Semantic Convergence: Investigating Shared Representations Across Scaled LLMs

Anonymous ACL submission

Abstract

We investigate feature universality in Gemma-2 language models (Gemma-2-2B & Gemma-2-9B), asking whether models with a fourfold difference in scale still converge on comparable internal concepts. Using the sparse autoencoder (SAE) dictionary learning pipeline, we used pretrained SAEs on each model's residual-stream activations, aligned the resulting monosemantic features via activation correlation, and compared the matched feature spaces with metrics such as SVCCA and RSA. Middle layers yield the strongest overlap, indicating that this is where both models most similarly represent concepts, while early and late layers show much less similarity. Preliminary experiments extending the analysis from single tokens to multi-token subspaces show that semantically similar subspaces tend to interact similarly with LLMs. These results offer further evidence that large language models carve the world into broadly similar, interpretable features despite size differences, reinforcing universality as a foundation for cross-model interpretability.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Guo et al., 2025) have demonstrated increasing reasoning abilities across many tasks (Bubeck et al., 2023). However, our understanding of the internal representations and computations that support this behavior remains limited (Bereska and Gavves, 2024).

Previous work (Lan et al., 2024) has shown that models with the same tokenizer rely on similar internal representations and structures, indicating that universal feature spaces might exist. We define a feature as being universal if its activation corresponds to the same semantic concept regardless of model size or architecture. Universal feature spaces may be encoded across different layers or neurons in different models, but there would exist a sparse direction in each model's activation space that consistently "lights up" on these concepts. Understanding universal feature spaces is a crucial step in learning if general rules govern how LLMs internally structure and use their nodes. This key would increase the generalizability of interpreting different LLMs and may accelerate LLM training efficiency as well as LLM safety (Chughtai et al., 2023; Gurnee et al., 2024; Bricken et al., 2023).

Comparing features between LLMs is challenging because nodes in the model usually represent multiple features, rather than one specific feature. This is called polysemanticity (Elhage et al., 2022). In this paper, we build on the methods shown in Lan et al. (2024) that leverage Sparse Autoencoders (SAEs) to transform LLM node activations into lower dimensional spaces that are easier to interpret. The main advantage of using SAEs is that they have the ability to decompose the complex, polysemantic representations in an LLM into distinct features that can be interpreted more easily (Cunningham et al., 2023; Bricken et al., 2023). Then, representational space similarity metrics are used on these SAE features to check for similarities in the internal structure of the LLM.

Although results for feature universality in Lan et al. (2024) were promising, only single token words, in a limited number of semantic subspaces, were tested for the semantic experiments. Furthermore, the experiments were only carried out against

similar sized models, namely Pythia-70m with Pythia-160m (Biderman et al., 2023) and Gemma-1-2B with Gemma-2-2B (Team et al., 2024b,a). Therefore, in this paper, we will further investigate the universality of feature spaces through the following key experiments:

1. Probe universality in multi-token semantic subspaces, including overlaps of related concepts, to see whether phrase-level and higher-order features align across models.
2. Quantify universality across a 4× model-size gap and compare similarity measures (e.g., SVCCA, RSA) to test how metric choice affects the result.

For our experiments, we use models with a four fold size difference. Our results demonstrate that the similarity in internal feature representations remains across these models despite difference in complexity. Furthermore, in our semantic subspaces studies, we show that there are certain groups of overlapping concepts that the models internally represent similarly. This is another indication of feature universality.

This work opens up several branches of future research that we believe are worth studying. Training SAEs on multiple model layers can reveal internal representations that are not captured in a single layer. In addition, comparing the internal representations of SAEs trained on MLP layers may provide deeper insights about the universality of MLP features. These findings can accelerate AI reasoning and safety training (Hendrycks et al., 2023). Through understanding the similarities between models, and their differences, a more complete picture of how LLMs process, reason and understand natural language would be formed (Lan et al., 2024).

2 Background

Sparse Autoencoders. Sparse Autoencoders (SAEs) are a type of neural network used to learn efficient, sparse representations of input data (Makhzani and Frey, 2013). Unlike other autoencoders, SAEs incorporate a sparsity constraint, typically an L1 penalty on the hidden layer activations or a KL divergence term, which pushes most hidden units to be inactive (i.e. any output values close to zero) for any input given. This leads to features that are more interpretable and disentangled. The aim is to discover a basis of features, similar to dictionary learning (Olshausen and Field, 1997), where each feature activates for semantically meaningful concepts.
Mathematically, an input $x \in \mathbb{R}^n$ is given to the neural network which is reconstructed into $\hat{x}$ using $\hat{x} = W'\sigma(Wx + b)$, where $W \in \mathbb{R}^{hxn}$ is the encoder weight matrix, b is the bias term, σ is a nonlinear activation function, and W' is the decoder matrix, which often uses the transpose of the encoder weights. SAE training seeks to both encourage sparsity in the activations $h = \sigma(Wx+b)$ and to minimize the reconstruction loss $L_{\text{rec}}(x, \hat{x}) = \|x - \hat{x}\|^2$.

3 Methods

3.1 Feature Pairings

To determine whether different models of varying sizes converge on similar internal representations, generalizations of feature spaces, spaces formed by feature groups, and feature relations must be explored. To quantitatively measure these similarities, we follow the methods of Lan et al. (2024). Overall, we compare an SAE trained on layer A_i from LLM A with another SAE trained on layer B_j from LLM B for every layer pairing.

However, accurate comparisons between spaces hinges on solving two issues:
Permutation issue. To solve the permutation issue, we find neuron pairings that are the most similar in SAE_A and SAE_B. Since the mapping of features is unknown due to arbitrary neuron indexing and some features may not have a "similar" feature in the other SAE, we pairwise match them using a correlation metric.
Rotational Alignment issue. Even after permutation alignment, each SAE may use its own orthonormal basis for latent space. To ensure that the true relational similarity is captured, we apply rotation-invariant similarity measures, namely SVCCA and RSA.

To score the results against a baseline, randomly paired features are obtained. Then,

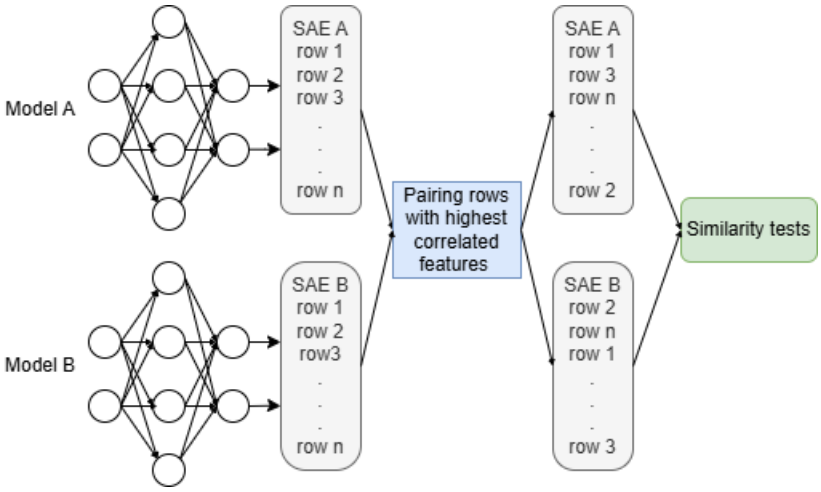

Figure 1: Workflow of pairing rows with the highest correlated features between two models (Gemma-2-2B and Gemma-2-9B) and performing similarity tests to assess feature alignment.

the score of the features paired by correlation (referred to as "paired features") is compared with the average score of N runs of randomly paired features to obtain a p-value score.

3.1.1 1-to-1 vs. Many-to-1 Feature Matching

Following Lan et al. (2024), we consider two ways of pairing SAE features from layer A_i of the first model with layer B_j of the second.

1-to-1 (bijective) matching. We iteratively build a one–to–one assignment: at each step we pick the still–unmatched pair of features with the highest Pearson correlation. Each feature is used at most once, yielding a bijection of size $K = \min(|A_i|, |B_j|)$.

Many-to-1 (injective) matching. To probe whether the entire dictionary of the smaller layer can be embedded inside the larger one, we relax the uniqueness constraint on layer B_j. Every feature in A_i is matched to its most-correlated partner in B_j, even if that target has already been claimed by other sources. Thus one feature in B_j may receive multiple links, while each feature in A_i is still matched exactly once.

Unless stated otherwise, all correlations are computed with Pearson correlation; the aligned pairs returned by the chosen strategy are then fed into the subsequent SVCCA and RSA calculations.

3.2 Representational Similarity Metrics

3.2.1 Singular Value Canonical Correlation Analysis (SVCCA)

Singular Value Canonical Correlation Analysis (Raghu et al., 2017) is a variation of the Canonical Correlation Analysis CCA (Hotelling, 1936) which finds a pair of the most correlated variables, u_i and v_i from two sets of variables $X \in \mathbb{R}^{nxd_1}$ and $Y \in \mathbb{R}^{nxd_2}$. Before applying CCA, SVCCA reduces noise by applying Singular Value Decomposition (SVD) to X and Y using $X = U_X S_X V_X^T$ and $Y = U_Y S_Y V_Y^T$, where U_X and U_Y are the matrices containing the left singular vectors (informative directions), and S_X and S_Y are diagonal matrices containing the singular values. After CCA is applied on the new data, correlation scores between the most informative components are obtained, which are then averaged to get a similarity score. SVCCA measures how well subspaces of two SAE weight matrices align, essentially quantifying the global feature space overlap.

3.2.2 Representational Similarity Analysis (RSA)

Representational Similarity Analysis (Kriegeskorte et al., 2008) calculates, for each space, a Representation Dissimilarity Matrix (RDM) $D \in \mathbb{R}^{nxn}$. Each element in this matrix represents the dissimilarity between every pair of data points in the space. Following RDM, a correlation metric such as Spearman's rank correlation coefficient is used

to compute a similarity score.

3.3 Semantic Subspaces

In addition to the layer-wise SAE comparisons, we also test semantic subspaces, collections of words defined by a high-level concept that contain concept-specific keywords. For example, "emotions" is a subspace with concept-specific keywords like "happy"and "sad". By testing these subspaces, we can evaluate whether LLMs encode the same semantic categories.

For each high-level concept, we first use GPT-4o (Achiam et al., 2023) to generate three independent lists of representative keywords. We then intersect these lists and retain only terms that are unambiguous (each having a single and clear meaning). Next, we add to the keyword set their hyponyms from WordNet (wor, 2010). This combined collection of keywords plus and their hyponyms defines the final semantic subspace for that concept.

To evaluate semantic subspaces more rigorously, we combine two different subspaces in the following ways:

Multi-token subspaces: In the multi-token subspaces, we concatenate keywords from different concepts together. For instance, "happy" (from "emotions") and "child" (from "person") becomes "happy child". In these types of subspaces, we aim to understand if different LLMs internally process longer sentences similarly. Furthermore, we concatenated unlikely pairs of concepts, such as "calender" and "emotions", which the LLMs were unlikely to see during their training to check whether the LLMs process previously unseen data in a similar manner.

Overlapping subspaces: Overlapping subspaces are formed by taking the union of whole subspaces together. For instance, the "emotions" and "person" subspaces would yield ("happy", "teacher", "sad", "child", ...). This aims to test if the different LLMs interact with multiple concepts similarly.

4 Experiments and Results

Layer-wise similarity of full SAE spaces

1-to-1 (5 run mean): In Fig. 2a the diagonal band of Paired SVCCA now peaks at 0.73 (Gemma-2-2B L14 and Gemma-2-9B L19) and stays consistently high across contiguous mid-layer pairs (0.64 – 0.71). Early layers sit at 0.35 ± 0.05, and the last decoder layer of Gemma-2-9B (L39) experiences a drop to an average of 0.374. Through these visualizations (Fig. 2a & Fig. 2b), it is observable that the middle layers between both models share the most similarity compared to other layers.

Paired RSA (Fig. 2b) follows the same shape but at roughly one-third of the magnitude: maxima of 0.22 and a midlayer plateau of 0.15 – 0.20, with edges staying less than 0.08. Yet again, the last decoder layer of Gemma-2-9B (L39) experiences a drop. This figure further displays the pattern of middle layers sharing the most similarity, especially compared to early and late layers.

Many-to-1 (Single run): When we allow duplicates (Fig. 4a) the peak SVCCA softens to 0.69 (Gemma-2-2B L14 and Gemma-2-9B L19) and the mid-layer plateau narrows (0.54 – 0.66). RSA follows suit, topping out at around 0.18 - 0.2. This confirms that when features are matched more than once, the alignment scores drop slightly, but the overall pattern is maintained.

Many-to-1 (5 run mean): Averaging five random initializations barely changes the picture (Fig. 6a): peak SVCCA = 0.69, peak RSA = 0.20. The variance across runs is <0.02 for every cell, indicating that the many-to-one procedure is stable, but still consistently lower than the ceiling of the 1-to-1 strategy.

Random-pair baselines and significance. Across the three experiments, the mean random SVCCA spans 0.005 – 0.034, with a majority of cells below 0.02 (Fig. 8a, Fig. 9a, Fig. 10a). Consequently, every empirical SVCCA score beyond the first residual layer lands in the 0.0 p-value bucket ($\leq 0.1\%$ chance of getting such a good alignment by random) (Fig. 8b, Fig. 9b, Fig. 10b). In other words, even the weakest observed alignment (SVCCA ≈ 0.30) is simply too strong to be by chance.

Effect of filtering non-concept features. Mean activation correlation before

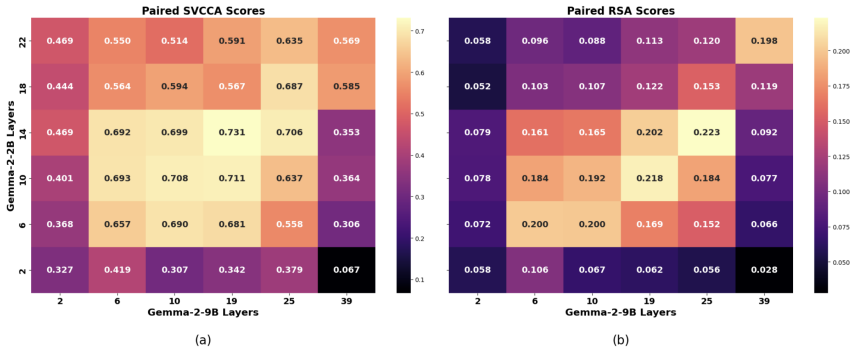

Figure 2: (a) SVCCA and (b) RSA 1-to-1 paired scores of SAEs for layers in Gemma-2-2B vs Gemma-2-9B. Note the pattern of higher scores between the middle layer pairings indicating similarity in middle layers between both models.

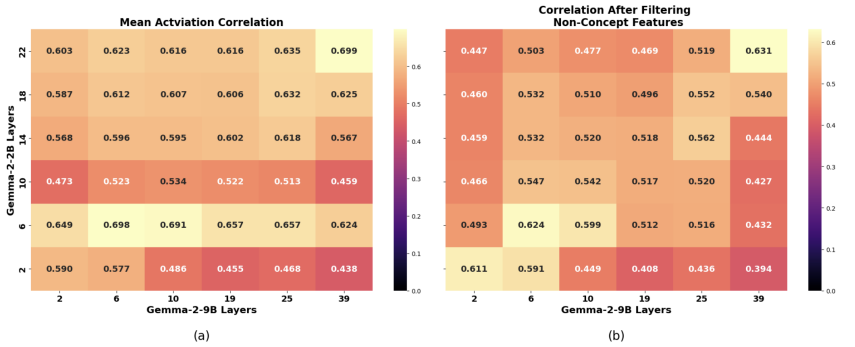

Figure 3: (a) 1-to-1 Mean Activation Correlation before and (b) after filtering non-concept features for Gemma-2-2B vs Gemma-2-9B. Note these patterns generally contrast from those of the SVCCA and RSA scores in Figure 1, indicating that these metrics each reveal different patterns not shown previously such as the pattern of middle layers between both models exhibiting higher correlation.

filtering peaks at 0.70 (L22 & L39) and averages 0.60 ± 0.07 on the mid-layer block (Fig. 3a). After removing unimportant features (Fig. 3b) the pattern of middle layers having an increased correlation between the two models becomes more evident, while more surrounding random matches fall by 0.10 – 0.15, raising the peak correlation from 0.70 to 0.74. Another result to note is that early layers and late layers of both models share strikingly higher similarity compared to other layers; for example L2 of both models, L6 of both models, and Gemma-2-2B L22 & Gemma-2-9B L39 all have the highest correlation (Fig. 3b, Fig. 5b, Fig. 7b). In other words, removing low-level features (such as punctuation) made strong alignments clearer and more meaningful, without just artificially boosting scores.

Semantic-subspace alignment. When we fix a single Gemma-2-2B layer and correlate every semantic-concept row against layers 2, 6, 10, 19, 25, and 39 for Gemma-2-9B, the same mid-on-mid pattern re-emerges: mid-stack layers in both models align best. Among the 2-2B sources we tried, the layer centred around L14 most consistently exhibited the highest SVCCA and RSA scores, reinforcing the idea that internal concept geometry more commonly converges in the middle of the networks. Full heat-maps for every source layer and metric are collected in Appendix B.

Overlapping Semantic Subspaces. When two semantic subspaces are combined, there are two different trends in the results based on the compatibility of the subspaces. For instance, combining the subspaces "country" and "people" in Table 2 which is a non-ideal pair results in low average SVCCA and RSA scores across the layers, indicating that there is insignificant correlation in how the models internally represent this subspace. This phenomenon could be caused by the fact that the models are unlikely to group these subspaces together during training.

Overlapping Concept	Paired SVCCA Mean	Random Shuffling Mean	p-value
Emotion and Time	0.62	0.13	0.0
Nature and People	0.63	0.17	0.0

Table 1: Comparison of paired SVCCA, random shuffling mean, and p-values for reasonable pairs of concepts at layers 10 of both Gemma-2-2B and Gemma-2-9B.

Overlapping Concept	Paired SVCCA Mean	Random Shuffling Mean	p-value
Country and People	0.03	0.13	0.02

Table 2: Comparison of paired SVCCA, random shuffling mean, and p-values for bad pairs of concepts at layers 10 of Gemma-2-2B and layer 19 of Gemma-2-9B. These results indicate that these pairs are not encoded similarly

However, when the pair of subspaces being combined makes sense, such as "nature" and "people" in Table 1, both the SVCCA and RSA scores are high, leading to the conclusion that both models represent these subspaces very similarly. All of the results are in Appendix C.

Multi-token Semantic Subspaces. Despite resource constraints limiting our evaluations to the "emotions time" subspace, our preliminary results on multi-token subspace (see Table 3) provide key insights. Notably, high SVCCA scores remained in the early and middle layers, providing strong empirical evidence that models sometimes encode multi-token concepts. Furthermore, the SVCCA scores are drastically higher than "emotions" or "time" alone in the early layers (see Appendix B), indicating that earlier layers represent multi-token subspaces rather than single-token ones. This result challenges the popular, underlying assumption that models internally encode single-token concepts (Dehouck, 2023; Valois et al., 2024). Hence, we believe that meaning is sometimes distributed across multiple tokens, and that semantic subspaces are the better level of analysis.

Distance metrics. During layer-to-layer SAE feature analysis, we have used the Pearson correlation as done in Lan et al. (2024). We tested other metrics such as the cosine similarity and euclidean distance; however, changing the similarity metric did not yield any statistically significant changes in the results, implying that the distance metric used does not affect the accuracy of our results.

5 Related Works

Superposition and Sparse Autoencoders. Previous studies have shown that, when there are more features to be represented than available parameters, feature representations are distributed across multiple parameters, leading to polysemantic neurons (Elhage et al., 2022). Polysemanticity causes challenges in interpreting models, which is crucial for AI safety in identifying goal misgeneralization (Shah et al., 2022; Langosco et al., 2022) as well as deceptive misalignment (Hubinger et al., 2024; Greenblatt et al., 2024). For these reasons, Sparse Autoencoders (SAEs) have been used to transform polysemantic neuron activations into monosemantic feature neurons that usually correspond to one feature (Makhzani and Frey, 2013; Cunningham et al., 2023; Gao et al., 2024; Rajamanoharan et al., 2024a,b). It is much easier to conduct quantitative interpretability studies on these monosemantic features.

Feature Universality. The existance of "universal" neurons across LLMs were first uncovered in a study of GPT-2 (Gurnee et al., 2024). Furthermore, previous studies that have performed quantitative analysis using SAEs to test for feature universality (Lan et al., 2024; Bricken et al., 2023) have shown universality in analogous features and representational features (Olah et al., 2020; Yosinski et al., 2014; Gurnee et al., 2024; Kornblith et al., 2019). These are not measures of "true features", which are stricter ground-truth features (Bricken et al., 2023) that represent atomic linear directions (Till, 2023).

Previous research on SAEs to test for

Multi-token Concept	Paired SVCCA Mean	Random Shuffling Mean	p-value
Emotion and Time (L6 vs L2)	0.27	0.02	0.0
Emotion and Time (L6 vs L10)	0.53	0.02	0.0

Table 3: Comparison of paired SVCCA, random shuffling mean, and p-values for bad pairs of concepts at layers 6 of Gemma-2-2B and layer 2 and 10 of Gemma-2-9B. These results indicate that multi-token inputs are encoded similarly.

feature universality (Lan et al., 2024) has demonstrated that, after aligning neurons via mean activation correlation, there exists statistically significant alignment ($p < 0.05$) for almost all non-input layers. The middle layers exhibited the strongest correspondence, indicating that distinct LLMs learn a shared set of features. Beyond layer-wise comparisons, semantically defined subspaces were tested by filtering features whose top activating tokens match curated keyword lists linked to a conceptual category such as "Emotions"or "Time". These subspaces yielded high SVCCA scores with $p \ll 0.05$, illustrating that semantic concept feature groups are more consistent across models.

Mechanistic Interpretability. Interpretting neurons and MLP analysis have become increasingly popular (Foote et al., 2023; Garde et al., 2023) techniques to understand the inner workings of LLMs. Other mechanistic interpretability methods have used SAEs (Lan et al., 2024; Nanda and Conmy, 2024) because of their more interpretable feature spaces.

6 Conclusion

In this paper, we provided a comprehensive study on feature universality in models of different sizes. Our experiments reveal that, in many cases, internal representations are similar across models regardless of size. In addition, our semantic subspace experiments revealed that different models encode pairs of subspaces and multi-token subspaces similarly, further enforcing the concept of LLM universality. These findings are a key step in forming a comprehensive understanding of how LLMs internally function.

7 Limitations

This study focused only on the Gemma-2-2B and Gemma-2-9B models due to resource constraints, particularly GPU availability and compute time. While this size comparison captures meaningful differences in model scale, further work could extend our approach to a wider range of architectures, parameter counts, and layer comparisons. Additionally, the number of SAE-derived subspaces we analyzed was limited to keep manual inspection and downstream evaluations manageable. Expanding this analysis to a larger, more diverse set of subspaces could help further characterize the extent and nature of feature space universality. Furthermore, our experiments with multi-token subspaces only tested a single subspace due to further resource limitations.

References

2010. Wordnet: Lexical database for english. https://wordnet.princeton.edu/. Accessed: 2025-05-18.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety – a review. *Preprint*, arXiv:2404.14082.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv:2304.01373.

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. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. A toy model of universality: Reverse engineering how networks learn group operations. *Preprint*, arXiv:2302.03025.

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

Mathieu Dehouck. 2023. Challenging the" one single vector per token" assumption. In *The SIGNLL Conference on Computational Natural Language Learning*, pages 498–507.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/toy_model/index.html.

Alex Foote, Neel Nanda, Esben Kran, Ioannis Konstas, Shay Cohen, and Fazl Barez. 2023. Neuron to graph: Interpreting language model neurons at scale. *Preprint*, arXiv:2305.19911.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *Preprint*, arXiv:2406.04093.

Albert Garde, Esben Kran, and Fazl Barez. 2023. Deepdecipher: Accessing and investigating neuron activation in large language models. *Preprint*, arXiv:2310.01870.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. 2024. Alignment faking in large language models. *arXiv preprint*. ArXiv:2412.14093 [cs].

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *Preprint*, arXiv:2306.12001.

Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, Adam Jermyn, Amanda Askell, Ansh Radhakrishnan, Cem Anil, David Duvenaud, Deep Ganguli, Fazl Barez, Jack Clark, Kamal Ndousse, Kshitij Sachan, Michael Sellitto, Mrinank Sharma, Nova DasSarma, Roger Grosse, Shauna Kravec, Yuntao Bai, Zachary Witten, Marina Favaro, Jan Brauner, Holden Karnofsky, Paul Christiano, Samuel R. Bowman, Logan Graham, Jared Kaplan, Sören Mindermann, Ryan Greenblatt, Buck Shlegeris, Nicholas Schiefer, and Ethan Perez. 2024. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *arXiv preprint*. ArXiv:2401.05566 [cs].

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. *Preprint*, arXiv:1905.00414.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2024. Sparse autoencoders reveal universal feature spaces across large language models. *arXiv preprint arXiv:2410.06981*.

Lauro Langosco Di Langosco, Jack Koch, Lee D. Sharkey, Jacob Pfau, and David Krueger. 2022. Goal Misgeneralization in Deep Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12004–12019. PMLR. ISSN: 2640-3498.

Alireza Makhzani and Brendan Frey. 2013. k-sparse autoencoders. *arXiv preprint arXiv:1312.5663*.

Neel Nanda and Arthur Conmy. 2024. Progress update #1 from the gdm mech interp team: Full update. Accessed: 2024-06-19.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Bruno A. Olshausen and David J. Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Narain Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Neural Information Processing Systems*.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. 2024a. Improving dictionary learning with gated sparse autoencoders. *Preprint*, arXiv:2404.16014.

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

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals. *arXiv preprint*. ArXiv:2210.01790 [cs].

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024a. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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 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. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Demian Till. 2023. Do sparse autoencoders find 'true features'? Accessed: 2025-01-29.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Pedro HV Valois, Lincon S Souza, Erica K Shimomoto, and Kazuhiro Fukui. 2024. Frame representation hypothesis: Multi-token llm interpretability and concept-guided text generation. *arXiv preprint arXiv:2412.07334*.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3320–3328, Cambridge, MA, USA. MIT Press.

A Additional Results

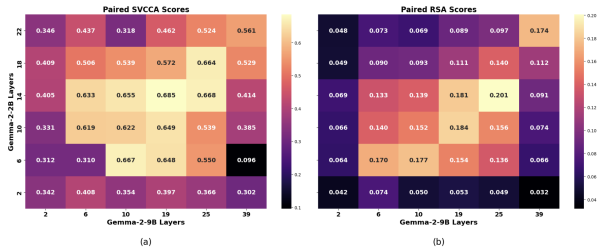

Figure 4: (a) SVCCA and (b) RSA Many-to-1 paired scores of SAEs for layers in Gemma-2-2B vs Gemma-2-9B. Note the pattern of higher scores in the middle layers indicating similarity in middle layers between both models.

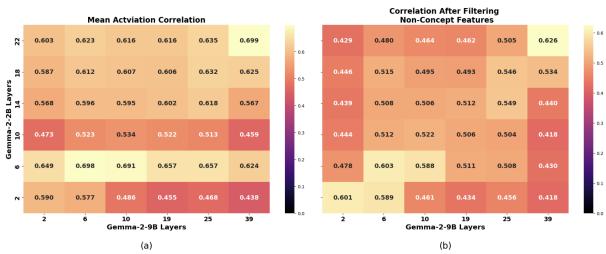

Figure 5: (a) Many-to-1 Mean Activation Correlation before and (b) after filtering non-concept features for Gemma-2-2B vs Gemma-2-9B. Note these patterns generally contrast from those of the SVCCA and RSA scores in Figure 1, indicating that these metrics each reveal different patterns not shown previously.

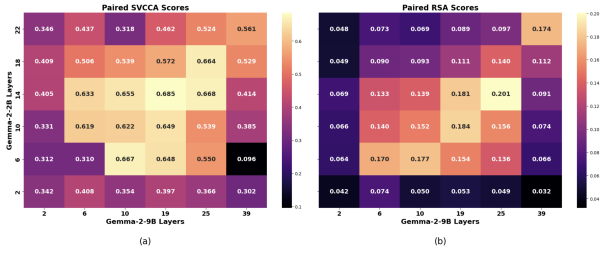

Figure 6: (a) SVCCA and (b) RSA Many-to-1 (5 run average) paired scores of SAEs for layers in Gemma-2-2B vs Gemma-2-9B. Note the pattern of higher scores in the middle layers indicating similarity in middle layers between both models. This pattern is a trend in layer similarity between the three experiments as seen in Fig. 2a, Fig. 2b, Fig. 10a, and Fig. 4b.

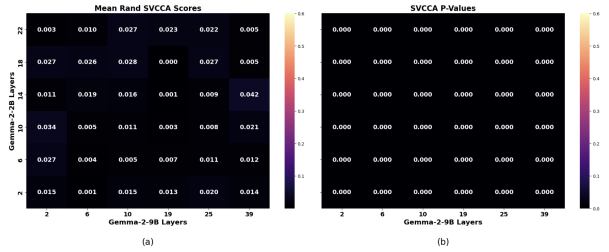

Figure 9: (a) Mean Randomly Paired SVCCA Many-to-1 scores and (b) SVCCA Many-to-1 P-values of SAEs for layers in Gemma-2-2B vs Gemma-2-9B.

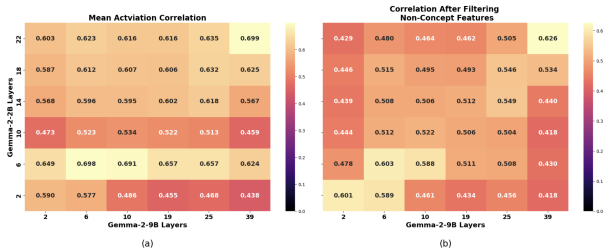

Figure 7: (a) Many-to-1 (5 run average) Mean Activation Correlation before and (b) after filtering non-concept features for Gemma-2-2B vs Gemma-2-9B. Note these patterns generally contrast from those of the SVCCA and RSA scores in Figures 2, 4, and 6, indicating that these metrics each reveal different patterns not shown previously.

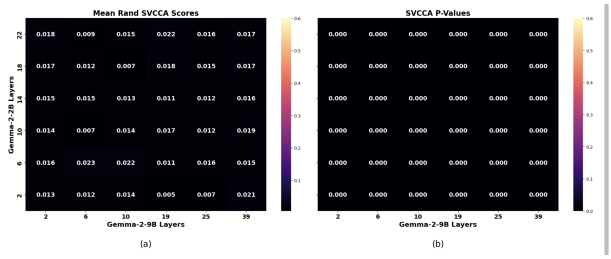

Figure 8: (a) Mean Randomly Paired SVCCA 1-to-1 scores and (b) SVCCA 1-to-1 P-values of SAEs for layers in Gemma-2-2B vs Gemma-2-9B.

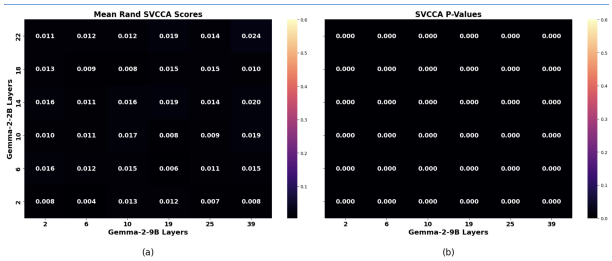

Figure 10: (a) Mean Randomly Paired SVCCA Many-to-1 scores and (b) SVCCA Many-to-1 (5 run average) P-values of SAEs for layers in Gemma-2-2B vs Gemma-2-9B.

B Semantic Subspace Similarity (Gemma-2-2B → Gemma-2-9B).

The following figures visualize concept-wise alignment from fixed layers of Gemma-2-2B to all layers of Gemma-2-9B. Each heatmap row corresponds to a semantic category (e.g., *Emotions*, *Biology*) and each column is a layer in Gemma-2-9B.

We show both SVCCA and RSA 1-to-1 results for 2-2B source layers L6, L10, L14, and L17. These help assess which layers in Gemma-2-9B best match the concept geometry of the smaller model. For most layers, similarity scores peak in the mid-stack (L10–L19), further supporting the cross-model alignment trend observed in Section 4.

** Some concepts may not appear across all rows if they lacked sufficient matched features or token coverage.*

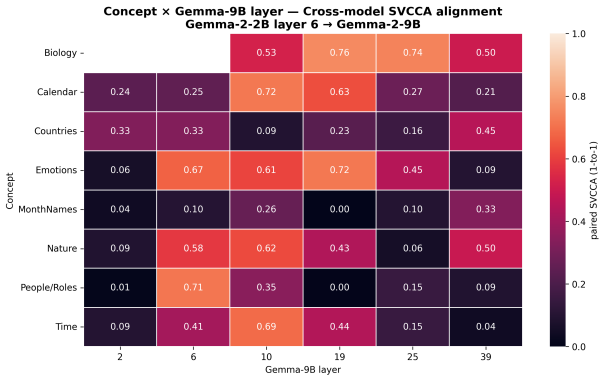

Figure 11: Scores rise into the mid-layers, peaking near 0.40 at L10–L19.

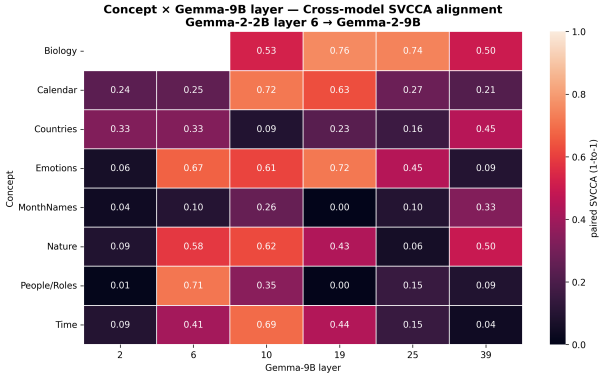

Figure 12: Scores rise into the mid-layers, peaking near 0.40 at L10–L19.

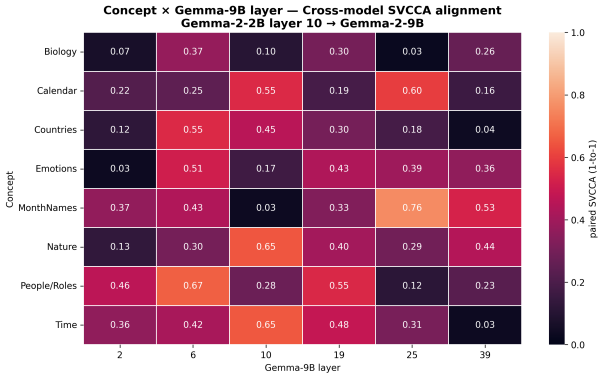

Figure 13: Mid-layer alignment strengthens; several concepts surpass 0.60.

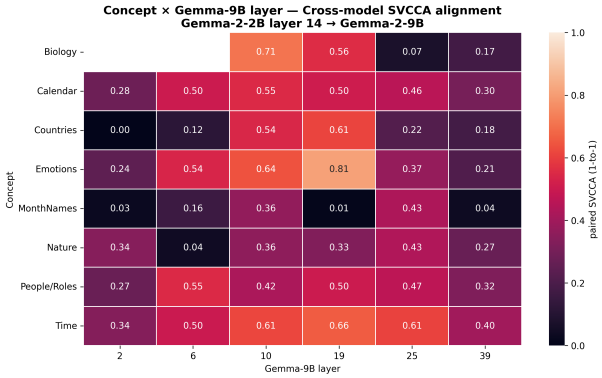

Figure 14: Highest similarity sits in the center stack; edge layers lag.

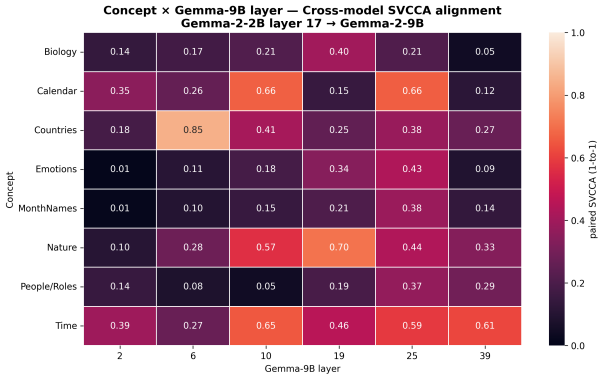

Figure 15: Alignment remains centered; deep and early layers score lower.

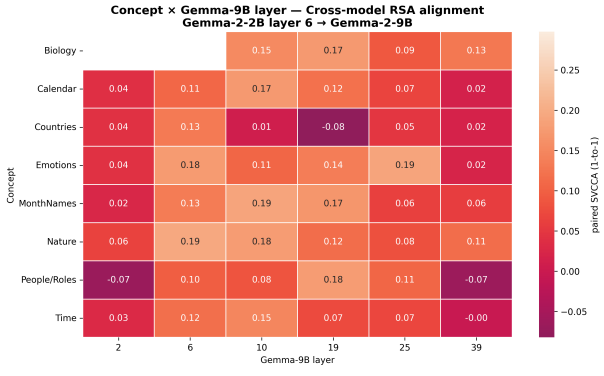

Figure 16: Mid-layer RSA peaks near 0.17; edge layers stay low. *People/Roles* dips below 0 at L2.

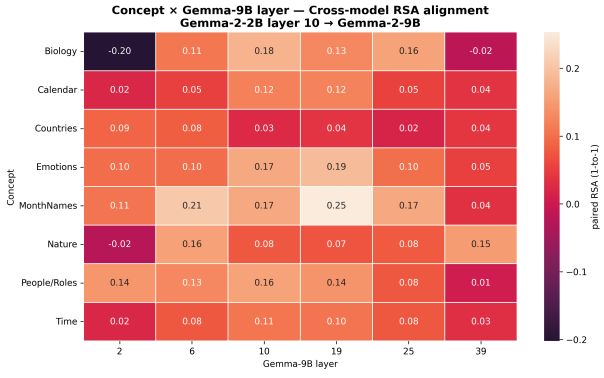

Figure 17: Mid-to-deep pairs score higher: *Biology*, *Month-Names* reach ∼0.25 at L19.

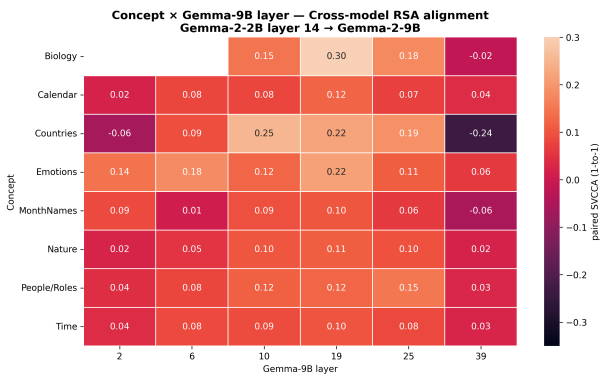

Figure 18: Highest scores at 9B L19: *Biology* ∼0.30, *Countries* ∼0.25. L2 and L39 remain near zero.

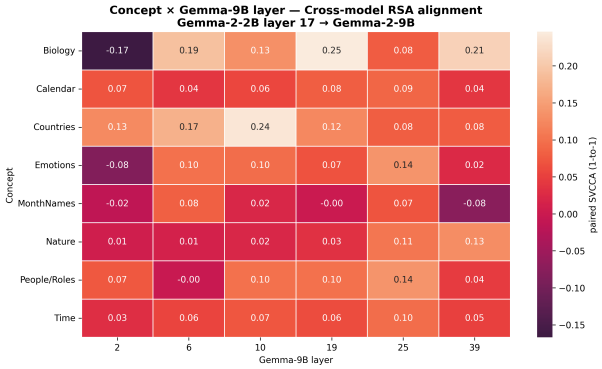

Figure 19: Strongest alignment at L19; outer layers weak or negative.

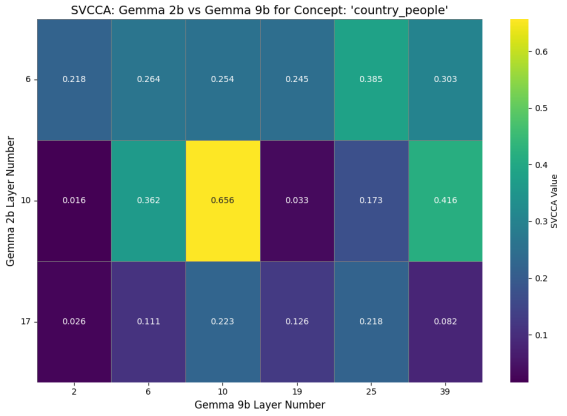

Figure 22: paired SVCCA 1-to-1 for country-people concept

C SVCCA by concept in Gemma-2-2B and Gemma-2-9B.

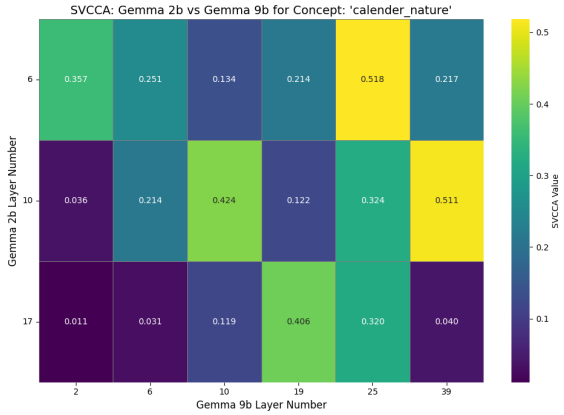

Figure 20: paired SVCCA 1-to-1 for calender-nature concept

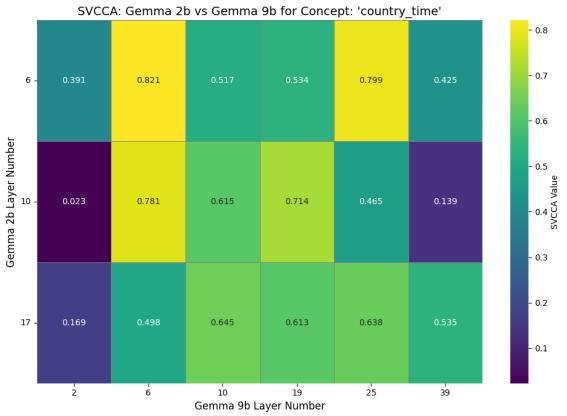

Figure 23: paired SVCCA 1-to-1 for country-time concept

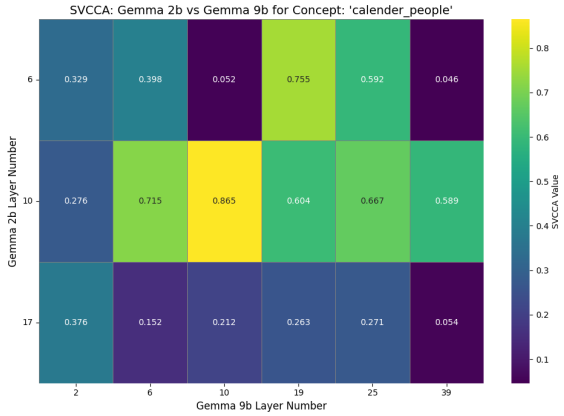

Figure 21: paired SVCCA 1-to-1 for calender-people concept

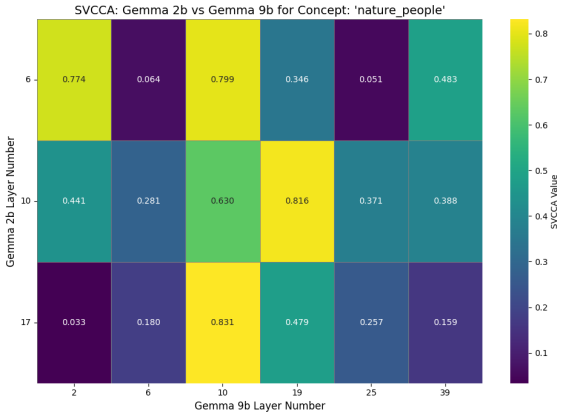

Figure 24: paired SVCCA 1-to-1 for nature-people concept

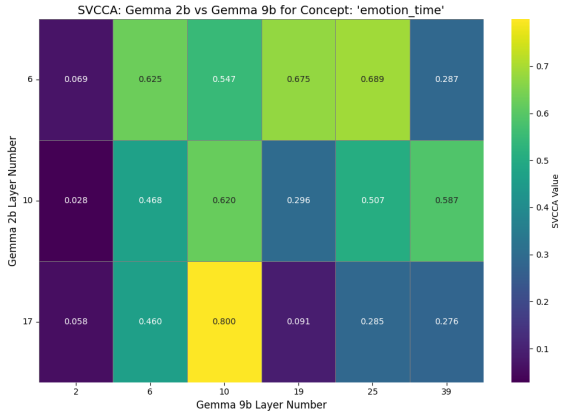

Figure 25: paired SVCCA 1-to-1 for emotion-time concept