Probing by Analogy: Decomposing Probes into Activations for Better Interpretability and Inter-Model Generalization

Anonymous Author(s)
Affiliation
Address
email

Abstract

Linear probes have been used to demonstrate that LLM activations linearly encode high-level properties of the input, such as truthfulness, and that these directions can evolve significantly during fine-tuning and training. However, despite their seeming simplicity, linear probes can have complex geometric interpretations, leverage spurious correlations, and lack selectivity. We present a method for decomposing linear probe directions into weighted sums of as few as 10 model activations, whilst maintaining task performance. These probes are also invariant to affine transformations of the representation space, and we demonstrate that, in some cases, poor base to fine-tune probe generalization performance is partially due to simple transformations of representation subspaces, and the structure of the representation space changes less than indicated by other methods. Anonymized code is available here.

1 Introduction

Linear probes, simple classifiers trained on the activations of frozen models, are a key tool for interpreting neural networks. Early work found that linear probes can recover high-level features of the input from model activations [Alain and Bengio, 2017], for example refusal [Arditi et al., 2024], sentiment [Tigges et al., 2024], spatial and temporal relationships [Gurnee and Tegmark, 2023], and truthfulness [Marks and Tegmark, 2023]. Applying linear probes at different checkpoints during training [Qian et al., 2024] and between base and fine-tuned models [Du et al., 2025, Mosbach et al., 2020] has revealed how model representations change during optimisation. They have also been used to show that models know when they are being tested [Nguyen et al., 2025], and predict when models will produce jail-broken answers [Zou et al., 2023], provide harmful information to a malicious user [Roger et al., 2023], or output falsehoods [Orgad et al., 2024].

Given the importance of linear probes in understanding and monitoring model behavior, we would want them to be reliable and interpretable. However, they can leverage spurious correlations in their training dataset and even overfit in high dimensions [Belinkov, 2022]. There is also limited empirical research into the features they capture [Kunz and Kuhlmann, 2020, Choi et al., 2024, Sharkey et al., 2025].

In this work we introduce a method for decomposing linear probes called analogous probing, that is based on the MetaSAEs method for decomposing sparse autoencoder (SAE) decoder matrices [Leask et al., 2025a], and the Inference-Time Decomposition of Activations (ITDA) method for decomposing unseen model activations into a dictionary of sampled model activations in a sparse dictionary learning setting [Leask et al., 2025b]. These decompositions are based on relative representation methods [Moschella et al., 2022], which are invariant to linear transformations. We use this property

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

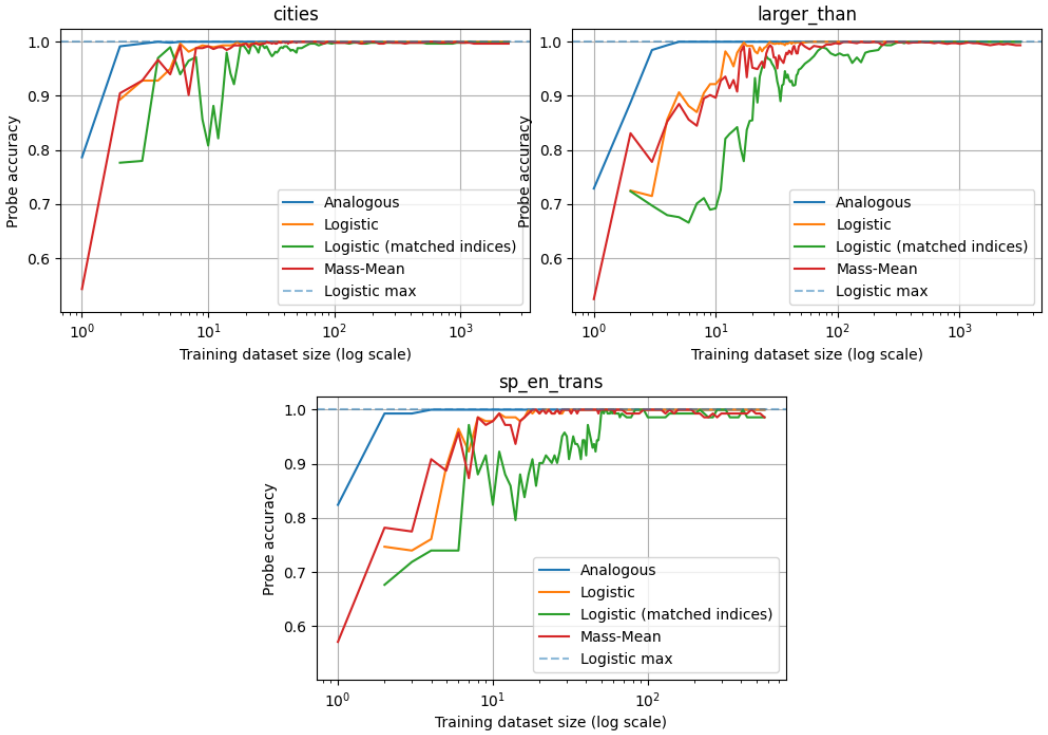

Figure 1: The performance of probes when constructed from a fixed number of training samples, for each of the three datasets, averaged over models. We trained mass-mean probes and logistic probes on 30 random subsets of the dataset and selected the best performing of those at each sparsity level. We also trained a logistic probe using the deterministically selected examples that are used in the analogous probe decomposition. The dashed line represents the performance achieved by a logistic regressor on the full training dataset. See Section C.3 for details of the probing methods.

to demonstrate that some of the difference in probe performance when transferring probes between base and fine-tuned models, and different pretraining checkpoints, is due to linear transformations of the representation space. This gives new insight into how the model representation space develops throughout training, and suggests the changes are less significant than previously thought.

Analogous probes have a simple geometric interpretation, similar to activation engineering [Zou et al., 2023] approaches to probing, such as mass-mean probing [Marks and Tegmark, 2023], whilst achieving performance similar to trained linear probes. Simpler explanations are generally preferred in the interpretability literature, and we propose that these sparse probe decompositions may be useful for constructing interpretable probes. We provide examples of decompositions, but leave proper interpretability experiments to future work (see Section E).

See Appendix A for a detailed literature review of sparse dictionary learning, probing, relative representation similarity methods, and activation engineering.

2 Analogous Probing

Let LM be a frozen language model with L transformer layers. For an input sequence $x = (x_1, \ldots, x_T)$ we write $\mathbf{h}_{\ell,t}(x) \in \mathbb{R}^d$ for the normalized hidden state of token t at layer ℓ with $1 \leq \ell \leq L,\ 1 \leq t \leq T$. In the binary setting, a linear probe s is a shallow classifier that maps this representation to a supervised label $y \in \{0, 1\}$ (e.g. TRUTHFUL $= 1$, FALSE $= 0$) without updating the weights of LM.

$$s(x) = \theta^\top \mathbf{h}_{\ell,t}(x) + b \tag{1}$$

source_model	Linear	Aligned Linear	Mass-Mean	Analogous	Δ Analogous
gemma-2-2b	79.59%	**86.29%**	75.29%	84.32%	-1.96%
gemma-2-2b-it	82.91%	**83.59%**	75.11%	78.81%	-4.78%
Llama-3-8B	69.05%	67.50%	69.55%	**84.61%**	15.06%
Llama-3-8B-Instruct	77.29%	77.98%	65.61%	**89.25%**	11.27%
Mistral	99.29%	**99.37%**	96.59%	98.77%	-0.59%
mistral-instruct	98.79%	98.73%	91.44%	**98.86%**	0.06%

Table 1: Performance of probes when transferred between base and fine-tuned variants of the same model. Δ Analogous is the difference between the performance of the analogous probe and the best performance of the other probes.

Where the weight vector $\theta \in \mathbb{R}^d$ and bias b are parameters that are trained on a labeled dataset $\mathcal{T}$ of model inputs using standard cross-entropy loss. It is this weight vector θ that we decompose.

Similarly to ITDA [Leask et al., 2025b], we decompose θ into a dictionary of activations $\mathbf{D}$. In our case, we are only interested in decomposing a single vector, so rather than collect a dictionary of activations as in ITDA, we use the entire probe training dataset $\mathcal{T}$ as our $\mathbf{D}$. We then solve the following sparse coding problem to obtain the coefficients $\mathbf{a}$:

$$\min_{\mathbf{a}\in\mathbb{R}^n} ||\theta - \mathbf{aD}|| \text{ subject to } ||\mathbf{a}||_0 \leq L_0 \tag{2}$$

where $||\cdot||$ is the l_0-pseudo-norm (the number of non-zero elements), and L_0 is a pre-specified sparsity level (the number of latents used to represent each θ). More specifically, we use orthogonal matching pursuit [Mallat and Zhang, 1993] (see Algorithm 1) to find an approximation to this solution.

Our analogous probe is then defined as the approximate solution $\mathbf{a}$ to the sparse coding problem, and we can reconstruct the probe direction through matrix multiplication. We call these probes as analogous as they are constructed by reference to examples of activations, rather than optimisation.

To transfer an analogous probe to another model $\hat{LM}$, we generate a training dataset $\mathcal{T}_{\hat{LM}}$ using the same dataset as was used to generate the original training dataset. This means that activations at the same index in each dataset corresponds to the activations of the two different models on the same prompt and token. This gives a dictionary $\mathbf{D}_{\hat{LM}}$ that we use to reconstruct our probe $\mathbf{aD}_{\hat{LM}}$.

3 Results

We evaluate analogous probe in two settings. Firstly, the setting where LM = $\hat{\text{LM}}$. Here, we are interested in the performance of our probes at various levels of sparsity, and example decompositions. Secondly, on the setting where we transfer probes between different models. Here, we are interested in the performance gap between analogous probes, which are invariant to linear changes of basis in the representation space [Moschella et al., 2022], and logistic probes, which are not (Section C.4).

In both sets of experiments we use six probing datasets from Marks and Tegmark [2023], which consist of true and false statements in three different domains. We use models from three families of LLM [Team et al., 2024, Dubey et al., 2024, Jiang et al., 2023]. Across our experiments, we train the probe on the final content token position, i.e. not including special tokens, and on layer 12. See Appendix C for more experimental details.

3.1 Same-Model Reconstruction Performance

When decomposing probes into the entire activation dataset, i.e. no sparsity, the reconstructed probe achieves similar performance to the original logistic probe, and significantly better performance than the mass-mean probe (Table 2). Figure 1 shows the average performance of the three probing methods when constructed from a fixed number of samples. The analogous probes can achieve the same performance as the logistic probes with an L0 in the single digits. In Section D.2 we provide example probe decompositions, however do not investigate the interpretability of these decompositions further.

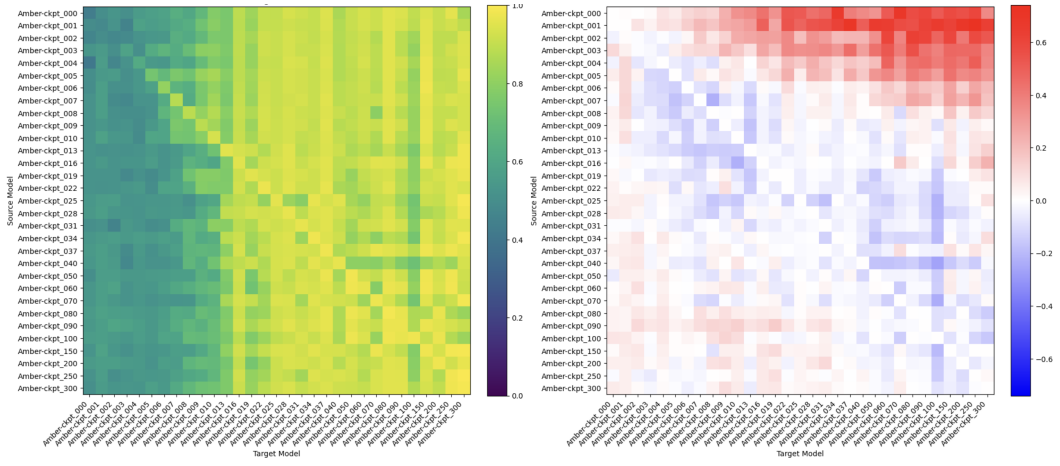

Figure 2: Performance of analogous probes when transferring between checkpoints of the LLM360 Amber model on the left, and the performance difference between the analogous probe and the naively transferred linear probe on the right. For linear probe performance see Figure 4.

Base and Fine-Tuned Models Various papers [Qian et al., 2024, Du et al., 2025, Mosbach et al., 2020] have used linear probes to evaluate the effect of training and fine-tuning on LLM representations of input features. A decrease in the performance of the linear probe is considered evidence that the model representations have somehow changed, however, this does not account for affine transformations (scaling, rotations) of the representation space, which decrease linear probe performance but may not reflect changes in model computation. These representation spaces can be aligned through linear regression [Mikolov et al., 2013b], however this is a global alignment and may not be sensitive to transformations of representation subspaces, and local alignment may not be possible due to the high dimensionality of the representations. In Table 1, we compare these methods to analogous probes, which are invariant to transformations of subspaces. We find that for Mistral 8b, the logistic probe transfers well anyway; and that for Gemma 2, none of the probes transfer well, suggesting changes to the structure of the representation space. However, on the Llama models we note that analogous probes perform significantly better than the other methods - this suggests that some of the drop in probe performance is due to a subspace transformation, rather than a structural change.

Pretraining Checkpoints We evaluate the generalization performance of probes trained on checkpoints of LLM 360 Amber [Liu et al., 2023]. Due to computational constraints we only evaluated analogous probes and naively transferred linear probes, i.e. without learning a linear regressor between the activations of the models, however will address this in future work. We find that there is a significant performance difference between the two probing methods when transferring from an early checkpoint of Amber to a late checkpoint and vice versa, with the linear model maintaining its performance in a broad region around the leading diagonal of the heatmap. This again suggests that truth may be established earlier during training that previously thought, but undergoes transformation throughout training.

4 Discussion

Analogous probes offer a novel perspective on probing: using a weighted sum of just a handful of activations, they can match the performance of logistic probes and far exceed that of mass-mean probes; they are also defined by relation to activations, rather than absolute vectors, and so are less impacted by transformations of representation spaces. In future work, we plan to develop a better understanding of probe interpretability, and how analogous probing can contribute to that agenda. In particular, we will use them in downstream interpretability tasks to validate their usefulness. We also want to understand better why analogous probes transfer so much better on Llama than other probes, and analogous probes on other LLMs. See Section E for an detailed list of limitations.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *International Conference on Learning Representations*, 2017.

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

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. *Advances in neural information processing systems*, 34:225–236, 2021.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Joseph Berkson. Application of the logistic function to bio-assay. *Journal of the American statistical association*, 39(227):357–365, 1944.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. *Advances in Neural Information Processing Systems*, 37:107286–107325, 2024.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *NeurIPS Workshop on Scientific Methods for Understanding Deep Learning (SciForDL)*, 2024.

Kwanghee Choi, Jee-weon Jung, and Shinji Watanabe. Understanding probe behaviors through variational bounds of mutual information. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5655–5659. IEEE, 2024.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. *arXiv preprint arXiv:2506.03093*, 2025.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *ICLR*, 2024.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

Hongzhe Du, Weikai Li, Min Cai, Karim Saraipour, Zimin Zhang, Himabindu Lakkaraju, Yizhou Sun, and Shichang Zhang. How post-training reshapes llms: A mechanistic view on knowledge, truthfulness, refusal, and confidence. *arXiv preprint arXiv:2504.02904*, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *NeurIPS 2024*, 2024.

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings, 2010.

Thomas Fel, Ekdeep Singh Lubana, Jacob S Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. *arXiv preprint arXiv:2502.12892*, 2025.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *ICLR*, 2025.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

Sai Sumedh R Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba Ba. Projecting assumptions: The duality between sparse autoencoders and concept geometry. *arXiv preprint arXiv:2503.01822*, 2025.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. 2008.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. doi: 10.48550/arXiv.2310.06825.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Arthur Conmy, Callum McDougall, Kola Ayonrinde, Matthew Wearden, Samuel Marks, and Neel Nanda. Saebench: A comprehensive benchmark for sparse autoencoders, December 2024. URL `https://www.neuronpedia.org/sae-bench/info`. Accessed: 2025-01-20.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. Interpreting attention layer outputs with sparse autoencoders. *ICML*, 2024a.

Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. Saes (usually) transfer between base and chat models. In *AI Alignment Forum*, 2024b.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

Jenny Kunz and Marco Kuhlmann. Classifier probes may just learn from linear context features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5136–5146, 2020.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders. *arXiv preprint arXiv:2410.06981*, 2024.

Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. *ICLR*, 2025a.

Patrick Leask, Neel Nanda, and Noura Al Moubayed. Inference-time decomposition of activations (itda): A scalable approach to interpreting large language models. *International Conference on Machine Learning*, 2025b.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *ICLR*, 2016.

Johnny Lin. Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. URL `https://www.neuronpedia.org`. Software available from neuronpedia.org.

Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, 2024.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, et al. Simple probes can catch sleeper agents. *Anthropic Research Updates*, 2024.

Aleksandar Makelov, George Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *ICLR*, 2025.

Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *ICLR*, 2025.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013b.

Marius Mosbach, Anna Khokhlova, Michael A Hedderich, and Dietrich Klakow. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. *arXiv preprint arXiv:2010.02616*, 2020.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *arXiv preprint arXiv:2209.15430*, 2022.

Jord Nguyen, Khiem Hoang, Carlo Leonardo Attubato, and Felix Hofstätter. Probing evaluation awareness of language models. *arXiv preprint arXiv:2507.01786*, 2025.

Chris Olah. Visualizing representations: Deep learning and human beings. 2015.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. *arXiv preprint arXiv:2410.02707*, 2024.

Gonçalo Paulo, Alex Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *arXiv preprint arXiv:2410.13928*, 2024.

Chen Qian, Jie Zhang, Wei Yao, Dongrui Liu, Zhenfei Yin, Yu Qiao, Yong Liu, and Jing Shao. Towards tracing trustworthiness dynamics: Revisiting pre-training period of large language models. *arXiv preprint arXiv:2402.19465*, 2024.

Alec Radford, R Jozefowicz, and I Sutskever. Learning to generate reviews and discovering sentiment. arxiv 2017. *arXiv preprint arXiv:1704.01444*, 2017.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders. *ICLR Workshop on Building Trust in Language Models and Applications*, 2025.

Fabien Roger, Ryan Greenblatt, Max Nadeau, Buck Shlegeris, and Nate Thomas. Benchmarks for detecting measurement tampering. *arXiv preprint arXiv:2308.15605*, 2023.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Adly Templeton. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic, 2024.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second International Conference on Machine Learning*, 2025.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Language models linearly represent sentiment. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

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

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. *ICLR*, 2020.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Related Work

Activation Probing: Alain and Bengio [2017] found that high-level concepts can be decoded from the middle layers of models using linear probes. This has been further validated on specific concepts in LLMs such as refusal [Arditi et al., 2024], sentiment [Tigges et al., 2024], spatial and temporal relationships [Gurnee and Tegmark, 2023], and truthfulness [Marks and Tegmark, 2023]. Furthermore, probes have been used to demonstrate that models know when they are being tested [Nguyen et al., 2025], and predict when models will produce jailbroken responses [Zou et al., 2023], harmful information to malicious users [Roger et al., 2023], or falsehoods [Orgad et al., 2024].

Limitations of Probing: Probes can leverage spurious correlations in their training dataset, rather than true features of the model; and even linear probes can overfit when the representation dimension is high enough Belinkov [2022]. There is also little empirical research into the features that linear probes capture [Kunz and Kuhlmann, 2020, Belinkov, 2022, Choi et al., 2024, Sharkey et al., 2025]. Linear probes also transfer imperfectly between base and fine-tuned models [Du et al., 2025, Mosbach et al., 2020], and have variable performance across training checkpoints [Qian et al., 2024].

SAEs for Mechanistic Interpretability: Sparse Autoencoders (SAEs) have been used to recover sparse, monosemantic, and interpretable features from the representations of LLMs [Bricken et al., 2023, Cunningham et al., 2024, Templeton, 2024]. The decompositions found by these SAEs are often manually inspected using feature dashboards [Bricken et al., 2023, Lin, 2023], automatically described through automated interpretability techniques [Gao et al., 2025, Paulo et al., 2024], and evaluated with the SAEBench benchmarking suite [Karvonen et al., 2024]. SAEs have been used for circuit analysis [Marks et al., 2025] in the vein of [Olah et al., 2020, Olsson et al., 2022]; to study the role of attention heads in GPT-2 [Kissane et al., 2024a]; and to replicate the identification of a circuit for indirect object identification in GPT-2 [Makelov et al., 2025]. Transcoders, a variant of SAEs, have been used to simplify circuit analysis and applied to the greater-than circuit in GPT-2 [Dunefsky et al., 2024]. Whilst the application of SAEs to mechanistic interpretability is supported by qualitative and quantitative evidence, their usefulness is highly dependent on hyperparameterisation [Leask et al., 2025a].

A number of SAE variants have been proposed that modify either the activation function or the loss term [Gao et al., 2025, Rajamanoharan et al., 2025, Bussmann et al., 2024, Rajamanoharan et al., 2024, Braun et al., 2024, Thasarathan et al., 2025, Fel et al., 2025, Hindupur et al., 2025]. Leask et al. [2025b] introduced Inference-Time Decomposition of Activations (ITDA), which decomposes activations into an iteratively constructed dictionary of activations at inference time using matching pursuit [Mallat and Zhang, 1993]. [Costa et al., 2025] similarly applied matching pursuit as an encoder, but with an optimized dictionary.

In their investigation into the canonicality of SAE latents, Leask et al. [2025a] introduced MetaSAEs, a second-order application of SAEs, that are trained on the problem of reconstructing SAE decoder vectors using a sparse and overcomplete bottleneck. MetaSAEs decompose these latents into a sparse sum of second-order latents; for example, the decoder vector of a first-order latent activating on the token "Einstein" in a GPT-2 SAE was decomposed in second-order latents relating to "Germany", "Words starting with E-", "Prominent Figures", "Space and Galaxies", and "Science and Scientists". Whilst the first-order latents seem to correspond to atomic and interpretable concepts, so do the second-order latents.

Representation Engineering and Steering: Since the classic word2vec result that “king - man + woman = queen" [Mikolov et al., 2013a], there has been significant interest in performing arithmetic on model activations. Guiding the generations of models by modifying neuron activations at inference time has been used to mitigate gender bias in models Bolukbasi et al. [2016], and steer review sentiment [Radford et al., 2017]. Steering vectors that are added to model activations at inference time to modify behavior have been found through training classifiers [Dathathri et al., 2019] and simply taking the difference in activations between contrastive pairs of prompts Turner et al. [2023]. Contrastive pairs of prompts have also been used to construct simple prompts to detect sleep agents [MacDiarmid et al., 2024]. These methods form the foundation of the emerging field of representation engineering as described by [Zou et al., 2023]. Contrast-consistent search used pairs of positive and negative activations to construct probes without optimizing for classification performance on a target

variable [Burns et al., 2022], and mass-mean probing uses the difference in means between positive and negative examples to find a probe direction [Marks and Tegmark, 2023].

Representation Similarity: A range of methods for comparing representations between neural networks has been developed. Inspired by Erhan et al. [2010], Olah [2015] applied t-SNE, a dimensionality reduction technique [Van der Maaten and Hinton, 2008], to the representations of vision and language models. Lenc and Vedaldi [2015], Bansal et al. [2021] stitched layers of two frozen models with a trained intermediate adapter layer, and evaluated the similarity of the model's representations by the performance of the stitched model. Representation similarity metrics compare the alignment of the representation subspaces of different models, and include Singular Vector Canonical Correlation Analysis (SVCCA) [Raghu et al., 2017] and Centered Kernel Alignment (CKA) [Kornblith et al., 2019]. The performance of linear probes trained on the representations of different models can provide insight into what information the representations represent [Alain and Bengio, 2017, Hewitt and Manning, 2019]. Li et al. [2016] investigated whether different neural networks converge to the same representations, and [Garipov et al., 2018, Zhao et al., 2020] find that different models are occupied by low-loss paths in the parameter space. Olah et al. [2020] provide examples of potential universal features, such as curve detectors, in vision models, and Olsson et al. [2022] find evidence of induction heads in language models of different sizes. Bricken et al. [2023] found similar SAE latents in different models, and Kissane et al. [2024b] found examples of SAEs that transfer between base and fine-tuned versions of the same language model. Lindsey et al. [2024] used Crosscoders, SAEs trained on the representations of multiple models, to find features present in a fine-tuned version of an LLM that were not present in the base model. Relative representation methods are kernel methods Hofmann et al. [2008] that measure similarity against a set of prototype inputs [Moschella et al., 2022], which avoids learning model specific parameters from absolute model representations. Lan et al. [2024] compared the feature spaces of different models using SAEs, and Leask et al. [2025b] did the same with ITDA.

B Orthogonal Matching Pursuit

We decompose probe directions using orthogonal matching pursuit [Mallat and Zhang, 1993] as described in Algorithm 1. Our dictionary A is the model activations on the training dataset, and our y are probe directions.

Algorithm 1 Orthogonal Matching Pursuit (OMP)

Require: $A \in \mathbb{R}^{m \times n}$ with columns a_j $\big(\|a_j\|_2 = 1\big)$, $y \in \mathbb{R}^m$, sparsity $k \in \mathbb{N}$; optional tolerance $\varepsilon \geq 0$ (default 0)
Ensure: Support $S \subseteq \{1, \ldots, n\}$, estimate $\hat{x} \in \mathbb{R}^n$
$S \leftarrow \emptyset$, $r \leftarrow y$, $t \leftarrow 0$
while $t < k$ **and** $\|r\|_2 > \varepsilon$ **do**
  $t \leftarrow t + 1$
  $j^\star \leftarrow \arg\max_{j \notin S} \; |a_j^\top r|$ ▷ Maximal correlation
  $S \leftarrow S \cup \{j^\star\}$
  $x_S \leftarrow \arg\min_{z \in \mathbb{R}^{|S|}} \|y - A_S z\|_2$ ▷ e.g., $x_S = A_S^\dagger y$
  $r \leftarrow y - A_S x_S$
end while
$\hat{x} \leftarrow 0 \in \mathbb{R}^n$; $\hat{x}_S \leftarrow x_S$
return $S, \hat{x}$

C Experiment Details

This section contains further details of the datasets, models, and probe transfer methods that we use in Section 3.

C.1 Datasets

We use six datasets of the twelve true/false datasets used in Marks and Tegmark [2023]. Details of these datasets are given in Table C.1. Unlike Marks and Tegmark [2023], we merge the positive and negative datasets for each task into a single dataset. The negative datasets consists of the negation of the positive dataset, eg. "The city of [city] is in [country]"becomes "The city of [city] is not in [country]". These datasets are small, and limited in that they probe for a single target variable, truth; we will improve the diversity of our datasets in future work.

Name	Description	Rows
`cities`	"The city of [city] is in [country]."	2992
`larger_than`	"x is larger than y."	3960
`sp_en_trans`	"The Spanish word '[word]' means '[English word]'."	708

C.2 Models

We use LLMs from the Gemma [Team et al., 2024], Llama [Dubey et al., 2024], and Mistral [Jiang et al., 2023] families. We also use a range of checkpoints from LLM360 Amber [Liu et al., 2023]. Details of the models are included in Table C.2.

Model ID	Family	Params	Hugging Face
Gemma 2 2B	Gemma [Team et al., 2024]	2B	HF
Gemma 2 2B IT	Gemma [Team et al., 2024]	2B	HF
Llama 3 8B	Llama [Dubey et al., 2024]	8B	HF
Llama 3 8B Instruct	Llama [Dubey et al., 2024]	8B	HF
Mistral 7B v0.3	Mistral [Jiang et al., 2023]	7B	HF
Mistral 7B v0.3 Instruct	Mistral [Jiang et al., 2023]	7B	HF
Amber 7B	LLM360 [Liu et al., 2023]	7B	HF

C.3 Probing Baselines

We compare analogous probes to logistic probes [Berkson, 1944] and mass-mean probes [Marks and Tegmark, 2023], which are described below, with details on how they are transferred between models.

Mass-mean probing Given labeled data $\mathcal{D} = \{(x_i, y_i)\}$ with $y_i \in \{0, 1\}$, compute the class means $\mu^+ = \frac{1}{|\{i:y_i=1\}|} \sum_{i:y_i=1} x_i$ and $\mu^- = \frac{1}{|\{i:y_i=0\}|} \sum_{i:y_i=0} x_i$. Define the direction $\theta_{\mathrm{mm}} = \mu^+ - \mu^-$ and the probe

$$p_{\mathrm{mm}}(x) = \sigma(\theta_{\mathrm{mm}}^\top x).$$

For IID evaluation use a covariance–whitened variant

$$p_{\mathrm{mm}}^{\mathrm{iid}}(x) = \sigma(\theta_{\mathrm{mm}}^\top \Sigma^{-1} x),$$

where Σ is the covariance of the class–centered dataset $\mathcal{D}^c = \{x_i - \mu^+ : y_i = 1\} \cup \{x_i - \mu^- : y_i = 0\}$. We refer to p_{mm} and $p_{\mathrm{mm}}^{\mathrm{iid}}$ as *mass-mean probes*. In our experiments we evaluate only mass-mean probes, rather than whitened mass-mean probes, which we will also evaluate in future work.

Unwhitened mass-mean probes can be understood as analogous probes where, for a training dataset consisting of P positive examples and N negative examples, the weight given to each positive activation is $1/P$ and each negative activation is $-1/N$. When we transfer unwhitened mass-mean probes between models, we use these weights in the same way we would for analogous probes; however, because the weights are independent of the values of the activations, this is the same as retraining the mass-mean probe on the target model.

When comparing mass-mean probes to analogous probes at a fixed L0, as we do in Section 3.1, we construct them from a training subset of size L0. This is because selecting the optimal subset of training data points for mass-mean probing is of combinatorial complexity. However, we select the best performing of the probes trained on 30 subsets of the training data for comparison. This value of 30 subsets was chosen to balance performance with the computational cost of training on many random subsets, a case study of the number of random subsets required to achieve analogous probe performance is presented in Figure 3.

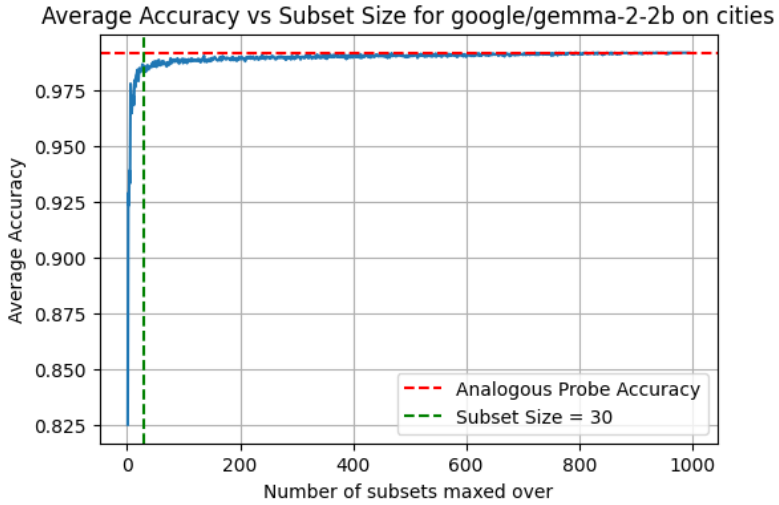

Figure 3: Maximum performance of linear probes on a specific model and task when trained on x subsets. I.e. for x=30, linear probes were trained on 30 different subsets of the dataset, and the plotted value is the performance of the best probe. To achieve within 0.1% of the performance of the analogous probe, the maximum must be taken over 336 random probes, which is too computationally expensive to run for all the experiments in this paper, and unlikely to be useful in practice.

Logistic probing Given labeled data $\mathcal{D} = \{(x_i, y_i)\}$ with $y_i \in \{0, 1\}$, fit a linear logistic-regression probe on the representations x_i. The probe is

$$p_{\text{lr}}(x) = \sigma(\theta^\top x + b), \qquad \sigma(t) = (1 + e^{-t})^{-1}.$$

Estimate (θ, b) by minimizing the negative log-likelihood

$$\min_{\theta,b} \; -\sum_i \Big[y_i \log p_{\text{lr}}(x_i) + (1 - y_i) \log\big(1 - p_{\text{lr}}(x_i)\big) \Big] + \lambda \|\theta\|_2^2,$$

with optional ℓ_2 regularization $\lambda \geq 0$. For hard predictions use $\mathbb{I}[p_{\text{lr}}(x) \geq 1/2]$. We refer to this as a *logistic probe*.

When comparing linear probes to analogous probes at a fixed L0, we also select the best performing probe from 30 random subsets of the training data again because choosing the best training dataset of a fixed size is of combinatorial complexity. We also train linear probes on the training samples that are used in the analogous probe decompositions.

When we transfer probes from a source to target model, we always use the full training dataset to do so. Analogous probes we transfer as described in Section 2. Mass-Mean probes can be thought of analogous probes where each positive and negative training sample is given the same weight when constructing the probe coefficient, so we construct these from the activations of the target model directly. We transfer linear probes naively by not updating their coefficient vector, but also by learning a logistic regressor between the activation spaces of the source and target models that we trained on a generic pretraining dataset [Gao et al., 2020]. In all of these cases, we estimate a new bias for the probe on the target model training data as below, as some methods do not estimate this parameter directly.

Given a fixed weight vector $w \in \mathbb{R}^d$ and data $X \in \mathbb{R}^{n \times d}$ with labels $y \in \{0,1\}^n$, compute the projections $p_i = x_i^\top w$ for $i = 1, \ldots, n$. Fit only a bias $b \in \mathbb{R}$ by maximizing empirical accuracy of the threshold classifier

$$\hat{y}_i(b) = \mathbb{I}[\, p_i > b \,].$$

The estimate is

$$b^\star \in \arg\max_{b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\big(\hat{y}_i(b) = y_i\big).$$

Implementation detail: initialize b at $\mathrm{median}(p_1, \ldots, p_n)$ and minimize the negative accuracy with Powell's method. The routine returns the scalar $b^\star$.

C.4 Invariance of Analogous Probes to Affine Transforms

Let $D = [h_1, \ldots, h_n] \in \mathbb{R}^{d \times n}$ collect activations $h_i \in \mathbb{R}^d$ and let $\theta \in \mathbb{R}^d$ be the probe. Consider the sparse reconstruction

$$a^\star \in \arg\min_{a \in \mathbb{R}^n} \ \|\theta - Da\|_2 \quad \text{s.t.} \quad \|a\|_0 \le L_0,\ \mathbf{1}^\top a = 1,$$

and set $\hat{\theta} := Da^\star$. For any invertible affine map $T(x) = Ax + c$ with $A \in GL_d(\mathbb{R})$ and $c \in \mathbb{R}^d$, define

$$D' := AD + c\,\mathbf{1}^\top, \qquad \theta' := A\theta + c.$$

Then the reconstruction co-transforms:

$$\hat{\theta}' \ := \ D'a^\star \ = \ T(\hat{\theta}) \ = \ A\hat{\theta} + c.$$

By direct expansion and the constraint $\mathbf{1}^\top a^\star = 1$,

$$\hat{\theta}' \ = \ (AD + c\,\mathbf{1}^\top)a^\star \ = \ ADa^\star + c(\mathbf{1}^\top a^\star) \ = \ A\hat{\theta} + c.$$

If $c = 0$ (pure change of basis), the same statement holds without the constraint $\mathbf{1}^\top a = 1$: for $D' = AD$ and $\theta' = A\theta$, one has $D'a^\star = A(Da^\star)$.

In our setting, we do not have $c = 0$, however we center and normalize the activations to reduce the effect of translations. We confirmed this empirically by training linear and analogous probes on the activations of Gemma 2 2B in the cities probing task, both achieving 98.6% accuracy. We then applied 100 random affine transformations to the activations, keeping the target variable the same. The average performance of the linear regressor dropped to 57.7%, with the analogous probe only falling to 98.5% accuracy.

D Further Results

This section includes further results from the experiments in Section 3.

D.1 Reconstructed Probe Performance

Table 2 shows the performance of logistic, mass-mean, and analogous probes trained on the full training dataset and evaluate on the source model.

dataset	probe_type model	Logistic	Mass-Mean	Analogous
cities	google/gemma-2-2b	98.83%	97.00%	**99.17%**
	google/gemma-2-2b-it	98.83%	74.79%	**99.17%**
	meta-llama/Meta-Llama-3-8B	99.83%	99.33%	**99.83%**
	meta-llama/Meta-Llama-3-8B-Instruct	99.33%	98.50%	**99.33%**
	mistralai/Mistral	99.17%	98.16%	**98.66%**
	mistralai/mistral	99.67%	97.66%	**99.67%**
larger_than	google/gemma-2-2b	99.75%	88.89%	**99.62%**
	google/gemma-2-2b-it	100.00%	97.73%	**100.00%**
	meta-llama/Meta-Llama-3-8B	99.12%	87.50%	**99.12%**
	meta-llama/Meta-Llama-3-8B-Instruct	96.59%	80.43%	**96.72%**
	mistralai/Mistral	99.87%	98.36%	**99.87%**
	mistralai/mistral	99.75%	97.47%	**99.75%**
sp_en_trans	google/gemma-2-2b	95.77%	69.01%	**96.48%**
	google/gemma-2-2b-it	97.18%	78.87%	**97.18%**
	meta-llama/Meta-Llama-3-8B	97.89%	88.03%	**98.59%**
	meta-llama/Meta-Llama-3-8B-Instruct	97.89%	82.39%	**97.89%**
	mistralai/Mistral	97.89%	78.17%	**97.89%**
	mistralai/mistral	98.59%	92.96%	**98.59%**

Table 2: Comparison of analogous probes to the original logistic probe and mass-mean probe when using all training samples in the reconstruction, excluding a held-out validation set of samples from the dataset. The row-wise maximum of the analogous probe and mass-mean probe is bold.

D.2 Probe Decompositions

Tables 3, 4, and 5 show the decomposition by orthogonal matching pursuit of the linear probes on three models for the cities task into training dataset examples. The probes are trained on the final position residual stream value after layer 12, and decomposed into those values again, with the Statement referring to the prompt that caused that hidden state. For a more in-depth explanation of how this works, and the justification for these dashboards as an interpretability tool, see Leask et al. [2025b].

Statement	Label	Weight
The city of Anqing is in Brazil.	0	9.623209
The city of Vilnius is in Lithuania.	1	4.178972
The city of Nagpur is in India.	1	4.011485
The city of Sambhaji Nagar is in India.	1	3.853475
The city of Sanmenxia is in China.	1	3.150570
The city of Managua is in Nicaragua.	1	2.030811
The city of Cankaya is in Turkey.	1	-1.492452
The city of Pyongyang is not in North Korea.	0	-2.965898
The city of Kota Kinabalu is in Malaysia.	1	-5.339108
The city of Belo Horizonte is not in Brazil.	0	-5.700376

Table 3: Decomposition of the Gemma 2 2B cities probe with L0=10.

Statement	Label	Weight
The city of Sapporo is not in Mexico.	1	12.681614
The city of Sharjah is not in China.	1	9.578495
The city of Saratov is in Indonesia.	0	7.903430
The city of Kagoshima is in Japan.	1	4.766314
The city of Linyi is not in Iraq.	1	3.978264
The city of Luohe is not in China.	0	-5.803697
The city of Perm is not in China.	1	-5.939833
The city of Conakry is not in Belarus.	1	-6.163413
The city of Bien Hoa is in Vietnam.	1	-6.799433
The city of Ashgabat is in Bangladesh.	0	-9.745727

Table 4: Decomposition of the Gemma 2 2B Instruct cities probe with L0=10.

Statement	Label	Weight
The city of Mbuji-Mayi is in Pakistan.	0	10.500116
The city of Jalandhar is not in India.	0	3.402439
The city of Bursa is in Turkey.	1	3.336683
The city of Malatya is not in Russia.	1	3.074517
The city of Hamadan is in Iran.	1	2.856083
The city of Macapa is not in Brazil.	0	-3.195982
The city of Chandigarh is in India.	1	-3.892706
The city of Fort Worth is not in Russia.	1	-3.902153
The city of Taguig is in the Philippines.	1	-4.046275
The city of Kampung Baru Subang is in the Philippines.	0	-5.100050

Table 5: Decomposition of the Llama 3 8B Instruct cities probe with L0=10.

D.3 Pretraining Generalization

Figure 4 shows a heatmap of the generalization performance when transferring linear probes between different checkpoints of LLM360 Amber [Liu et al., 2023], averaged across datasets. Note the worse performance on transfers far off the leading diagonal.

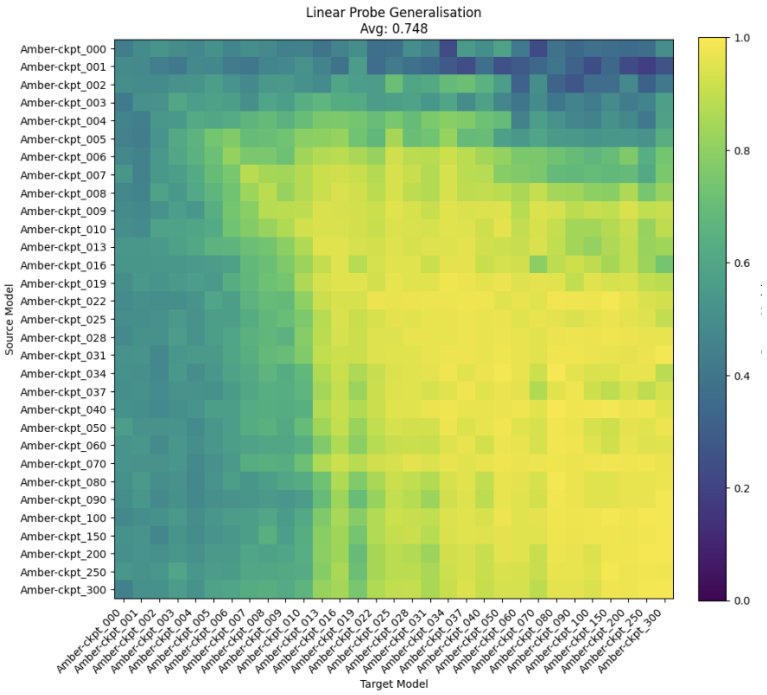

Figure 4: Performance of linear probes when transferred between checkpoints of the LLM360 Amber model. See Figure 2 for analogous probing results and a comparison.

Figure 5 shows the generalization performance of linear and analogous probes as a function of the number of steps between the source and target training checkpoint.

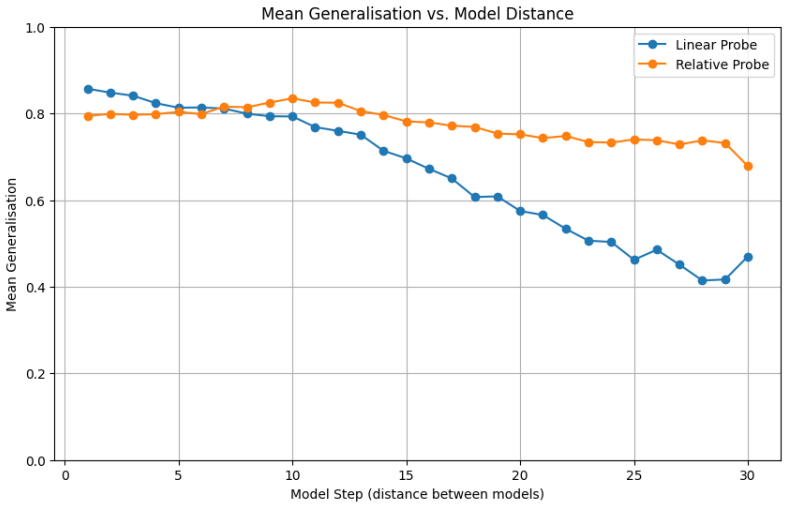

Figure 5: Performance of linear and analogous probes when transferring between checkpoints of certain separations, i.e. when transferring from the model checkpoint at step i to that at step j, then the x-axis is equal to $j - i$, and the y-axis is the average performance of the accuracies at that x value.

E Limitations

1. Whilst we hypothesize that analogous probes transfer better because of their invariance to transformation [Moschella et al., 2022], we do not investigate these results deeply. For example, we wonder why analogous probes outperform other methods by so much on the Llama models.
2. We only use six of the datasets, merging both positive and negative variants, from Marks and Tegmark [2023]. In future work we intend to include the rest of the datasets, and further datasets from other papers.
3. We do not evaluate the interpretability of our probe decompositions on downstream tasks. We intend to address this in future work using selectivity experiments: i.e., given the decompositions of a real probe and one trained to predict an unrelated target, can a human differentiate the probes.
4. There are other probing methods against we could compare, such as whitened mass-mean and contrast-consistent search [Burns et al., 2022]; and we also need to include results for linearly-aligned baseline to the Amber results. We note the general lack of literature on this problem from which to construct strong baselines, however.
5. Whilst we evaluated the method on diverse model families, all the chosen models are small and we only targeted a single layer in each of the models. Further ablations would increase our confidence in our results further, especially finding more cases where analogous probe performance is dramatically different to other probes.