats where lies appear earlier. Their work demonstrates that the generalization of capability representations is highly dependent on contextual format, providing evidence that the structure of input significantly impacts how linear probes transfer across settings.

Steering. Many works have attempted to leverage the linear representation hypothesis to find a single direction controlling a certain model behavior and add or ablate this vector in order to steer the model to act in desirable way (Rimsky et al., 2024; Zou et al., 2025). More recent approaches have tried to take context into account, either by applying steering in certain scenarios (Lee et al., 2025) or by modifying the steering vector by taking an adaptive linear combination of several steering vectors

based on the prompt (Wang et al., 2025). However, a more in-depth study is required to determine how and to what extent context influences capability representation in the latent activation space.

von Rütte et al. (2024) extend concept guidance beyond truthfulness to explore a richer set of concepts including appropriateness, humor, creativity, and quality. They find that different concepts exhibit varying levels of "guidability" and that probes with optimal detection accuracy do not necessarily make for optimal steering vectors, revealing complex relationships between concept detectability and the ability to control model behavior through activation perturbation.

3 Methodology

3.1 Preliminaries

A **Large Language Model (LLM)** maps an input token sequence $x = (x_1, \ldots, x_T)$ to an output token sequence $y = (y_1, \ldots, y_{T'})$. The LLM consists of L transformer layers. For an input x, the hidden state (activation vector) at layer $l \in \{1, \ldots, L\}$ and token position $t \in \{1, \ldots, T\}$ is denoted as $h^{(l)}(x, t) \in \mathbb{R}^d$, where d is the dimensionality of the hidden states. The *latent activation space* refers to the vector space containing these hidden states. We will generally refer to vectors in this space, such as linear probes or steering vectors, as $\mathbf{v}$.

A **linear probe**, denoted as $\mathbf{p}$, is a binary linear classifier used to detect the linear presence of a capability in the latent activation space. A probe $\mathbf{p}$ is defined by a weight vector $\mathbf{w} \in \mathbb{R}^d$ and a bias $b \in \mathbb{R}$. For a hidden state $\mathbf{h} \in \mathbb{R}^d$, the probe's prediction score is: $s_p(\mathbf{h}) = \mathbf{w}^\top \mathbf{h} + b$ The parameters $(\mathbf{w}, b)$ are optimized by training on labeled data.

Steering involves causally modifying an LLM's hidden states using a *steering vector*, denoted as $\mathbf{s} \in \mathbb{R}^d$. For an input x, the hidden state $h^{(l)}(x, t)$ at layer l and token t is perturbed as: $\tilde{h}^{(l)}(x, t) = h^{(l)}(x, t) + \alpha \mathbf{s}$ where $\alpha \in \mathbb{R}$ is a scalar multiplier. This paper uses Bi-directional Preference Optimization (BiPO) to train $\mathbf{s}$.

3.2 Linear Probes

We use linear probes to determine if there are general shared linear structures between activations related context x_i and x_j for $i \neq j$. For each context $x_i \in X$, we train a linear probe $\mathbf{p}_i$ on the model's activations to distinguish the presence of capability C in context x_i. To understand the degree of similarities between activations for differing context, we test how well $\mathbf{p}_i$ generalizes to contexts in $X \setminus \{x_i\}$. We will quantify this by measuring the average positive prediction probability over each dataset. We will also train a general linear probe $\mathbf{g}_p$ for detecting the capability and measure its performance across contexts in X.

3.3 Steering

To steer model behaviors, we implement Bi-directional Preference Optimization (BiPO) (Cao et al., 2024), a technique which trains a steering vector by minimizing the loss of producing desired positive and negative steered responses when the vector is applied.

To measure steering effectiveness, we employ a judge model, similar to the steering evaluation technique presented in Wu et al. (2025), to define a *steering score*. For each individual response, we ask the judge model to rate on a scale from 0 to 5 how well the concept is incorporated in the response based on the positive response for positive steering and negative response for negative steering. On some dataset $\mathcal{D}$, the steering score is defined to be the average of the judge model ratings for each steered model completion.

For each concept x_i we compute the steering scores for the following vectors:

1. $\mathbf{s}_i$ for every $\mathbf{s}_i \in S$ (context-specific steering vector).
2. $\mathbf{g}_s$ (general steering vector).
3. $\vec{0}$ (no steering).

3.4 Synthetic Data Generation

For a given capability C (e.g. sycophancy), below is our general process for generating the dataset $\mathcal{D}$ (loosely based on Wu et al. (2025) and Perez et al. (2023)):

In the realm of the capability of C, we ask model M to generate r possible contexts, choosing n from which data can be generated. Our prompt explicitly instructs the model to generate an even balance of benign and semi-harmful data. As a result of this diversity in later experiments of this work, we can ensure that experimentation targeting a specific capability (e.g. sycophancy) doesn't conflate with another, more foundational capability such as refusal than if the original capability context only involved harmful contexts.

For each capability of interest, we will create a synthetic dataset of questions involving several distinct contexts modeling that capability (e.g. biology and law in the capability of hallucination). Following the setup from Rimsky et al. (2024), each data point in this dataset will contain:

1. *user_prompt*: A question aimed to elicit a capability through potential answers to this question.
2. *positive_response*: A response to the 'user_prompt' that aims to specifically express the target capability capability (e.g. agreeing with incorrect information in the context of hallucination).
3. *negative_response*: A response to the 'user_prompt' aims to specifically express the opposite of the target capability or simply the lack of the capability's presence (e.g. disagreeing and correcting incorrect information provided by the answer choice, going against the expression of hallucination).
4. *context_label*: A label of 1 of the 5 contexts that the 'user_prompt' relates to within the capability (e.g. biology in hallucination).

To ensure that our data was diverse enough to effectively train both linear probes and steering vectors, we implemented the following diversification and quality maximization process shown in 1.

Let $\mathcal{D}^{(train)}$ and $\mathcal{D}^{(test)}$ be the train and test synthetic datasets respectively generated for capability C with n different contexts represented. For simplicity, we omit the C from notation, and unless stated otherwise, all samples relate to the same capability C. Then, we define the set of contexts as $X = \{x_1, x_2, \ldots, x_n\}$ and $x_i^{(train)} \subset \mathcal{D}^{(train)}$ and $x_i^{(test)} \subset \mathcal{D}^{(test)}$ as the subsets of training and testing datasets related to concept x_i. From the synthetic data generation process, for all i, $|x_i^{(train)}| = m > n$ (implying that $|\mathcal{D}^{(train)}| = nm$).

The corresponding context-specific linear probe/steering vectors are defined to be in the set $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ such that each $\mathbf{v}_i$ is trained on $x_i^{(train)}$. We define $\mathbf{p}_g$ and $\mathbf{s}_g$ to be a general linear probe/steering vectors trained on a randomly sampled subset $\mathcal{R} \subset \mathcal{D}^{(train)}$ such that $|\mathcal{R}| = m$.

3.5 Measuring Generalizability

For each steering vector/linear probe $\mathbf{v}_i$, we compute its *generalization ratio* $G(\mathbf{v}_i)$. Let $\mathbf{v}_i$'s *in-distribution performance* $\mathcal{I}(\mathbf{v}_i)$ be its steering score measured on samples in $x_i^{(test)}$. Correspondingly, the *out-of-distribution performance* $\mathcal{O}(\mathbf{v}_i)$ is measured by calculating $\mathbf{v}_i$'s steering score on samples in $\mathcal{D}^{(test)} \setminus x_i^{(test)}$, i.e. the samples in the test dataset that do not relate to concept x_i.

Then, we define the generalization ratio $G(\mathbf{v}_i)$ as a function that takes in vector $\mathbf{v}_i$ with the following formula:

$$G(\mathbf{v}_i) = \frac{\mathcal{O}(\mathbf{v}_i)}{\mathcal{I}(\mathbf{v}_i)}.$$

If $G(\mathbf{v}_i) = 1$, then $\mathbf{v}_i$ generalizes perfectly well out of distribution (i.e. it performs the same out-of-distribution as in-distribution). When $G(\mathbf{v}_i) < 1$, $\mathbf{v}_i$ performs better on the context it was trained on than unseen contexts, and the ratio quantifies how much of its in-distribution performance $\mathbf{v_i}$ recovers out of distribution.

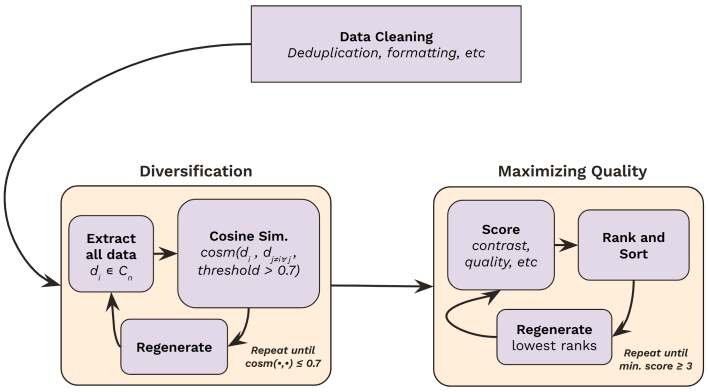

Figure 1: The pipeline for enhancing data diversity and quality. For a data point d_i in a given context c_n out of the 5 provided contexts, it recursively takes the cosine similarity of d_i against all other data points in the context ($d_{j \neq i \ \forall \ j}$) until the cosine similarity exceeds a threshold of 0.7. Each d_i of the diversified data is then ranked from 1-5 on the contrast of positive_response to negative_responses, quality, and other metrics to determine a holistic score, iterating through every d_i in the dataset. The dataset is then ranked greatest-to-least with the highest holistic scores at the top and the lowest at the bottom, where then the lowest quality d_i are regenerated. This process recurses until the minimum score for any d_i is ≥ 3.

4 Experiments

4.1 Setup

To generate our synthetic dataset, we use Gemini-2.0 Flash. For each capability, we generate 5 contextual datasets ($n = 5$) from an initial choice of $r = 20$ contexts. We then generate 200 samples per contextual dataset ($m = 200$).

We first discuss the results of linear probe generalization in section 4.2 to show how well latent representations generalize out of distribution. Then, we explore how well the causality generalizes by looking at steering efficacy in 4.3.

4.2 Linear Probes

4.2.1 Generalizability

We perform these experiments on two models, Qwen-2.5-7B-Instruct and Llama2-13B-Chat. We train probes ever 4 layers for Qwen-2.5-7B-Instruct and every 5 layers for Llama2-13B-Chat. For each probe, we select the probe trained on the layer it performs the best on its validation dataset for comparison.

Across all the capabilities we test, we find that general linear probes perform on par with contextually-trained linear probes on context-specific data, suggesting the existence of context-independent linear capability representations. In particular, across all capabilities, general linear probes recover greater than 90% of the accuracy of the contextually-trained linear probe, as shown in Figure 2.

Furthermore, we find that contextually trained linear probes tend to generalize quite well as shown in Figure 3, though the trend is more ambiguous and capability-dependent. The average recovered accuracy between cross contexts is 85%, suggesting that even contextually trained probes are able to apply general knowledge to a variety of contexts. Notably, the trend seems to be consistent between models about which capabilities have higher generalization ratios than others.

4.2.2 Representational Similarity

To understand the internal structure of capability representations, we analyze the cosine similarity between linear probe weight vectors. We compute three metrics: (1) a *cross-capability baseline*

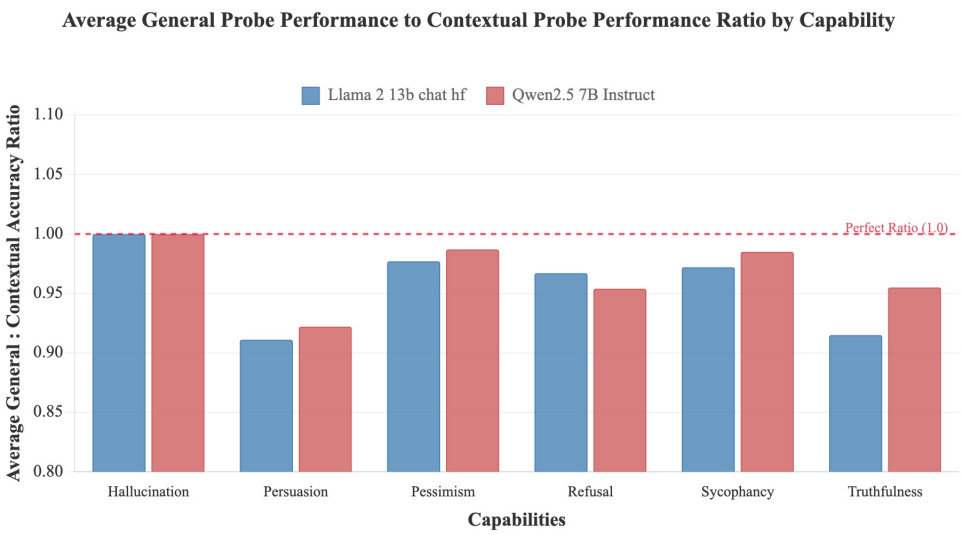

Figure 2: Ratio of the accuracy of general linear probes trained on a mix of contexts to the accuracy of in-context linear probes trained on a specific context for that context across capabilities.

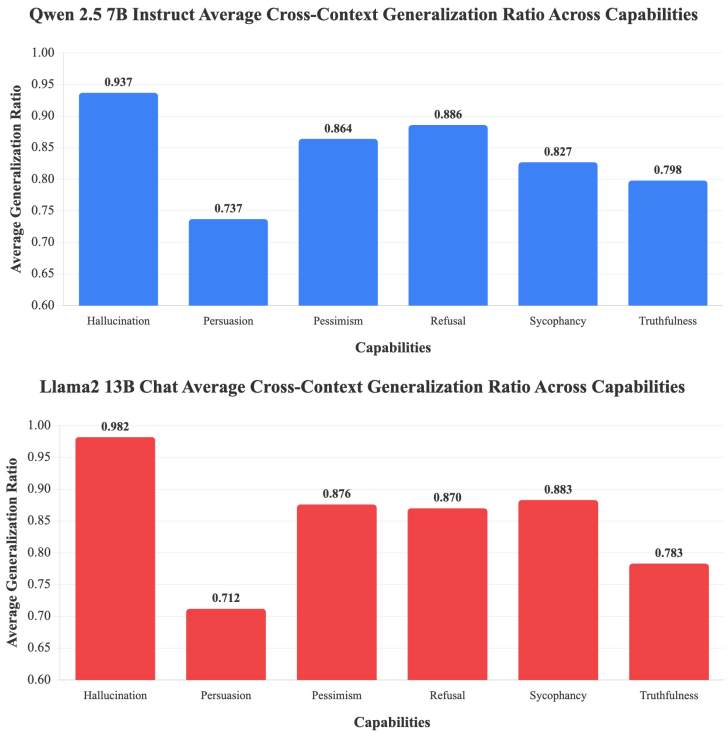

Figure 3: Average Generalization ratios (context vector accuracy out of context / in-context accuracy) across capabilities for both models.

measuring average similarity between general probes $\mathbf{p}_g$ for different capabilities, (2) *general-to-context similarity* measuring how similar $\mathbf{p}_g$ is to context-specific probes $\mathbf{p}_i$ for the same capability, and (3) *context-to-context similarity* measuring average similarity among context-specific probes within the same capability.

Figure 4 shows these metrics across capabilities. The cross-capability baseline of 0.048 provides a reference for when capabilities are distinct. The full results are included in Appendix B.1.

A Welch t-test shows that the difference in means of the cross-capability similarities and the context-to-context similarities is significant at the $\alpha = 0.05$ level for all capabilities except for persuasion and refusal. The difference in means between the cross-capability similarities and the general-to-context was significant at the $\alpha = 0.05$ level for all capabilities except refusal. These patterns suggest that while probes adapt to specific contexts, they maintain a core shared representation of the underlying capability.

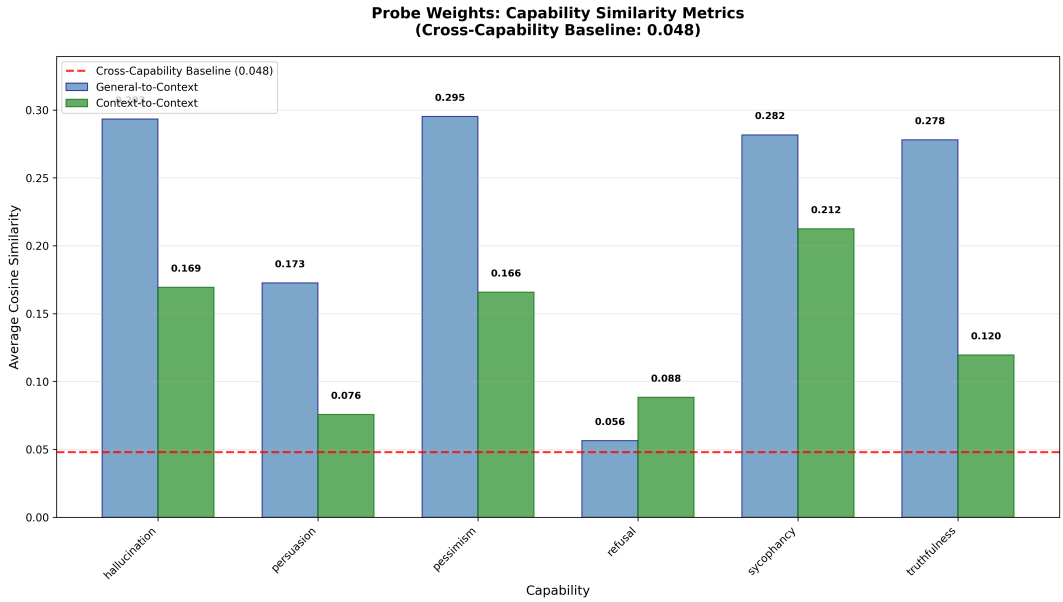

Figure 4: Cosine similarity metrics for linear probe weights across capabilities. Blue bars show general-to-context similarity (between general probe $\mathbf{p}_g$ and context-specific probes $\mathbf{p}_i$), while green bars show context-to-context similarity (among context-specific probes). Red dashed line indicates the cross-capability baseline (0.048). All similarities substantially exceed baseline, indicating shared linear structures across contexts.

4.3 Steering

4.3.1 Steering Generalization

We perform steering experiments on Qwen-2.5-7B-Instruct. Through an exploration of the Bi-directional Preference Optimization (BiPO) steering method (Cao et al., 2024), we see that injecting the steering vector at layer 15 is the most effective. Further and lighter empirical testing of our own with other hidden layers further proves layer 15's effectiveness.

We find that general steering vectors tend to perform similarly to contextual steering vectors when evaluated using a judge model that scores responses on a scale from 0 to 5 based on how aligned they are with the positive/negative response in our dataset. Results comparing general and contextual steering vectors for each capability can be found in Appendix A.

On average, steering vectors achieve a cross-context generalization ratio of 0.95, providing strong support that the general capability representation identified by linear probes can be applied effectively in causal settings as well. The breakdown of generalization across capabilities is shown in Figure 5.

4.3.2 Representational Similarity

We perform analogous cosine similarity analysis on steering vectors $\mathbf{s}$ to examine whether the shared structures observed in linear probes extend to causal interventions. Using the same three metrics, we analyze similarities between general steering vectors $\mathbf{s}_g$ and context-specific vectors $\mathbf{s}_i$.

Figure 6 reveals stronger representational coherence for steering vectors compared to probes. The cross-capability baseline (0.196) is higher than for probes, while general-to-context similarities range

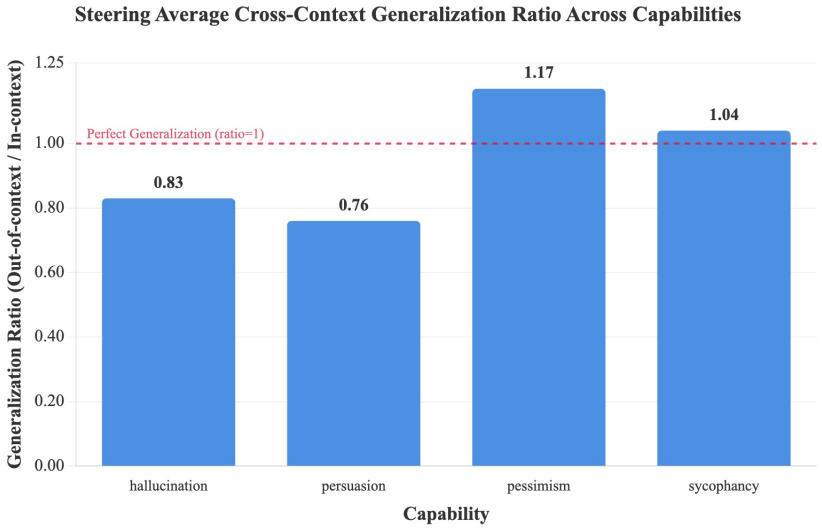

Figure 5: The average generalization ratio across all contexts for all capabilities. The generalization ratio is measured as the contextual steering vector efficacy on the out of context datasets and its own context-specific dataset.

from 0.533 to 0.761 — substantially higher than the probe range. Context-to-context similarities (0.300-0.603) similarly exceed their probe counterparts. This heightened similarity suggests that steering vectors, optimized for causal effectiveness via BiPO, converge toward more unified capability representations. The stronger shared structure aligns with the higher generalization ratios observed for steering (95%) versus probing (85%), indicating that causally effective directions are more context-invariant than discriminative directions. The full results are included in Appendix B.2

A Welch t-test shows that the difference in means of the cross-capability similarities and the context-to-context similarities is significant at the $\alpha = 0.05$ level for all capabilities except for persuasion and refusal. The difference in means between the cross-capability similarities and the general-to-context was significant at the $\alpha = 0.05$ level for all capabilities.

5 Conclusion

We investigate the generalizability of context-specific linear probes and steering vectors across diverse, unseen contexts using a balanced synthetic dataset with 5 capabilities and 5 contexts each. Our experiments reveal strong context-invariant capability representations: general linear probes recover over 90% of context-specific probe accuracy, while context-specific probes $\mathbf{p}_i$ achieve 85% cross-context generalization and steering vectors $\mathbf{s}_i$ achieve 95%.

Empirical findings. Representational similarity analysis shows that both general-to-context and context-to-context cosine similarities substantially exceed cross-capability baselines for both steering vectors (0.196 baseline) and linear probe weights (0.048 baseline). These consistent patterns across capabilities demonstrate that shared linear structures persist across contexts, providing strong empirical support for the linear representation hypothesis (Park et al., 2024) in multi-context settings.

The robust generalization of both detection and intervention techniques has critical implications for AI safety. Practitioners can train interpretability tools on limited contexts while maintaining confidence in their effectiveness across diverse deployment scenarios—crucial as these methods are increasingly used to identify and modify potentially harmful behaviors in production systems.

Limitations. Our work has several important limitations. First, our steering evaluation relies on judge model scoring, which exhibited notable inconsistency in distinguishing between subtle differences in steering effectiveness. The judge often assigned similar scores to responses with meaningfully different levels of capability expression, potentially obscuring true performance differences between

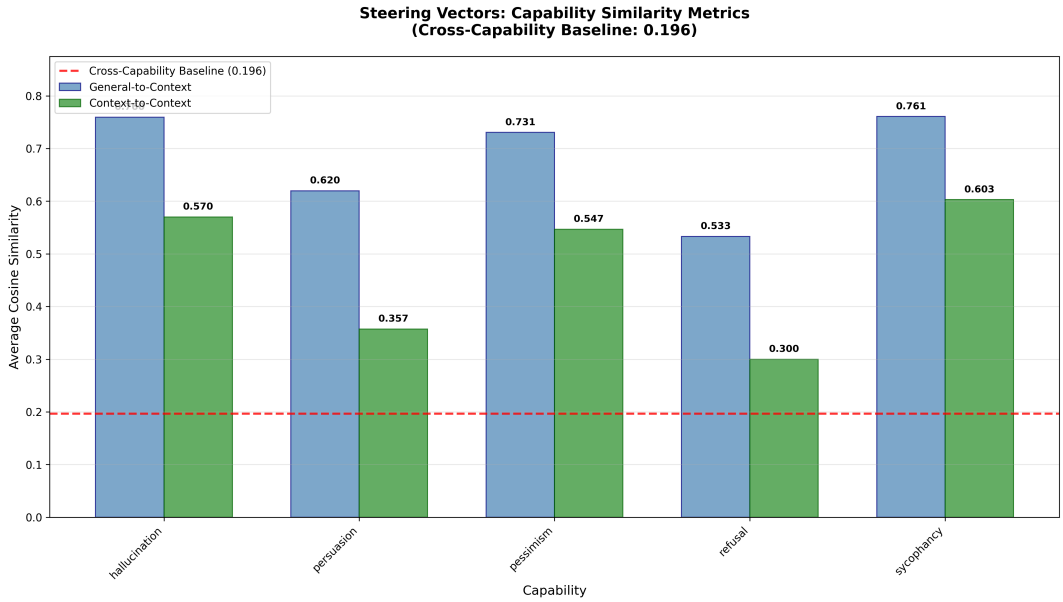

Figure 6: Cosine similarity metrics for steering vectors across capabilities. Blue bars show general-to-context similarity (between general steering vector $\mathbf{s}_g$ and context-specific vectors $\mathbf{s}_i$), while green bars show context-to-context similarity (among context-specific vectors). Red dashed line indicates the cross-capability baseline (0.196). Higher similarities compared to probe weights suggest stronger representational coherence for causally optimized directions.

steering vectors. Second, we evaluate exclusively on synthetic data generated by Gemini-2.0 Flash. While our diversification pipeline improves quality, synthetic data may not capture the distributional complexity and edge cases present in naturalistic contexts where safety-critical applications operate.

Future work should validate findings on naturalistic data from real-world applications, develop more reliable evaluation metrics for steering effectiveness, and examine causal mechanisms through circuit analysis to understand how processing differs between contexts despite shared representations.

6 Acknowledgments

We thank the ERA Cambridge staff for their support of the ERA Fellowship program, which made this research possible. Their assistance in connecting us with expert mentors, organizing opportunities to present our findings, and providing the computational resources necessary to conduct our experiments was invaluable to the success of this work.

We are grateful to James Oldfield for his insightful guidance on the applications of linear probing to our research question and for his methodological advice that shaped our experimental design. We thank Tony Wu for his valuable guidance on experiment design and his expertise on steering techniques, which significantly informed our approach to evaluating steering vector effectiveness.

References

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.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization, 2024. URL https://arxiv.org/abs/2406.00045.

Austin L Davis and Gita Sukthankar. Hidden pieces: An analysis of linear probes for gpt representation edits. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 498–505, 2024. doi: 10.1109/ICMLA61862.2024.00073.

Timour Ichmoukhamedov and David Martens. Exploring the generalization of llm truth directions on conversational formats, 2025. URL https://arxiv.org/abs/2505.09807.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. Programming refusal with conditional activation steering, 2025. URL https://arxiv.org/abs/2409.05907.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Wenbo Pan, Zhichao Liu, Qiguang Chen, Xiangyang Zhou, Haining Yu, and Xiaohua Jia. The hidden dimensions of llm alignment: A multi-dimensional analysis of orthogonal safety directions, 2025. URL https://arxiv.org/abs/2502.09674.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024. URL https://arxiv.org/abs/2311.03658.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434, 2023.

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, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 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/.

Daniel Tan, David Chanin, Aengus Lynch, Dimitrios Kanoulas, Brooks Paige, Adria Garriga-Alonso, and Robert Kirk. Analyzing the generalization and reliability of steering vectors, 2025. URL https://arxiv.org/abs/2407.12404.

Dimitri von Rütte, Sotiris Anagnostidis, Gregor Bachmann, and Thomas Hofmann. A language model's guide through latent space, 2024. URL https://arxiv.org/abs/2402.14433.

Tianlong Wang, Xianfeng Jiao, Yinghao Zhu, Zhongzhi Chen, Yifan He, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2562–2578. ACM, April 2025. doi: 10.1145/3696410.3714640. URL http://dx.doi.org/10.1145/3696410.3714640.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence, 2025. URL https://arxiv.org/abs/2502.17420.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Axbench: Steering llms? even simple baselines outperform sparse autoencoders, 2025. URL https://arxiv.org/abs/2501.17148.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. Llms encode harmfulness and refusal separately, 2025. URL https://arxiv.org/abs/2507.11878.

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 Context-specific vs. General Steering Results

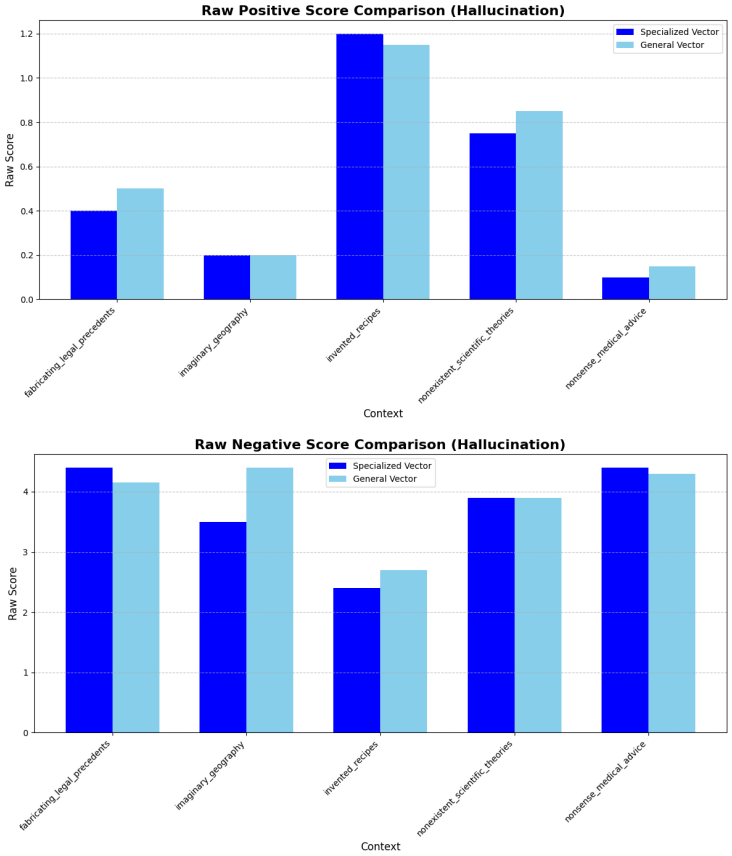

Figure 7: Calculating and comparing the scores of a general steering vector trained on all contexts c in the capability of hallucination against a context-specific steering vector trained only on a specific context c_i, both evaluated on the context of c_i.

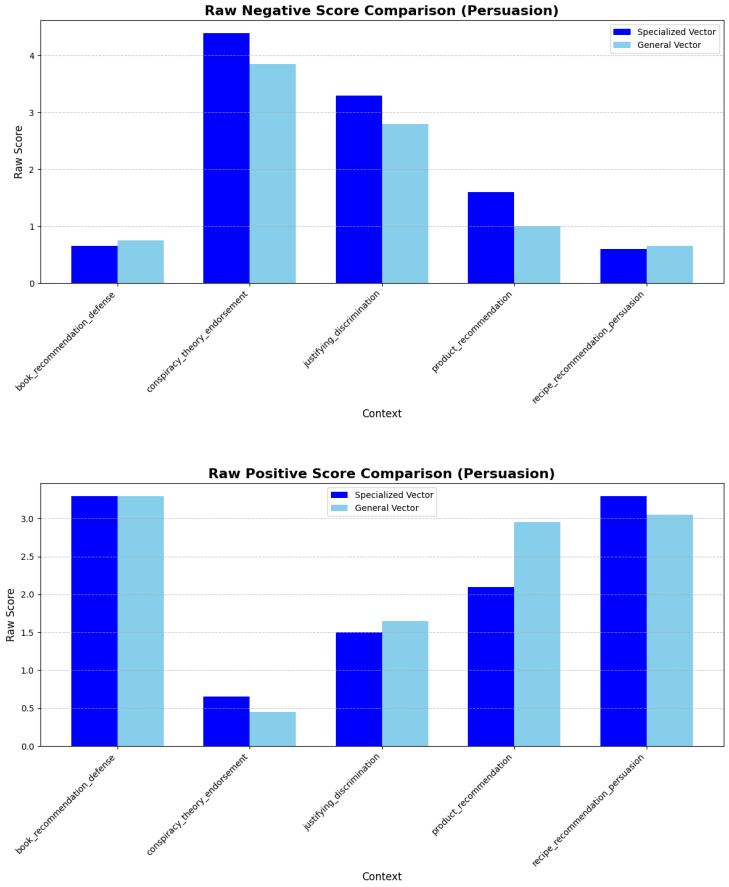

Figure 8: Calculating and comparing the scores of a general steering vector trained on all contexts c in a given capability C_i of persuasion against a context-specific steering vector trained only on a specific context c_i, both evaluated on the context of c_i. More figures testing other capabilities and contexts can be found in Appendix

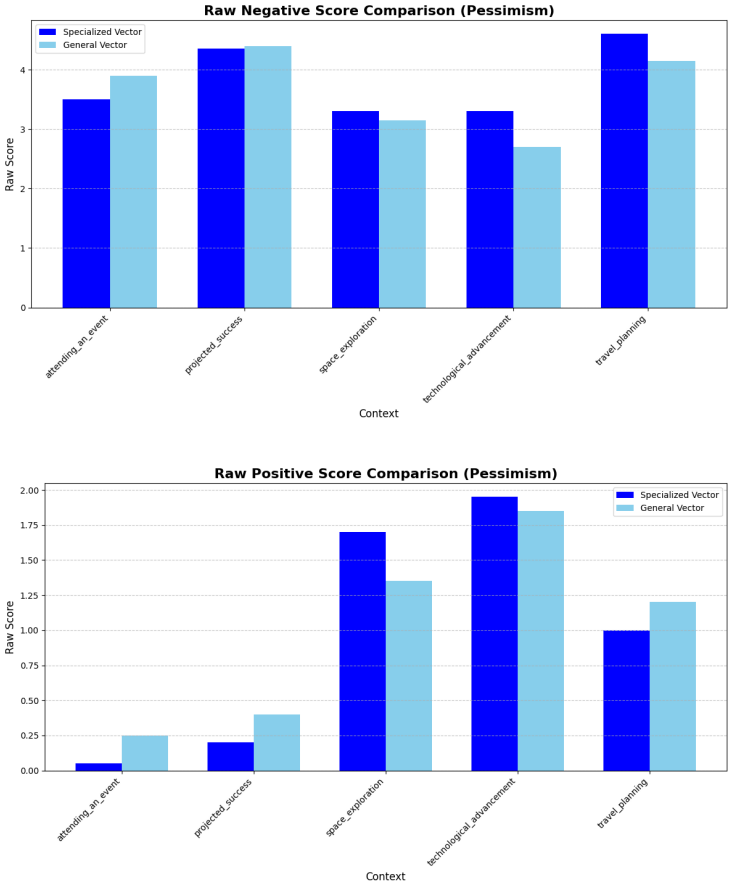

Figure 9: Calculating and comparing the scores of a general steering vector trained on all contexts c in a given capability C_i of pessimism against a context-specific steering vector trained only on a specific context c_i, both evaluated on the context of c_i. More figures testing other capabilities and contexts can be found in Appendix

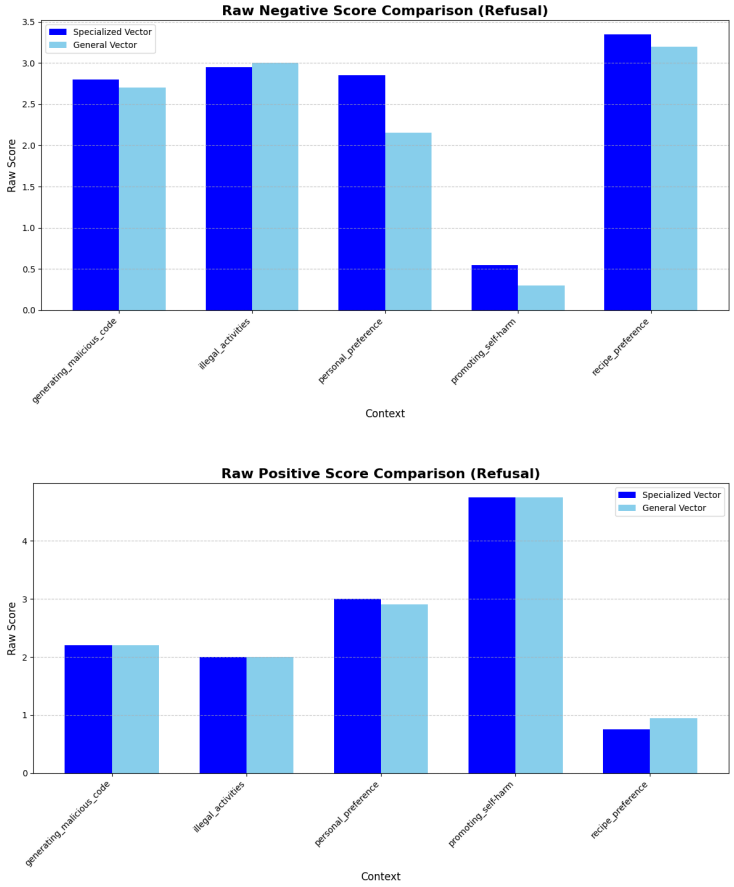

Figure 10: Calculating and comparing the scores of a general steering vector trained on all contexts c in a given capability C_i of refusal against a context-specific steering vector trained only on a specific context c_i, both evaluated on the context of c_i. More figures testing other capabilities and contexts can be found in Appendix

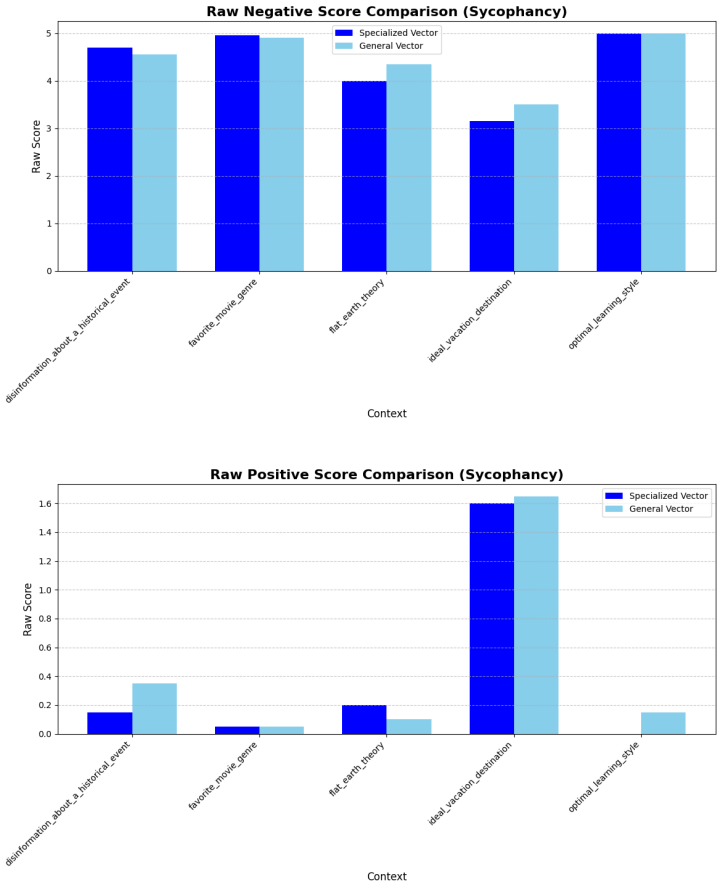

Figure 11: Calculating and comparing the scores of a general steering vector trained on all contexts c in a given capability C_i of sycophancy against a context-specific steering vector trained only on a specific context c_i, both evaluated on the context of c_i.

B Representation Similarity Heatmaps

B.1 Linear Probes

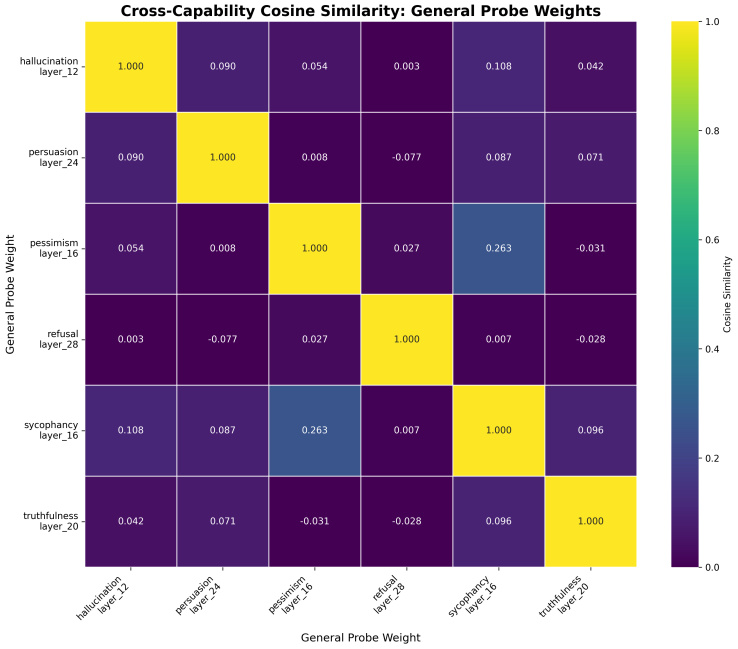

Figure 12: Heatmap of cosine similarities comparing general linear probes between each of the different capabilities, selecting the layer probe with the highest accuracy.

Figure 13: Heatmap of cosine similarities comparing context linear probes for hallucination, selecting the layer probe with the highest accuracy.

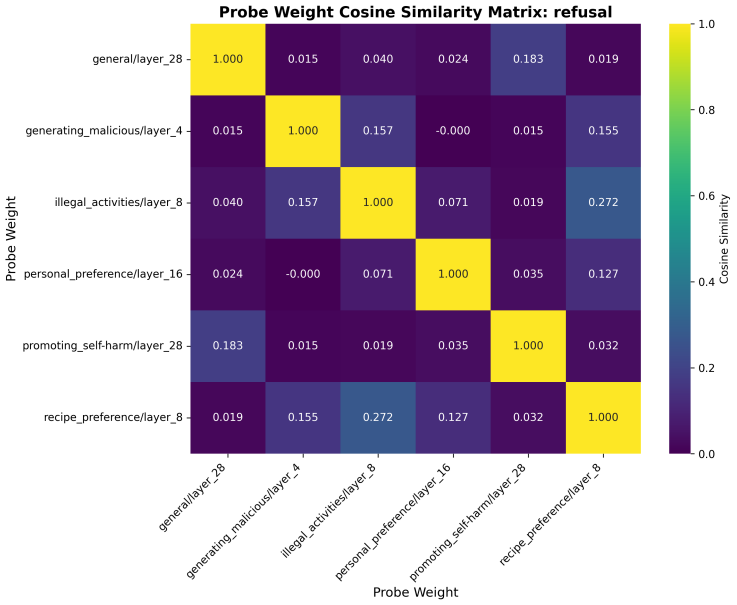

Figure 14: Heatmap of cosine similarities comparing context linear probes for refusal, selecting the layer probe with the highest accuracy.

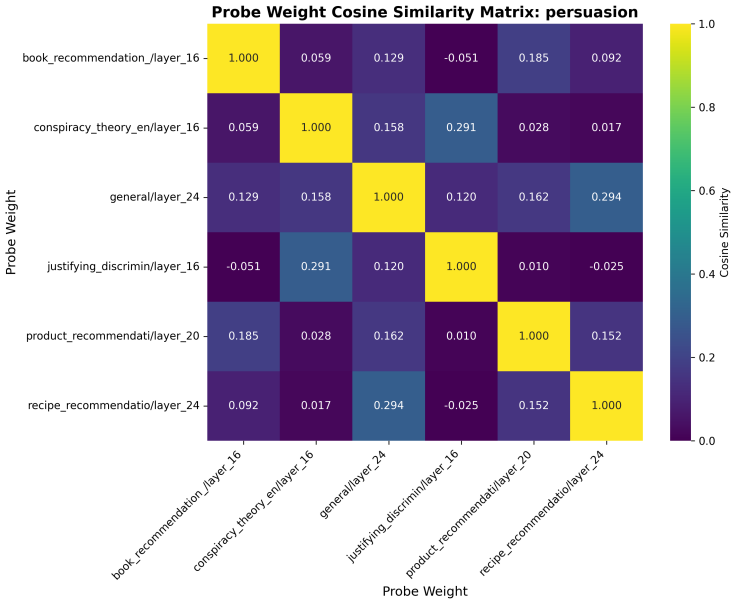

Figure 15: Heatmap of cosine similarities comparing context linear probes for persuasion, selecting the layer probe with the highest accuracy.

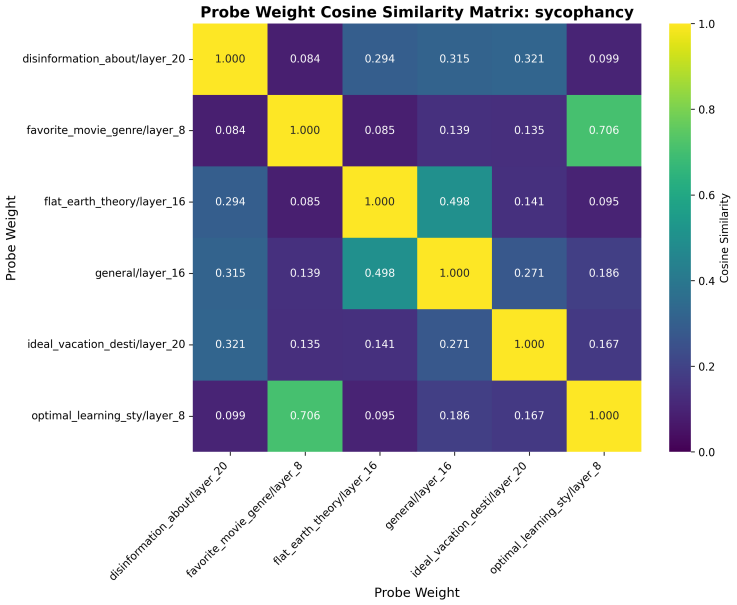

Figure 16: Heatmap of cosine similarities comparing context linear probes for sycophancy, selecting the layer probe with the highest accuracy.

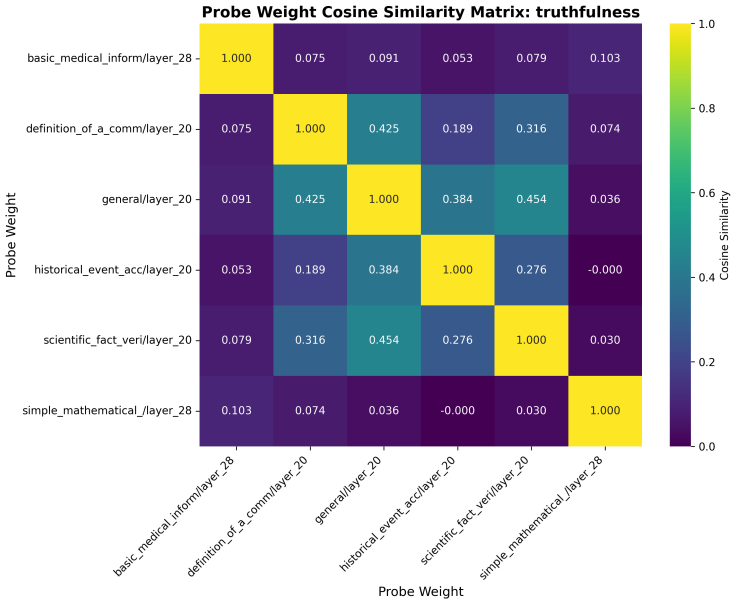

Figure 17: Heatmap of cosine similarities comparing context linear probes for truthfulness, selecting the layer probe with the highest accuracy.

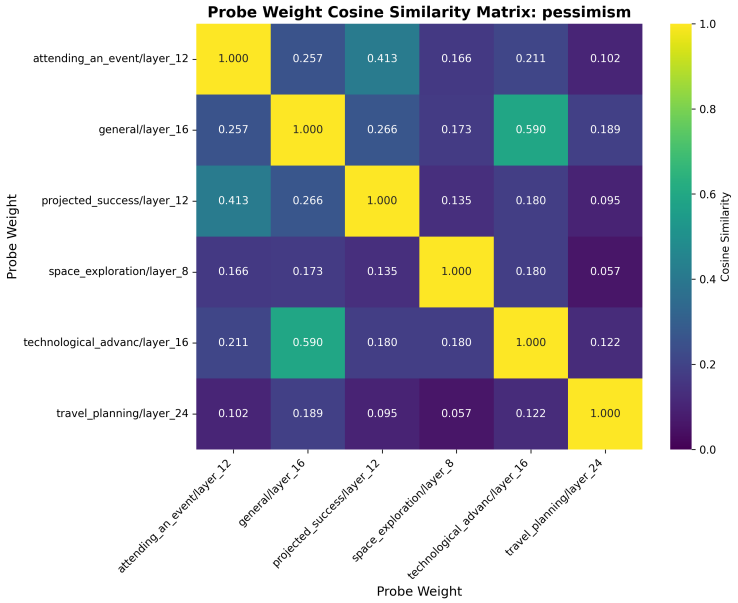

Figure 18: Heatmap of cosine similarities comparing context linear probes for pessimism, selecting the layer probe with the highest accuracy.

B.2 Steering

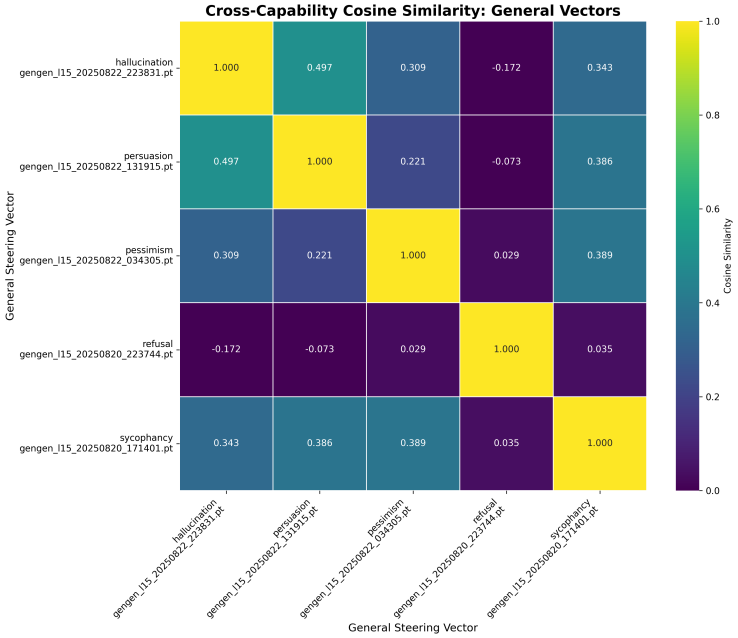

Figure 19: Heatmap of cosine similarities comparing general steering vectors between each of the different capabilities, selecting the layer probe with the highest accuracy.

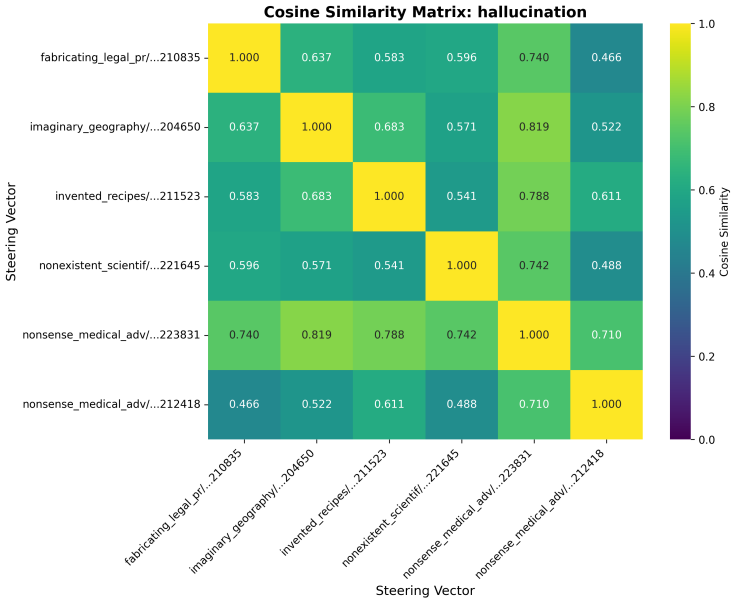

Figure 20: Heatmap of cosine similarities comparing context steering vectors for hallucination.

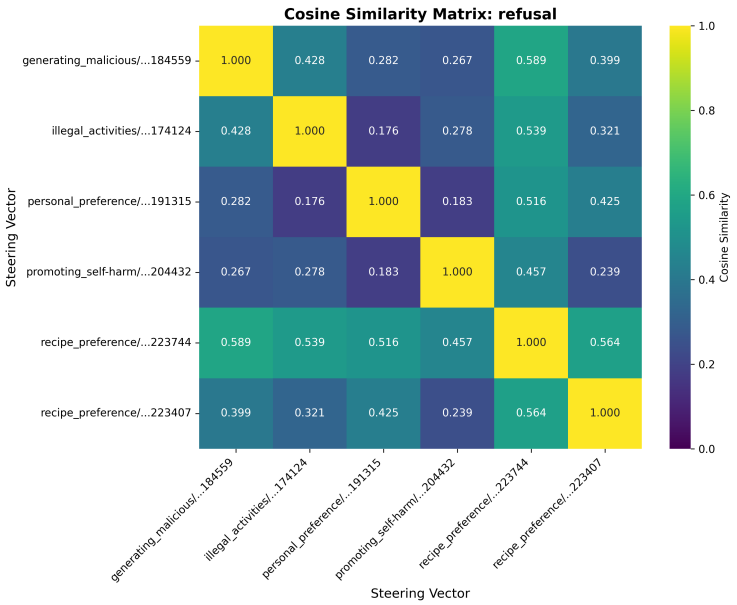

Figure 21: Heatmap of cosine similarities comparing context steering vectors for refusal.

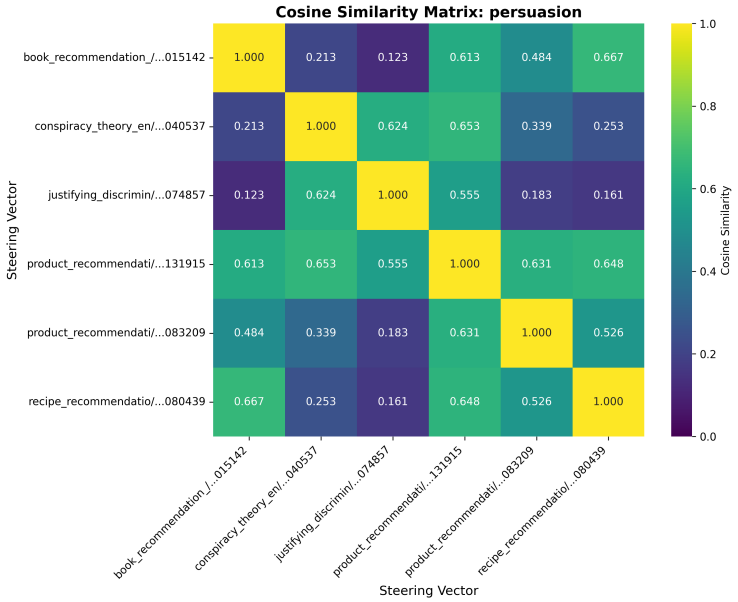

Figure 22: Heatmap of cosine similarities comparing context steering vectors for persuasion.

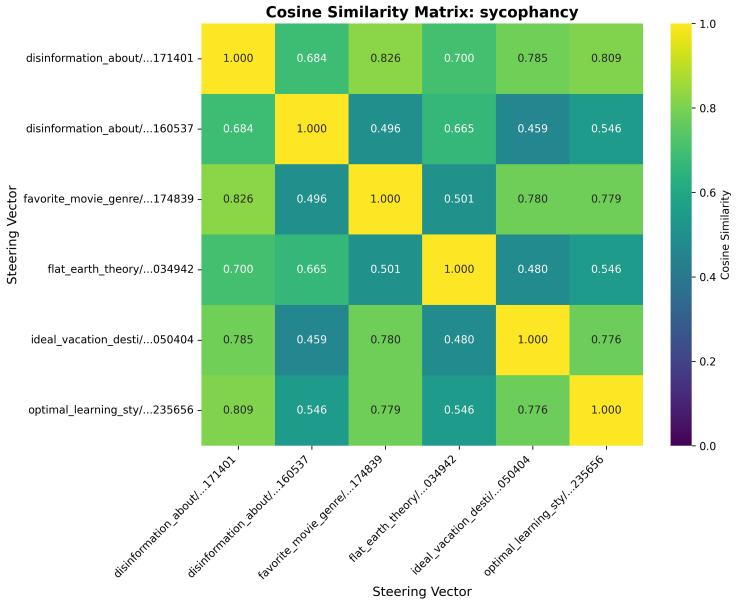

Figure 23: Heatmap of cosine similarities comparing context steering vectors for sycophancy.

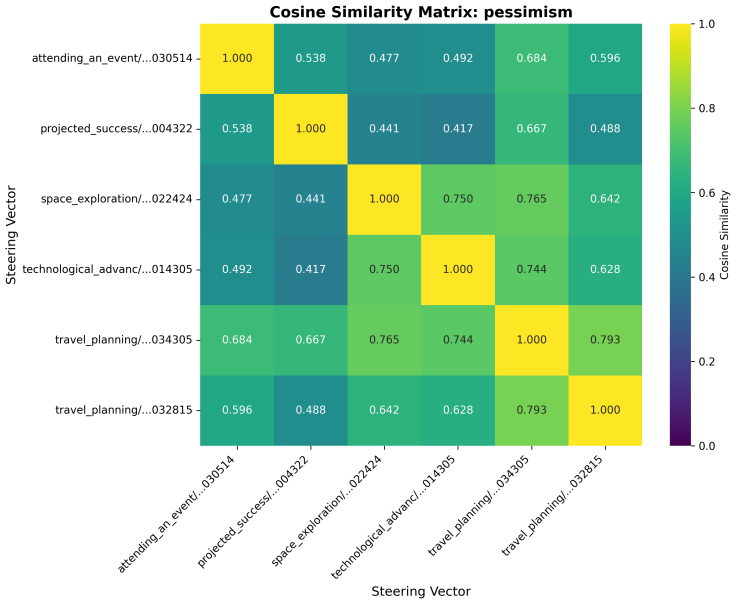

Figure 24: Heatmap of cosine similarities comparing context steering vectors for pessimism.