HIGH FREQUENCY LATENTS ARE FEATURES, NOT BUGS

Xiaoqing Sun
MIT
xqsun@mit.edu

Joshua Engels
MIT
jengels@mit.edu

Max Tegmark
MIT
tegmark@mit.edu

ABSTRACT

Sparse autoencoders (SAEs) have shown success at decomposing language model activations into a sparse set of interpretable linear representations ("latents"). However, recent work identifies a challenge for SAEs: high frequency latents (HFLs) that are seemingly uninterpretable and occur on greater than 10% of tokens. In this work, we find that HFLs have many unique properties: 1) most HFLs have a "pair", another HFL pointing in the geometrically opposite direction that they never co-occur with; 2) the HFL subspace is robust to the SAE initialization seed, but HFLs themselves are not; 3) when an SAE is trained on activations with the HFL subspace ablated, no new HFLs are learned; and 4) HFLs have uniquely high similarity with the SAE bias vector. Our experiments lead us to hypothesize that the HFL subspace is not an artifact of SAE training, but instead represents a subspace of truly dense language model features. We present preliminary results interpreting this dense subspace, including finding HFLs that represent context position, HFLs that fire continuously on large blocks of text, HFLs that fire on topic sentences, and HFLs that fire on numeric data.

1 INTRODUCTION

Sparse autoencoders (SAEs) (Bricken et al., 2023; Cunningham et al., 2023a;b) are an unsupervised learning technique that identifies a "dictionary" of latents. Model activations on any one token are a sparse combination of these latents. Most SAE latents are monosemantic, meaning they correspond to human-interpretable concepts. For instance, Templeton et al. (2024) find a "Golden Gate Bridge" latent in Claude 3 Sonnet Anthropic (2024), which they amplify to steer model behaviour.

However, as noted by Anthropic (2024), SAEs often still learn a few high-frequency latents (HFLs) that occur on a significant proportion of input tokens ($f \gtrapprox 10\%$). While it seems possible to reduce the number of such learned latents by enforcing greater sparsity (Anthropic, 2024), this does not remove them entirely; we show in Figure 16 that even low sparsity SAEs have a handful of HFLs.

Even if we could make sparsity low enough to remove HFLs, this may be a mistake: there are dense features that language models may need to track for all input tokens, such as context position. **We hypothesize that HFLs span a subspace of these truly dense language model features.** However, SAEs rely on features being sparse. Thus, in this dense regime, SAEs may not learn HFLs that align exactly with the true features (see Section 2), explaining why prior work has found that HFLs are mostly uninterpretable. We explore this hypothesis in two high level steps:

1. In Section 2, we explore the *geometry* and *characteristics* of HFLs and the HFL subspace. We find evidence that the HFL subspace contains true features: HFLs come in geometrically opposite pairs, the same HFL subspace is learned between SAE seeds (but the same HFLs are not), and ablating the HFL subspace causes HFLs not to be learned.
2. In Section 3, we investigate *specific* HFLs, finding that HFLs are actually *somewhat* interpretable: we find HFLs that depend on context position, distance to digit tokens, distance to newlines, and whether the given HFL fired previously.

Additionally, we discuss related work in Appendix A, show example texts that HFL pairs fire on in Appendix D, and show that our results generalize in Appendix E. Overall, our results are promising evidence that the HFL subspace can be understood and contains dense language model features.

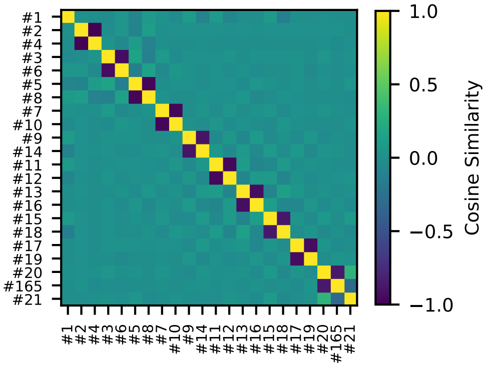


Figure 1: HFL cosine similarity matrix. For the most part, HFLs are perpendicular to each other, except that for all but the first HFL there is another HFL with -1 cosine similarity.

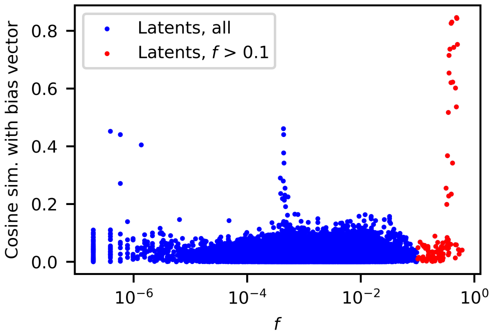


Figure 2: SAE latent frequency vs. cosine similarity with the SAE bias vector across all seeds. HFLs make up most of the latents with a high cosine similarity with the bias.[1]

2 Characteristics of High-Frequency Latents

Method: In this section, we examine properties of HFLs and their subspace. We use seeds 42, 43, 44, and 45 to train TopK (Gao et al., 2024) SAEs on layer 12 of Gemma-2-2B (Team et al., 2024; Team, 2024). We obtain frequency data from 5M tokens of RedPajama (Weber et al., 2024), discovering ~ 20 latents with frequency $f > 10\%$. See Appendix B for our full methodology. We note that our findings generalize well; we include reproductions on other SAEs in Appendix E. We denote by #i the latent with the ith largest frequency.

High-frequency latents come in geometrically opposite pairs: We find the cosine similarity between all latent directions within a single SAE. While most SAE latents are almost orthogonal to each other, we find that some latents that have a "pair" latent with cosine similarity ≈ -1. Interestingly, this group of latents is almost entirely HFLs, and almost all HFLs occur in a pair (see Figure 6). We plot the HFL pairs in Figure 1 for the seed=42 SAE, using a cutoff cosine similarity of < -0.85. These latents are not only geometrically opposing, but also functionally opposing — HFL pairs do not co-occur on the same token, despite individually having high frequencies (Table 2). Thus, each pair represents just one true model feature direction, but instead of the SAE learning one latent for this direction, the zero point (bias) is set at the center of this spread, and the SAE learns two latents in opposite directions.[2] For most of our experiments, we consider *pairs* of HFLs as the unit of study.

HFLs tend to be more bias-aligned: Most SAE latent directions are orthogonal to the decoder bias direction (which also tends to approximately equal the average input activation). However, a subset of HFLs have large absolute cosine similarity with the decoder bias, and they furthermore tend to be the only latents with such high similarity to the bias. We plot the absolute cosine similarity of each latent with the bias vector versus latent frequency for the seed 42 SAE in Figure 2.

The HFL subspace is rediscovered between SAEs: We train SAEs with 4 different seeds and find their HFL pairs. We then filter the HFLs to only pairs with a combined frequency of greater than 50% and take 1 member of each of these filtered pairs. We find 6-7 such HFLs across all seeds. In Figure 3 we plot the cosine similarity between the HFLs from each seed versus the HFLs from seed 42. We see that, while the HFLs are not always exactly rediscovered, the principal angles between the subspaces spanned by the HFLs from different seeds are extremely low. In other words, the HFL subspace seems to be consistently rediscovered, but not always along the same basis.

[1] The spike at moderate frequencies may be due to the presence of latents related to the <bos> token, since our SAEs were trained on all model activations including on <bos>.

[2] We are not sure why the SAE does not just learn the bias such that the "center" is at the corner of the HFL subspace; we suspect that the HFL "center" may be different from the "center" of most other sparse features. The finding that HFLs are similar to the bias supports this hypothesis.

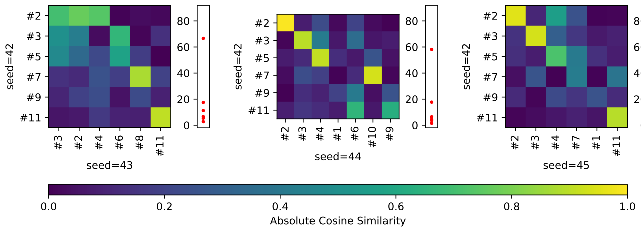


Figure 3: HFL to HFL cosine similarities (grids), and HFL subspace to HFL subspace principal angles (red dots) between seed=42 and seed=43, 44, and 45 SAEs.

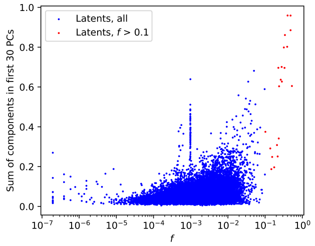


Figure 4: For each latent v_j, we find $\sum_{i=1}^{30}(v_j \cdot u_i)^2$ where u_i is the i-th PC component.[3]

Addressing Alternative Hypotheses: We address two alternative explanations for the HFL subspace. **Concern 1:** The HFL subspace is an artifact of SAE training. **Concern 2:** The HFL subspace is simply the SAE learning the highest variance dimensions.

To address concern 1, we train an SAE with the same seed $= 42$, but ablate the subspace spanned by all originally found HFLs with frequency $> 10\%$. Our results (in Figure 8) show that the new SAE discovers only one latent with frequency > 0.1 (at frequency $\approx 11\%$) and no other HFLs. Thus, HFLs truly arise from a specific language model subspace.

To address concern 2, if the HFLs are merely directions of high variance in the data, then we expect the HFL subspace (effectively ≈ 10 dimensions) and the top PCA subspace to be extremely similar. However, finding the components of HFLs along the top 30 PCA components (Figure 4), we see a significant proportion of the variance still unexplained. Furthermore, the HFL directions themselves do not simply correspond to the top PCA components (Figure 7). Thus, the SAE seems to be working well to break down the top PCA dimensions.

Linear combination hypothesis: The fact that the same subspace is repeatedly learned but the exact same HFLs are not implies an interesting hypothesis: the HFL subspace may be made up of "true" interpretable latents that are dense; however, since such latents occur on almost every token, the sparsity penalty may not be enough to learn the true basis, and the SAE might instead learn a "skewed" basis of the HFL subspace. In the next section, we still try to interpret individual HFLs, as they may be close to the true dense model features and thus still be somewhat interpretable; this partial interpretability is evidence that the true latents may be in the HFL subspace somewhere.

3 Interpretability of High-Frequency Subspace

In this section, we investigate top HFLs from seed=42 by combined frequency. We include a summary of our findings in Table 1, and example firing patterns for each HFL are in Appendix D.

Context position latent (#1): Examining the latent activations at each context position, we find that the top HFL across all seeds ($f \approx 0.5 - 0.6$) is exactly rediscovered (cosine similarity > 0.97 across seeds) and has a strong context position dependence, firing mostly and strongly early in contexts. This latent has a low-frequency (≈ 0.05) pair in two seeds, and no pair in the other two. In the top left of Figure 5, we plot latent firing count against context position.

Phrase-level Semantics (#2 and #4): The next two latents, #2 and #4, seem to have a meaningful firing pattern; latent #2 fires mostly on numeric data and short, common tokens, while latent #4 fires on most other tokens (total $f = 0.888$). We hypothesize that this HFL pair is related to a semantic axis similar to "structured text" vs. "prose / meaningful tokens". In the bottom left of Figure 5, we show that this latent pair fires at high frequency when it is near a digit token.

Context Indicator Hypothesis: We observe that for some HFL pairs, latent firings are "sticky" and occur in blocks. To quantify this, for a pair, we find the Markov transition matrix between the states

[3] PCA obtained from model activations on 1000 documents. The spike at $f \approx 10^{-3}$ is due to the <bos> token, which occurs about once every 1024 tokens, as the first PC direction is dominated by activations on the <bos>.

Latent	Total f	Hypothesis	Figures
#1	0.526	Context position	5, 10
#2, #4	0.888	Phrase-level semantic meaning	5, 11, 12
#3, #6	0.824	Context indicator	5, 13a
#5, #8	0.712	Context indicator	5, 13b
#7, #10	0.611	Context indicator	5, 13c
#9, #14	0.510	Body text vs. topic sentence	5, 14, 15

Table 1: HFL pairs with $f > 50\%$ and our hypotheses of their meanings, with figure links.

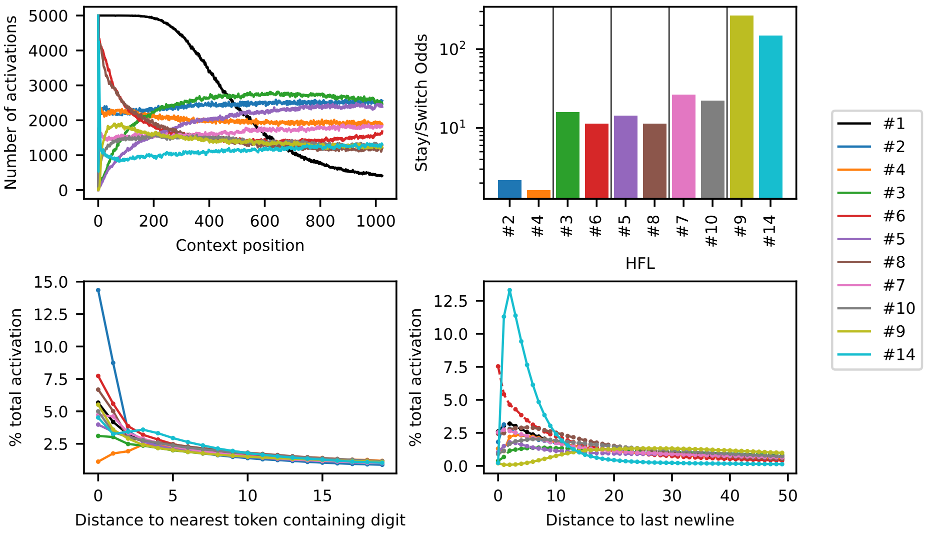


Figure 5: Unique aspects of latents. **Top left:** Context position vs. frequency; latent #1 has a clear dependence on context position. **Bottom left:** Percent of total activation strength vs. distance to the nearest digit; latents #2 and #4 have opposite correlations. **Top right:** "Stay/switch odds ratio" for HFL pairs; latents #2 and #4 are not sticky, but all other pairs are. **Bottom right:** Percent of total activation strength vs. distance to last newline token; latents #14 and #9 have opposite correlations.

"neither fires", "HFL1 fires", and "HFL2 fires". For each HFL, we compute the "stay/switch odds ratio", which is $P(\text{stay on}|\text{currently on})/P(\text{switch to pair}|\text{currently on})$. We indeed observe that for certain pairs, such as (#3,#6) and (#9,#14), once one latent is firing, the probability to switch to the other latent is very low; these results are shown in the top right of Figure 5. We hypothesize that these latents are "context indicators", and indicate some property of the text that persists across many tokens, such as "this is a paragraph about this concept" or "this context is of a certain type".

Body vs. Topic Sentence Detection (#9, #14): One particularly interpretable context indicator HFL pair is #9 and #14, which seems to indicate "body text" vs "topic sentences". We quantify this by examining the propensity to fire versus the distance from the last newline. The bottom right of Figure 5) shows that #14 fires strongly after newlines (but not trivially so, see Figure 15) while #9 fires minimally.

4 CONCLUSION

In this work, we argue that high frequency latents likely represent true dense language model features, and that because SAEs are optimized for sparsity, it is not surprising that these features are not learned cleanly by SAEs. Overall, we believe that our work provides a way forward to understand HFLs: if we carefully focus on each HFL one by one, they seem feasible to interpret. We also can try rotating the HFL space to result in maximally interpretable latents by some metric, as we hypothesize that the HFL subspace's current orientation may not line up with the model's ontology.

REFERENCES

Anthropic. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic, 2024.

Anthropic. Circuits Updates - June 2024, 2024. URL: `https://transformer-circuits.pub/2024/june-update/index.html`.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Róbert Csordás, Christopher Potts, Christopher D Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations. *arXiv preprint arXiv:2408.10920*, 2024.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023a. URL `https://arxiv.org/abs/2309.08600`.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023b.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

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. *arXiv preprint arXiv:2406.04093*, 2024.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*, 2024.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders, 2024.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Yuxiao Li, Eric J Michaud, David D Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *arXiv preprint arXiv:2410.19750*, 2024.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, 2024. URL `https://arxiv.org/abs/2408.05147`.

Jake Mendel. Sae feature geometry is outside the superposition hypothesis. *AI Alignment Forum*, 2024. URL https://www.alignmentforum.org/posts/MFBTjb2qf3ziWmzz6/sae-feature-geometry-is-outside-the-superposition-hypothesis.

Eric J Michaud, Isaac Liao, Vedang Lad, Ziming Liu, Anish Mudide, Chloe Loughridge, Zifan Carl Guo, Tara Rezaei Kheirkhah, Mateja Vukelić, and Max Tegmark. Opening the ai black box: program synthesis via mechanistic interpretability. *arXiv preprint arXiv:2402.05110*, 2024.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013a. URL https://arxiv.org/abs/1301.3781.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013b.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

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

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Abiy Tasissa, Manos Theodosis, Bahareh Tolooshams, and Demba E Ba. Discriminative reconstruction via simultaneous dense and sparse coding. *Transactions on Machine Learning Research*.

Gemma Team. Gemma. 2024. doi: 10.34740/KAGGLE/M/3301. URL https://www.kaggle.com/m/3301.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei

Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL `https://arxiv.org/abs/2408.00118`.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL `https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html`.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*, 2024.

Adam Yedidia. Gpt-2's positional embedding matrix is a helix, 2023a. URL `https://www.lesswrong.com/posts/qvWP3aBDBaqXvPNhS/gpt-2-s-positional-embedding-matrix-is-a-helix`. Accessed: 2024-09-16.

Adam Yedidia. The positional embedding matrix and previous-token heads: how do they actually work?, 2023b. URL `https://www.lesswrong.com/posts/zRA8B2FJLtTYRgie6/the-positional-embedding-matrix-and-previous-token-heads-how`. Accessed: 2024-09-17.

A Related Work

A.1 Investigating Sparse Autoencoder Latents

Cunningham et al. (2023a) and Bricken et al. (2023) demonstrate that SAEs can decompose model activations into a set of over-complete latents that are *more monosemantic* than the original neuron basis. To quantify this monosemanticity, these works perform extensive analysis of individual latents, both manually and using language models, and generate *explanations* of each latent using input text that they activate on. This work has been extended by Paulo et al. (2024), who develop more efficient methods for doing this "automated-interpretability" process at scale. Our work follows in this tradition of carefully examining individual SAE latents.

However, not all latents are interpretable, and SAEs do not capture all types of language model features. Dense latents, those that occur on greater than 10% of examples (and the subject of this work) were identified by both the early SAE work mentioned above and by Anthropic (2024). The phenomenon of "dark matter" in SAEs—features that exist in the model but are persistently not captured by the SAE as it scales—was first identified by Gao et al. (2024) and explored further by Engels et al. (2024).

SAEs also struggle with learning even sparse, interpretable features. Chanin et al. (2024) study feature splitting–the phenomenon that as SAEs scale the sparsity penalty causes interpretable latents to split into more specific versions–and feature absorption–the phenomenon that latents merge with other latents they frequently co-occur with. We hypothesize that these phenomenon occur to some extent with the HFL subspace identified in this work: if the sparsity is made extremely low, the HFL subspace is broken apart, even if it is truly dense.

A.2 Language Model Feature Structure

Understanding the geometry of language model representations has been an active area of research since the early days of word embeddings. Mikolov et al. (2013b) show that word embeddings exhibit linear structure and have semantic relationships that can be captured through vector arithmetic. Elhage et al. (2022) formalize this observation as the Linear Representation Hypothesis (LRH) and develop toy models to understand how networks might learn to represent an over-complete basis of linear features in a lower-dimensional space through superposition. Park et al. (2023) provide further theoretical grounding for when and why linear features might emerge in language models.

However, some recent work has challenged the universality of linear representations. Engels et al. (2024) demonstrate that some language model features are fundamentally multi-dimensional, and Csordás et al. (2024) find non-linear onion-like features in a toy RNN. On the other hand, while Li et al. (2022) study models trained on Othello and find that some representations seems to be nonlinearly encoded in the model, Nanda et al. (2023) find that with a change of basis these same variables are actually linearly encoded.

The geometry of features learned by sparse autoencoders has also received attention. Mendel (2024) argues that SAE features are not merely almost orthogonal random vectors, and that the relationships between them matter, while Li et al. (2024) studies SAE latent structure at many different layers of structure. In this work, we do not seek to propose an entirely new theory for how language model features and SAE latents are structured, but instead we just aim to explain high frequency features by postulating that they represent truly dense language model features.

A.3 Dense Model Representations

While much work has focused on sparse features in language models, there is a scattering of evidence for dense representations. Yedidia (2023a;b) finds that GPT-2's positional embeddings form a helix structure; the "height" of the helix is thus a feature that always activates. Gurnee et al. (2024) identify "universal neurons" that participate in many dense tasks, including encoding the context position and allowing the model to increase uncertainty ("entropy" neurons). Michaud et al. (2024) find structured lattice-like representations in toy models, an example of possible dense geometric structure. Mixed dense-sparse representations have also been explored in classical sparse coding:

Tasissa et al. develop dictionary learning methods for learning combined dense and sparse representations.

B METHODOLOGY

TopK SAEs: SAEs reconstruct model activations $\mathbf{x}$ as a sparse linear combination of a set of d_{SAE} latents, with $d_{\text{SAE}} \gg d_{\text{model}}$. The rows of the SAE decoder matrix correspond to SAE latent directions. We use TopK SAEs, which enforce sparsity by zeroing out the activations of all but the top k latent pre-activations $\mathbf{f}$:

$$\mathbf{f} = \text{ReLU}\left(W_e(\mathbf{x} - \mathbf{b_d}) + \mathbf{b_e}\right) \tag{1}$$

$$\hat{\mathbf{x}} = W_d\mathbf{f} + \mathbf{b_d} \tag{2}$$

For our main paper experiments, we train TopK SAEs with $d_{\text{SAE}} = 16384$ and $k = 64$.

Latent activations: We obttain latent activations by running Gemma on 5000 documents with context length of 1024 tokens from the shuffled RedPajama 1T sample (Weber et al., 2024), and capture all SAE latent activations on layer 12. High-frequency latents are defined as those that occur on $> 10\%$ of tokens.

SAE training: We train SAEs on the layer 12 activations on 1B tokens from the RedPajama 1T sample dataset, with 4 random seeds: 42, 43, 44, and 45. For our studies where we ablate part of the HFL subspace, we first subtract the learned $\mathbf{b_d}$, ablate the HFL subspace, add back $\mathbf{b_d}$, and then train the SAE on that.

C FURTHER ANALYSIS OF HFL CHARACTERISTICS

Linear probing: Since we suspect true dense features may be stored in the HFL subspace, but not exactly aligned with the learned HFLs, we attempt to probe true features by running a linear regression on the model activations with the "true" feature. We repeat this for the model activations projected only onto the HFL ($f > 0.1$) encoder subspace, and its complement.

Preliminary analysis finds that for context position, probing on the HFL subspace indeed performs better ($r^2 \approx 0.55$) than its complement ($r^2 \approx 0.45$) and on a random subspace spanned by the same number of latents ($r^2 \approx 0.05$), despite the context position not necessarily being linearly represented. However, attempting to probe other "true" features such as token length and log(token frequency) have not shown much success, and one reason could be that the "true" features are not necessarily token-level.

Quantifying semantic predictability: We train a `word2vec` (Mikolov et al., 2013a; Řehůřek & Sojka, 2010) model on all Gemma tokenized tokens (removing spaces) in the RedPajama dataset with window = 7, to generate the 300-dim embedding of the n tokens. This is concatenated into a $300n$-dim vector and passed through a logistic regression or simple MLP. Preliminary testing on some non-context-position HFLs such as #2 with $n = 5$ gives a balanced accuracy of around 70%, compared to the baseline test on the context position HFL #1 that gives a balanced accuracy of around 50%, suggesting a greater degree of semantic predictability.

Latent 1			Latent 2			Pair		
ID	#	Freq.	ID	#	Freq.	Cosine Sim.	Total Freq.	Jaccard Sim.
12723	1	0.526	-	-	-	-	0.526	-
13312	2	0.485	10028	4	0.403	-0.991	0.888	0
2837	3	0.483	600	6	0.341	-0.930	0.824	0
5019	5	0.393	10751	8	0.319	-0.972	0.712	0
8286	7	0.333	3817	10	0.279	-0.968	0.611	0
15161	9	0.283	6324	14	0.227	-0.892	0.510	0
3208	11	0.263	2242	12	0.238	-0.957	0.501	0
9929	13	0.233	12628	16	0.209	-0.925	0.442	0
11298	15	0.218	9900	18	0.155	-0.874	0.373	0
6279	17	0.176	14012	19	0.148	-0.934	0.324	0
11549	20	0.138	10430	165	0.026	-0.861	0.163	1.59×10^{-6}
522	21	0.102	-	-	-	-	0.102	-

Table 2: Table with HFLs, shown by their IDs and frequency ranks. For pairs, we show their cosine similarity, total frequency, and Jaccard similarity.

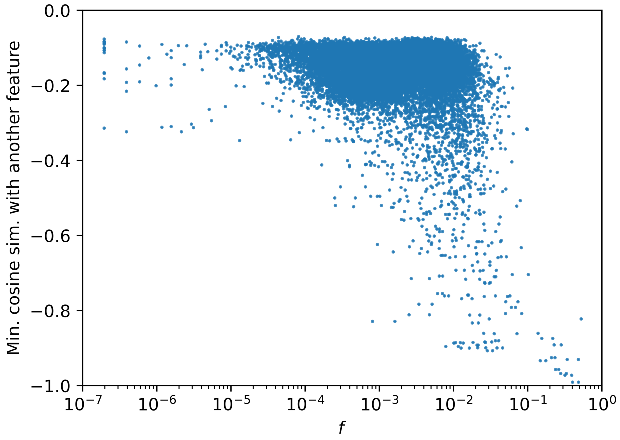


Figure 6: For each latent in seed=42, we find the minimum cosine similarity it has with another feature, and plot against frequency of the latent. We observe a distinct high-frequency group with highly negative minimum cosine similarity, while average features have minimum cosine similarity closer to zero.

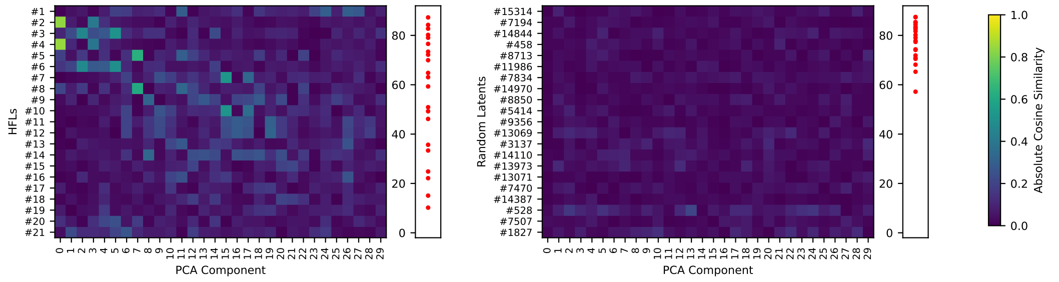


Figure 7: **Left:** Absolute cosine similarity between HFLs and the first 30 PC components, with the principal angles between the subspaces shown. **Right:** A similar plot for random features. We notice that while the HFLs are not closely aligned with the PCA components, they span a more similar subspace than random latents.

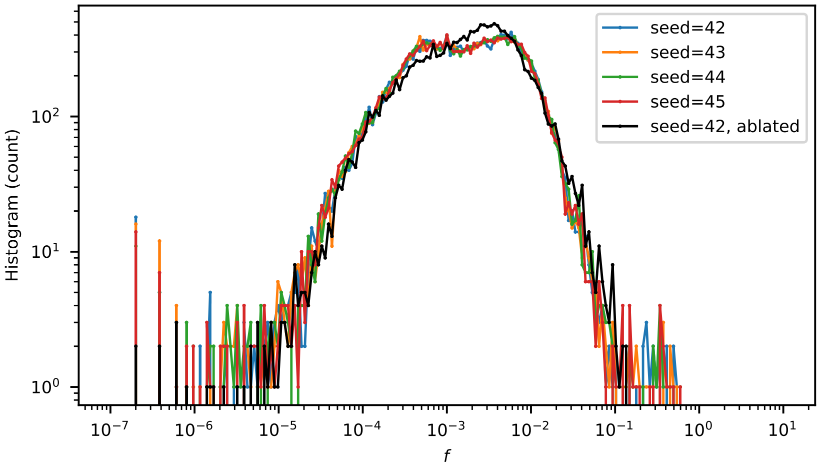
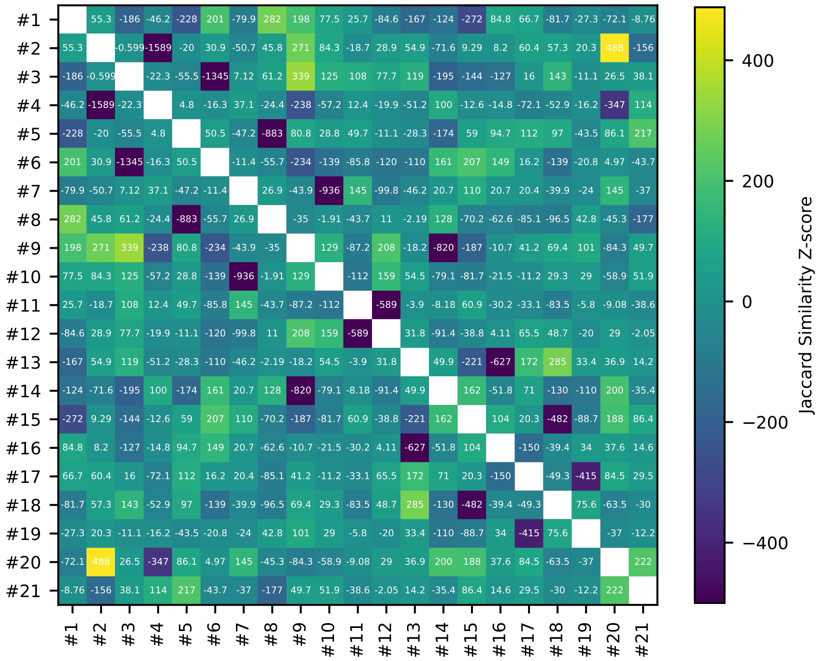


Figure 8: Histogram (count) of feature frequencies, in 200 log-spaced bins from 10^{-7} to 10^0. We observe that across all seeds, there are a handful of HFLs that fire $> 10\%$ of the time. However, when this HFL subspace is ablated, almost no latents discovered fire $> 10\%$.



	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	#18	#19	#20	#21
#1		55.3	-186	-46.2	-228	201	-79.9	282	198	77.5	25.7	-84.6	-167	-124	-272	84.8	66.7	-81.7	-27.3	-72.1	-8.76
#2	55.3		-0.599	-1589	-20	30.9	-50.7	45.8	271	84.3	-18.7	28.9	54.9	-71.6	9.29	8.2	60.4	57.3	20.3	488	-156
#3	-186	-0.599		-22.3	-55.5	-1345	7.12	61.2	339	125	108	77.7	119	-195	-144	-127	16	143	-11.1	26.5	38.1
#4	-46.2	-1589	-22.3		4.8	-16.3	37.1	-24.4	-238	-57.2	12.4	-19.9	-51.2	100	-12.6	-14.8	-72.1	-52.9	-16.2	-347	114
#5	-228	-20	-55.5	4.8		50.5	-47.2	-883	80.8	28.8	49.7	-11.1	-28.3	-174	59	94.7	112	97	-43.5	86.1	217
#6	201	30.9	-1345	-16.3	50.5		-11.4	-55.7	-234	-139	-85.8	-120	-110	161	207	149	16.2	-139	-20.8	4.97	-43.7
#7	-79.9	-50.7	7.12	37.1	-47.2	-11.4		26.9	-43.9	-936	145	-99.8	-46.2	20.7	110	20.7	20.4	-39.9	-24	145	-37
#8	282	45.8	61.2	-24.4	-883	-55.7	26.9		-35	-1.91	-43.7	11	-2.19	128	-70.2	-62.6	-85.1	-96.5	42.8	-45.3	-177
#9	198	271	339	-238	80.8	-234	-43.9	-35		129	-87.2	208	-18.2	-820	-187	-10.7	41.2	69.4	101	-84.3	49.7
#10	77.5	84.3	125	-57.2	28.8	-139	-936	-1.91	129		-112	159	54.5	-79.1	-81.7	-21.5	-11.2	29.3	29	-58.9	51.9
#11	25.7	-18.7	108	12.4	49.7	-85.8	145	-43.7	-87.2	-112		-589	-3.9	-8.18	60.9	-30.2	-33.1	-83.5	-5.8	-9.08	-38.6
#12	-84.6	28.9	77.7	-19.9	-11.1	-120	-99.8	11	208	159	-589		31.8	-91.4	-38.8	4.11	65.5	48.7	-20	29	-2.05
#13	-167	54.9	119	-51.2	-28.3	-110	-46.2	-2.19	-18.2	54.5	-3.9	31.8		49.9	-221	-627	172	285	33.4	36.9	14.2
#14	-124	-71.6	-195	100	-174	161	20.7	128	-820	-79.1	-8.18	-91.4	49.9		162	-51.8	71	-130	-110	200	-35.4
#15	-272	9.29	-144	-12.6	59	207	110	-70.2	-187	-81.7	60.9	-38.8	-221	162		104	20.3	-482	-88.7	188	86.4
#16	84.8	8.2	-127	-14.8	94.7	149	20.7	-62.6	-10.7	-21.5	-30.2	4.11	-627	-51.8	104		-150	-39.4	34	37.6	14.6
#17	66.7	60.4	16	-72.1	112	16.2	20.4	-85.1	41.2	-11.2	-33.1	65.5	172	71	20.3	-150		-49.3	-415	84.5	29.5
#18	-81.7	57.3	143	-52.9	97	-139	-39.9	-96.5	69.4	29.3	-83.5	48.7	285	-130	-482	-39.4	-49.3		75.6	-63.5	-30
#19	-27.3	20.3	-11.1	-16.2	-43.5	-20.8	-24	42.8	101	29	-5.8	-20	33.4	-110	-88.7	34	-415	75.6		-37	-12.2
#20	-72.1	488	26.5	-347	86.1	4.97	145	-45.3	-84.3	-58.9	-9.08	29	36.9	200	188	37.6	84.5	-63.5	-37		222
#21	-8.76	-156	38.1	114	217	-43.7	-37	-177	49.7	51.9	-38.6	-2.05	14.2	-35.4	86.4	14.6	29.5	-30	-12.2	222	

Figure 9: Jaccard similarity Z-score between latents' activations, ranked by their frequencies. The expected Jaccard similarity is calculated by randomly shuffling the activations across all tokens, repeating 50 times to obtain a mean and standard deviation. We observe that HFL pairs which never co-occur have an extremely low z-score. Between some non-pair HFLs, Jaccard similarity z-scores are high, indicating some relationship in their firing pattern.

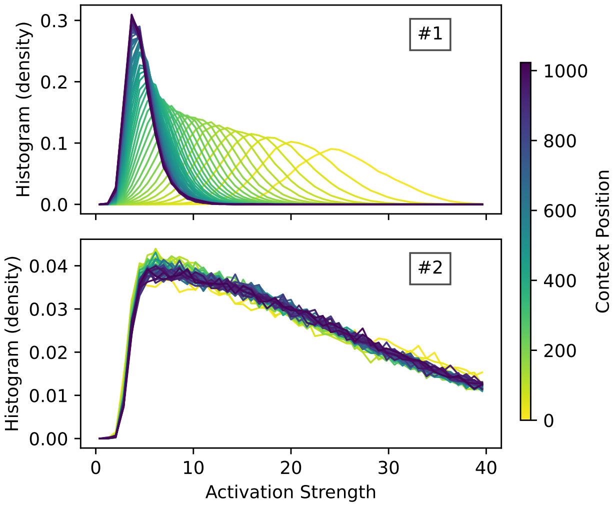
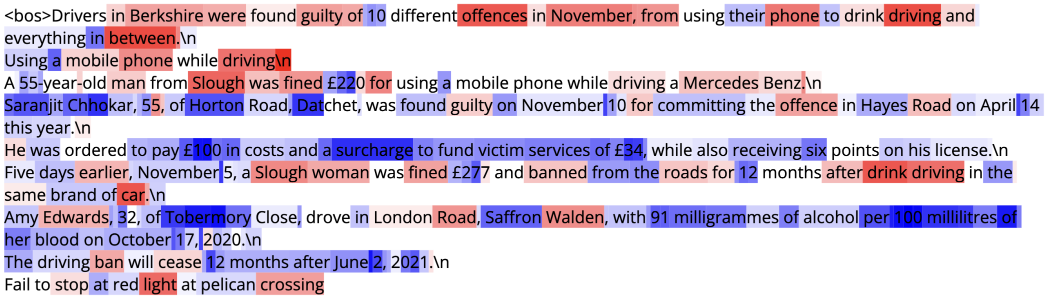


Figure 10: Histogram of activation strength when active, binned by context position in groups of 20. Not only does #1 fire more often early in contexts as shown in Figure 5, it also fires strongly early in contexts as shown here. Meanwhile, #2 has a more semantic meaning, and thus shows no clear dependence in firing strength on context position.

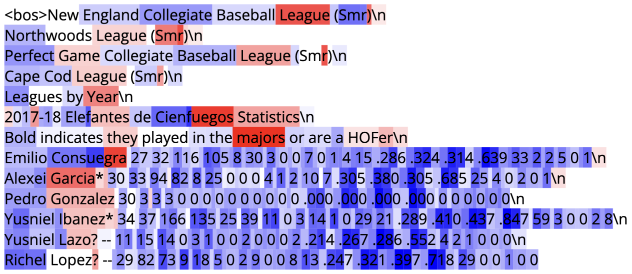
D HFL Examples

In this section, we present some examples of text, shaded by a pair's firing pattern. Pairs never activate on the same token, and un-highlighted text means neither latent activated. The highlight color is scaled by the maximum activation in that document's context (excluding the <bos> and first tokens which tend to have high activations).



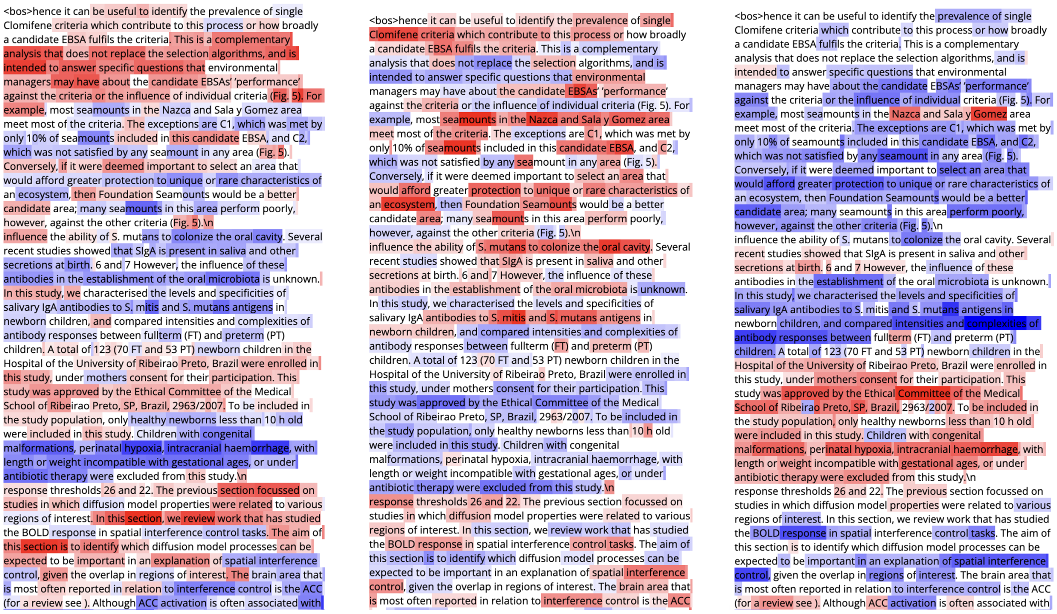
<bos>Drivers in Berkshire were found guilty of 10 different offences in November, from using their phone to drink driving and everything in between.\n
Using a mobile phone while driving\n
A 55-year-old man from Slough was fined £220 for using a mobile phone while driving a Mercedes Benz.\n
Saranjit Chhokar, 55, of Horton Road, Datchet, was found guilty on November 10 for committing the offence in Hayes Road on April 14 this year.\n
He was ordered to pay £100 in costs and a surcharge to fund victim services of £34, while also receiving six points on his license.\n
Five days earlier, November 5, a Slough woman was fined £277 and banned from the roads for 12 months after drink driving in the same brand of car.\n
Amy Edwards, 32, of Tobermory Close, drove in London Road, Saffron Walden, with 91 milligrammes of alcohol per 100 millilitres of her blood on October 17, 2020.\n
The driving ban will cease 12 months after June 2, 2021.\n
Fail to stop at red light at pelican crossing

Figure 11: Blue = #2, red = #4. We observe that #2 tends to fire on numbers and short, common words, as well as some names, while #4 fires more on "meaningful" words in the context.



<bos>New England Collegiate Baseball League (Smr)\n
Northwoods League (Smr)\n
Perfect Game Collegiate Baseball League (Smr)\n
Cape Cod League (Smr)\n
Leagues by Year\n
2017-18 Elefantes de Cienfuegos Statistics\n
Bold indicates they played in the majors or are a HOFer\n
Emilio Consuegra 27 32 116 105 8 30 3 0 0 7 0 1 4 15 .286 .324 .314 .639 33 2 2 5 0 1\n
Alexei Garcia* 30 33 94 82 8 25 0 0 0 4 1 2 10 7 .305 .380 .305 .685 25 4 0 2 0 1\n
Pedro Gonzalez 30 3 3 3 0 0 0 0 0 0 0 0 0 0 .000 .000 .000 .000 0 0 0 0 0 0\n
Yusniel Ibanez* 34 37 166 135 25 39 11 0 3 14 1 0 29 21 .289 .410 .437 .847 59 3 0 0 2 8\n
Yusniel Lazo? -- 11 15 14 0 3 1 0 0 2 0 0 0 2 .214 .267 .286 .552 4 2 1 0 0 0\n
Richel Lopez? -- 29 82 73 9 18 5 0 2 9 0 0 8 13 .247 .321 .397 .718 29 0 0 1 0 0

Figure 12: Blue = #2, red = #4. #2 fires strongly in the table with numeric data.



<bos>hence it can be useful to identify the prevalence of single Clomifene criteria which contribute to this process or how broadly a candidate EBSA fulfils the criteria. This is a complementary analysis that does not replace the selection algorithms, and is intended to answer specific questions that environmental managers may have about the candidate EBSAs' 'performance' against the criteria or the influence of individual criteria (Fig. 5). For example, most seamounts in the Nazca and Sala y Gomez area meet most of the criteria. The exceptions are C1, which was met by only 10% of seamounts included in this candidate EBSA, and C2, which was not satisfied by any seamount in any area (Fig. 5). Conversely, if it were deemed important to select an area that would afford greater protection to unique or rare characteristics of an ecosystem, then Foundation Seamounts would be a better candidate area; many seamounts in this area perform poorly, however, against the other criteria (Fig. 5).\n

influence the ability of S. mutans to colonize the oral cavity. Several recent studies showed that SIgA is present in saliva and other secretions at birth. 6 and 7 However, the influence of these antibodies in the establishment of the oral microbiota is unknown. In this study, we characterised the levels and specificities of salivary IgA antibodies to S. mitis and S. mutans antigens in newborn children, and compared intensities and complexities of antibody responses between fullterm (FT) and preterm (PT) children. A total of 123 (70 FT and 53 PT) newborn children in the Hospital of the University of Ribeirao Preto, Brazil were enrolled in this study, under mothers consent for their participation. This study was approved by the Ethical Committee of the Medical School of Ribeirao Preto, SP, Brazil, 2963/2007. To be included in the study population, only healthy newborns less than 10 h old were included in this study. Children with congenital malformations, perinatal hypoxia, intracranial haemorrhage, with length or weight incompatible with gestational ages, or under antibiotic therapy were excluded from this study.\n

response thresholds 26 and 22. The previous section focussed on studies in which diffusion model properties were related to various regions of interest. In this section, we review work that has studied the BOLD response in spatial interference control tasks. The aim of this section is to identify which diffusion model processes can be expected to be important in an explanation of spatial interference control, given the overlap in regions of interest. The brain area that is most often reported in relation to interference control is the ACC (for a review see). Although ACC activation is often associated with

(a) Blue = #3, red = #6

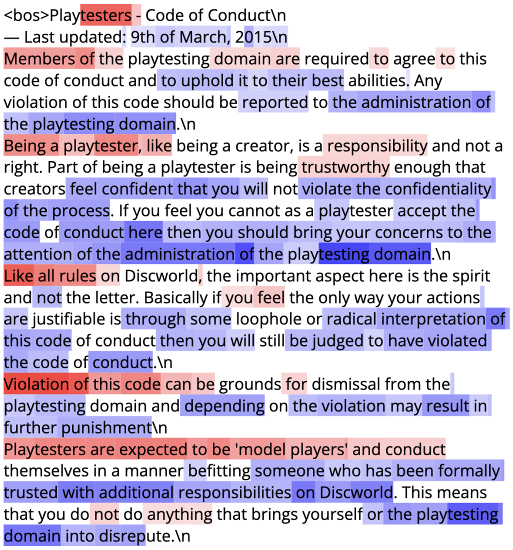
<bos>hence it can be useful to identify the prevalence of single Clomifene criteria which contribute to this process or how broadly a candidate EBSA fulfils the criteria. This is a complementary analysis that does not replace the selection algorithms, and is intended to answer specific questions that environmental managers may have about the candidate EBSAs' 'performance' against the criteria or the influence of individual criteria (Fig. 5). For example, most seamounts in the Nazca and Sala y Gomez area meet most of the criteria. The exceptions are C1, which was met by only 10% of seamounts included in this candidate EBSA, and C2, which was not satisfied by any seamount in any area (Fig. 5). Conversely, if it were deemed important to select an area that would afford greater protection to unique or rare characteristics of an ecosystem, then Foundation Seamounts would be a better candidate area; many seamounts in this area perform poorly, however, against the other criteria (Fig. 5).\n
influence the ability of S. mutans to colonize the oral cavity. Several recent studies showed that SIgA is present in saliva and other secretions at birth. 6 and 7 However, the influence of these antibodies in the establishment of the oral microbiota is unknown. In this study, we characterised the levels and specificities of salivary IgA antibodies to S. mitis and S. mutans antigens in newborn children, and compared intensities and complexities of antibody responses between fullterm (FT) and preterm (PT) children. A total of 123 (70 FT and 53 PT) newborn children in the Hospital of the University of Ribeirao Preto, Brazil were enrolled in this study, under mothers consent for their participation. This study was approved by the Ethical Committee of the Medical School of Ribeirao Preto, SP, Brazil, 2963/2007. To be included in the study population, only healthy newborns less than 10 h old were included in this study. Children with congenital malformations, perinatal hypoxia, intracranial haemorrhage, with length or weight incompatible with gestational ages, or under antibiotic therapy were excluded from this study.\n
response thresholds 26 and 22. The previous section focussed on studies in which diffusion model properties were related to various regions of interest. In this section, we review work that has studied the BOLD response in spatial interference control tasks. The aim of this section is to identify which diffusion model processes can be expected to be important in an explanation of spatial interference control, given the overlap in regions of interest. The brain area that is most often reported in relation to interference control is the ACC

(b) Blue = #5, red = #8

<bos>hence it can be useful to identify the prevalence of single Clomifene criteria which contribute to this process or how broadly a candidate EBSA fulfils the criteria. This is a complementary analysis that does not replace the selection algorithms, and is intended to answer specific questions that environmental managers may have about the candidate EBSAs' 'performance' against the criteria or the influence of individual criteria (Fig. 5). For example, most seamounts in the Nazca and Sala y Gomez area meet most of the criteria. The exceptions are C1, which was met by only 10% of seamounts included in this candidate EBSA, and C2, which was not satisfied by any seamount in any area (Fig. 5). Conversely, if it were deemed important to select an area that would afford greater protection to unique or rare characteristics of an ecosystem, then Foundation Seamounts would be a better candidate area; many seamounts in this area perform poorly, however, against the other criteria (Fig. 5).\n
influence the ability of S. mutans to colonize the oral cavity. Several recent studies showed that SIgA is present in saliva and other secretions at birth. 6 and 7 However, the influence of these antibodies in the establishment of the oral microbiota is unknown. In this study, we characterised the levels and specificities of salivary IgA antibodies to S. mitis and S. mutans antigens in newborn children, and compared intensities and complexities of antibody responses between fullterm (FT) and preterm (PT) children. A total of 123 (70 FT and 53 PT) newborn children in the Hospital of the University of Ribeirao Preto, Brazil were enrolled in this study, under mothers consent for their participation. This study was approved by the Ethical Committee of the Medical School of Ribeirao Preto, SP, Brazil, 2963/2007. To be included in the study population, only healthy newborns less than 10 h old were included in this study. Children with congenital malformations, perinatal hypoxia, intracranial haemorrhage, with length or weight incompatible with gestational ages, or under antibiotic therapy were excluded from this study.\n
response thresholds 26 and 22. The previous section focussed on studies in which diffusion model properties were related to various regions of interest. In this section, we review work that has studied the BOLD response in spatial interference control tasks. The aim of this section is to identify which diffusion model processes can be expected to be important in an explanation of spatial interference control, given the overlap in regions of interest. The brain area that is most often reported in relation to interference control is the ACC (for a review see). Although ACC activation is often associated with

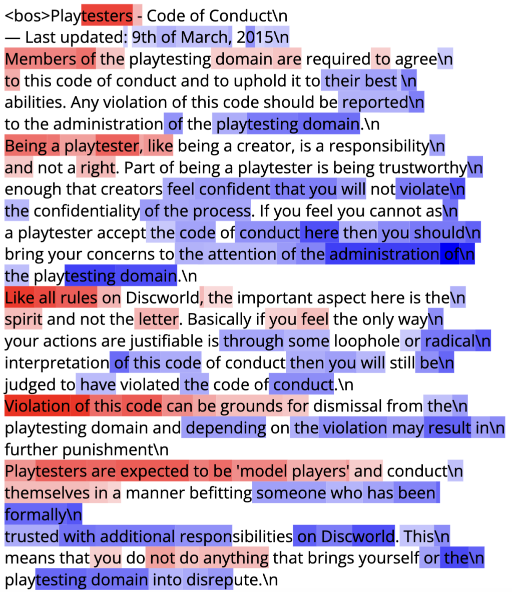
(c) Blue = #7, red = #10

Figure 13: Firing pattern of "sticky" feature pairs, on same text. While it is difficult to concretely interpret what these features represent, they seem to correspond to specific general concepts present in the context. For instance, biological facts, scientific analysis / self-reference in text, and information on the study or experiments.



<bos>Playtesters - Code of Conduct\n
— Last updated: 9th of March, 2015\n
Members of the playtesting domain are required to agree to this code of conduct and to uphold it to their best abilities. Any violation of this code should be reported to the administration of the playtesting domain.\n
Being a playtester, like being a creator, is a responsibility and not a right. Part of being a playtester is being trustworthy enough that creators feel confident that you will not violate the confidentiality of the process. If you feel you cannot as a playtester accept the code of conduct here then you should bring your concerns to the attention of the administration of the playtesting domain.\n
Like all rules on Discworld, the important aspect here is the spirit and not the letter. Basically if you feel the only way your actions are justifiable is through some loophole or radical interpretation of this code of conduct then you will still be judged to have violated the code of conduct.\n
Violation of this code can be grounds for dismissal from the playtesting domain and depending on the violation may result in further punishment\n
Playtesters are expected to be 'model players' and conduct themselves in a manner befitting someone who has been formally trusted with additional responsibilities on Discworld. This means that you do not do anything that brings yourself or the playtesting domain into disrepute.\n

Figure 14: Latent pair #9 and #14 firing on a context (#9 in blue and #14 in red). Latent #14 seems to fire on topic sentences, while latent #9 seems to fire on body text.



<bos>Playtesters - Code of Conduct\n
— Last updated: 9th of March, 2015\n
Members of the playtesting domain are required to agree\n
to this code of conduct and to uphold it to their best \n
abilities. Any violation of this code should be reported\n
to the administration of the playtesting domain.\n
Being a playtester, like being a creator, is a responsibility\n
and not a right. Part of being a playtester is being trustworthy\n
enough that creators feel confident that you will not violate\n
the confidentiality of the process. If you feel you cannot as\n
a playtester accept the code of conduct here then you should\n
bring your concerns to the attention of the administration of\n
the playtesting domain.\n
Like all rules on Discworld, the important aspect here is the\n
spirit and not the letter. Basically if you feel the only way\n
your actions are justifiable is through some loophole or radical\n
interpretation of this code of conduct then you will still be\n
judged to have violated the code of conduct.\n
Violation of this code can be grounds for dismissal from the\n
playtesting domain and depending on the violation may result in\n
further punishment\n
Playtesters are expected to be 'model players' and conduct\n
themselves in a manner befitting someone who has been
formally\n
trusted with additional responsibilities on Discworld. This\n
means that you do not do anything that brings yourself or the\n
playtesting domain into disrepute.\n

Figure 15: The latents seem robust to artificially adding newlines; they are doing something more than just detecting newlines.

E Generalization of Main Results

In Figure 16, Figure 17, and Figure 18 we reproduce our key results on feature geometry on the Gemma Scope (Lieberum et al., 2024) and Llama Scope (He et al., 2024) suites of SAEs, and find that HFLs are indeed discovered in different SAEs and different models (focusing on middle model layers).

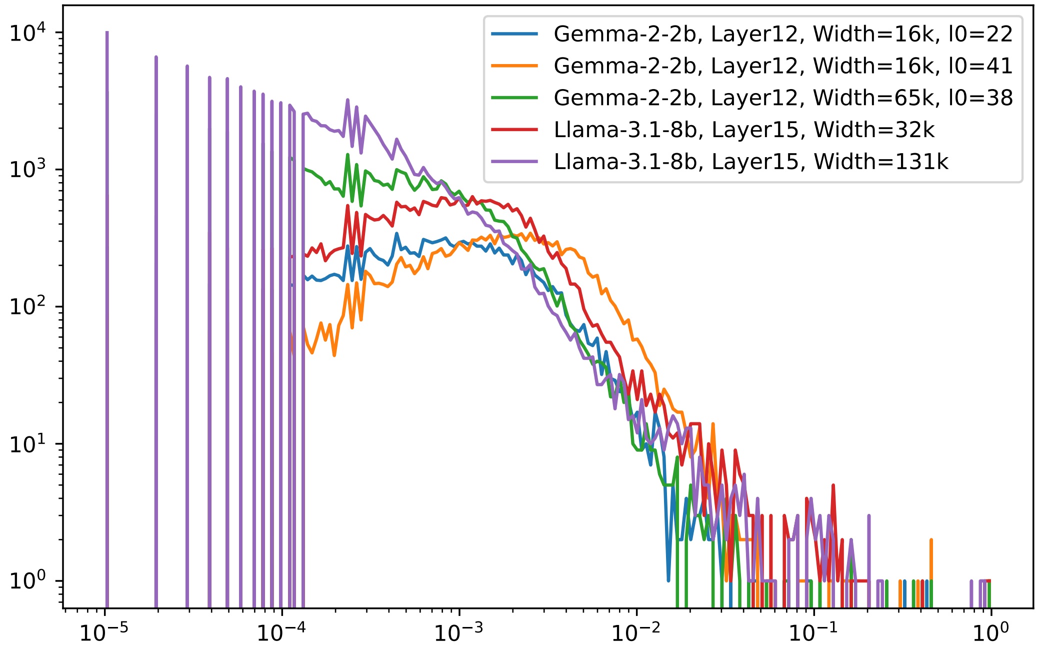


Figure 16: Frequency histogram for different SAEs on different models. We notice that in all SAEs, there are a handful of features with frequencies $> 10\%$.

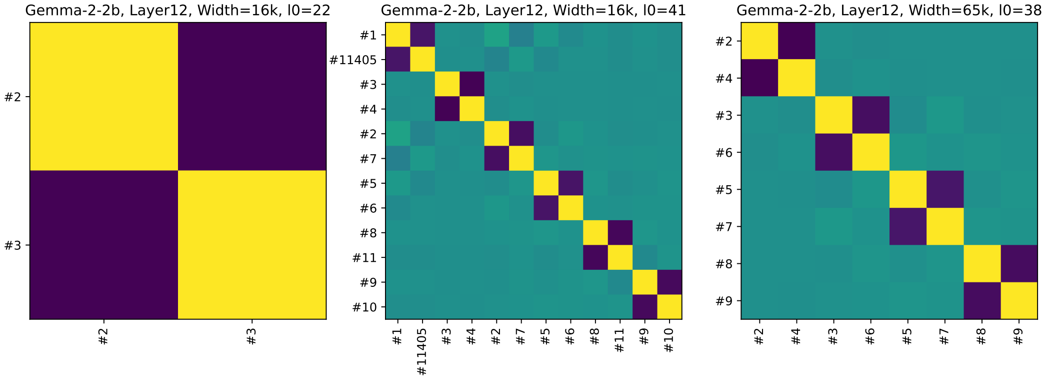


Figure 17: For each HFL ($f > 0.1$) in Gemma-2-2b SAEs of different widths and sparsities, we search for a geometrically opposite pair, and present their cosine similarities here.

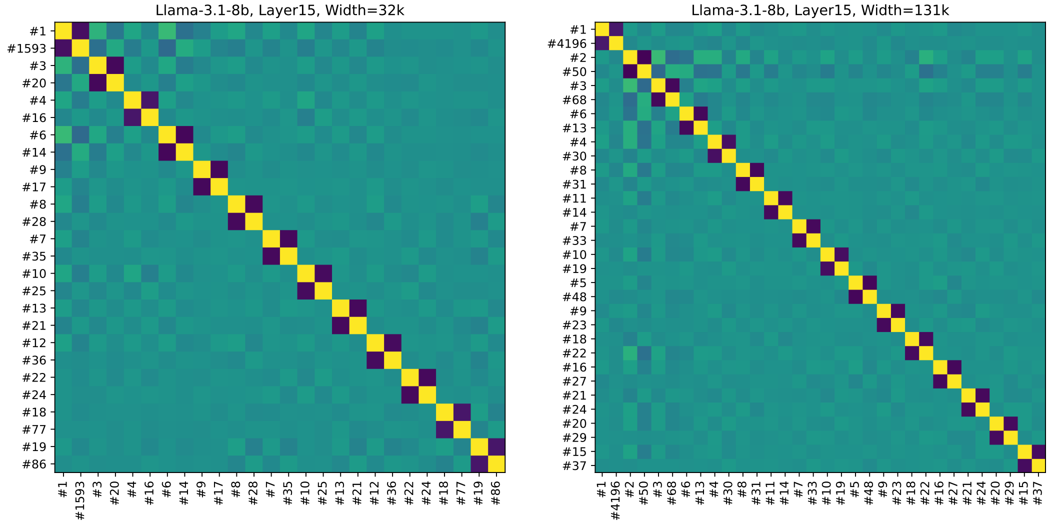


Figure 18: Similarly, for each HFL ($f > 0.1$) in Llama-3.1-8b SAEs of different widths, we search for a geometrically opposite pair, and present their cosine similarities.