TO BE OR NOT TO BE: MUTUALLY EXCLUSIVE FEATURES IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders have become a widespread tool for discovering interpretable features in language models and to intervene upon them. Yet the standard practice still treats features in isolation, leaving the structure of their relationships unexplored. We study these relationships through feature co-activation graphs, focusing on a particular suppressive relationship that we call **mutual exclusivity**: features that occur in similar contexts *but not together*. Across a wide range of models varying in size, we consistently find that mutually exclusive features form interpretable groups whose members represent alternative values of a common semantic type, such as numerals, pronouns, and months, to name a few. These groups exhibit consistent geometric structure: their decoder directions are highly similar, placing alternative values within narrow regions of activation space. Finally, we show that these features are causally relevant to model behavior: replacing a feature with another ME alternative predictably moves the model's output toward the concept represented by the replacement feature. Together, our findings demonstrate that relationships among features reveal structure in model representations and mechanisms that are difficult to discover by studying them in isolation.

1 INTRODUCTION

Sparse autoencoders (SAEs) have become a widely used tool for identifying interpretable features in language models (Bricken et al., 2023; Cunningham et al., 2024; Templeton et al., 2024; Gao et al., 2025). They offer a way to break complex model activations into more understandable components, with the hope that these components can reveal what information models represent, providing targets for analyzing or controlling their behavior (Marks et al., 2025; Arad et al., 2025; Yeon et al., 2025). However, these features are typically analyzed in isolation, one at a time. Model computation depends not only on which features activate, but also on how they relate and interact. Despite this, little is known about the structure formed by SAE features collectively.

A burgeoning and exciting line of work has begun to study interactions between features, including relationships between features that frequently co-activate (Li et al., 2025; Prieto et al., 2026). While prior work has focused primarily on positive co-activation (Bhalla et al., 2026; Prieto et al., 2026; Li et al., 2025; Park et al., 2025; Fel et al., 2026a), suppressive relationships between features have received much less attention. In this work, we study a particular suppressive relationship, which we call *mutual exclusivity*. Two features are mutually exclusive (ME) when they are negatively coupled—activation of one is associated with reduced activation of the other—while sharing similar co-activation neighborhoods. Intuitively, these features occur in similar contexts, but not together (see Figure 1B). A natural way to study these relationships more broadly is through a co-activation graph Bhalla et al. (2026): a graph in which nodes correspond to SAE features and weighted edges capture statistical relationships between their activation patterns (Figure 1A). Such a graph describes not only which features tend to co-occur, but also which features tend not to.

Across four language models spanning three different SAE architectures, we construct co-activation graphs and identify interpretable ME sets (pairs and cliques), e.g. digits 0-9, pronouns, months. We find that **mutually exclusive sets are interpretable as collectives**, and larger cliques are more interpretable (see Figure 2). Cliques of mutually exclusive features often correspond to alternative values of a common high-level type, suggesting that **they capture decision points in the model's generation process**. These features also exhibit a consistent geometric structure: decoder directions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

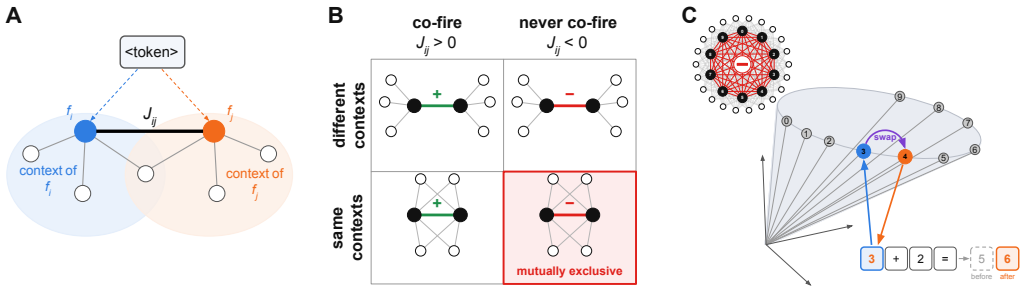

Figure 1: **Mutually exclusive features in the co-activation graph of SAE features. (A)** We fit an Ising model to SAE feature activations, yielding a *co-activation graph*. Nodes are features f_i; edges are couplings J_{ij} measuring how often two features co-fire. The shaded regions mark the neighbors of features f_i and f_j, i.e. their context. **(B)** Two graph properties characterize each feature pair: the sign of the direct coupling (columns) and the overlap of the pair's neighborhoods (rows). Pairs that share contexts but have strongly negative coupling (red) are *mutually exclusive*. **(C)** The digit features 0–9 form a mutually exclusive *clique*, arranged on a cone in activation space. Swapping "3" for "4" in "$3 + 2 =$" changes the output from "5" to "6".

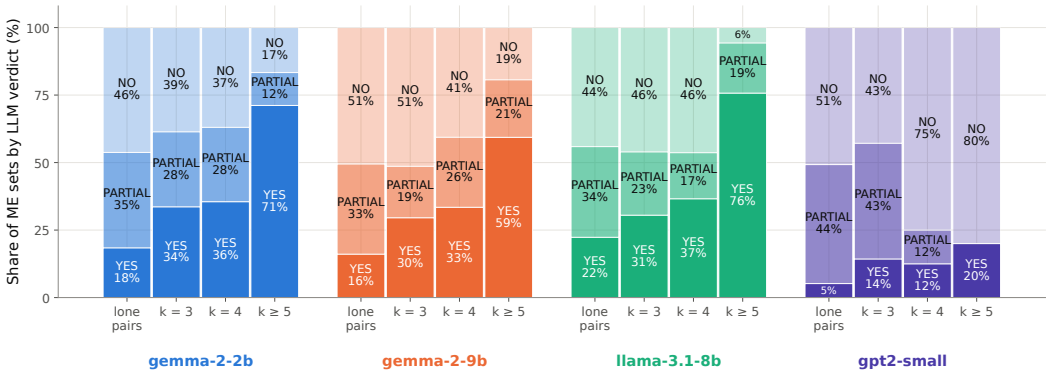

Figure 2: **LLM judge verdict across clique sizes.** We ask an LLM judge to provide a description of a set of ME features and a verdict as to that description's ability to describe all the features in the set. The verdict is a 3-value outcome among {`YES`, `PARTIAL`, `NO`}. We report the distribution of verdicts given by the judge for four categories of ME features: lone pairs ($k = 2$), cliques with $k = 3$, $k = 4$, and $k \geq 5$. For each k we color differently the fraction of times each verdict is given. We observe that larger cliques are easier to describe as a unit. The number of cliques for each k are reported in Table 3

within ME cliques have high cosine similarity, forming narrow cones in activation space. While superposition Elhage et al. (2022) states that two features which never co-activate might not be orthogonal, we find a stronger pattern: when such features also occur in similar contexts, they are close together.

Finally, we test whether this structure has a causal role in model behavior. When one active feature is replaced by another member of its ME clique, the model's output shifts toward the concept represented by the replacement feature (see Figure 1C for an illustrative example). Compared with features identified using a standard activation-based selection method, substitutions between ME features produce stronger targeted steering and are able to do so in later layers.

Together, these results establish mutual exclusivity as a useful criterion for discovering structured, interpretable groups of SAE features. SAE features are not merely a collection of independently meaningful directions: relationships among them can reveal higher-level organization, including sets of alternative values and the choices between them that shape model computation.

2 Related Work

Language models represent more concepts than they have orthogonal directions by storing features in superposition (Elhage et al., 2022). Under the linear representation hypothesis, each concept corresponds to a direction in activation space Park et al. (2023). SAEs recover such directions by learning a sparse, overcomplete dictionary (Cunningham et al., 2024; Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2025; Rajamanoharan et al., 2024; Bussmann et al., 2024), and they have become the standard tool for decomposing model activations into features (Sharkey et al., 2025).

Despite the popularity of SAEs, a growing body of work finds that many high-level concepts are not single directions. Concepts can be stored as convex regions (Park et al., 2025; Fel et al., 2026b; Tětková et al., 2025) or on low-dimensional manifolds (Modell et al., 2026; Kantamneni & Tegmark, 2025; Gurnee et al., 2025; Karkada et al., 2026; Feucht et al., 2026; Sarfati et al., 2026), such as the circles traced out by days of the week (Engels et al., 2025). This has motivated work on how groups of SAE directions jointly capture such concepts (Bhalla et al., 2026), and new featurizers that move beyond single directions (Francel, 2026; Dalili & Mahdavi, 2026; Shafran et al., 2026; Huang & Hahn, 2026; Fel et al., 2026a).

Co-activation has been used to recover groups of related features. Prieto et al. (2026) show that correlated features are arranged according to their co-activation patterns, which makes the interference between their directions constructive rather than noise to be filtered out. At a coarser scale, SAE features that co-occur have been found to cluster spatially into functional "lobes" (Li et al., 2025). Bhalla et al. (2026) show that SAEs can capture a concept manifold by grouping two ways: globally, with a compact group of features whose span contains the manifold, or locally, with features that each tile a small region of it. Block-sparse featurizers build this grouping into the architecture by forcing groups of directions to activate together (Fel et al., 2026a).

All of this prior work concerns *positive* associations: features that fire together are grouped together. Negative relationships, as in mutual exclusivity of features, have mostly been treated as training artifacts of SAE design, usually in the context of feature splitting and absorption Chanin et al. (2024); Bricken et al. (2023).

3 Method

3.1 Ising model

Consider an LLM with an SAE of width N. Each token passed through the LLM activates only a small subset of the SAE features, yielding a sparse vector $\mathbf{z} \in \mathbb{R}^N_{\geq 0}$. We discretize $\mathbf{z}$ to get $\mathbf{s} \in \{-1, 1\}^N$ by mapping $\mathbf{z}_i$ to 1 if $\mathbf{z}_i > 0$ and -1 otherwise.

The Ising model (Ising, 1925) describes how features relate while making minimal assumptions on the structure of their interactions (Jaynes, 1957). It assigns each discretized vector a score

$$p(\mathbf{s}) \propto \exp\Big(\sum_{i=1}^{N} h_i \mathbf{s}_i + \sum_{i<j} J_{ij} \mathbf{s}_i \mathbf{s}_j\Big), \tag{1}$$

h_i is a per-feature term reflecting how often feature i fires on its own, and J_{ij} is a pairwise coupling added whenever features i and j are conditionally dependent Besag (1974). J is symmetric and zero along the diagonal, providing us with a weighted graph, which we call the **co-activation graph**. J_i are feature i's connections to its neighbors. The subset of positively coupled neighbors to feature i will be denoted J_i^+. Details for how we fit this model are in Appendix B.

3.2 Selecting mutually exclusive pairs

Using the co-activation graph J introduced in Section 3.1, we measure two properties of each pair of features i and j: their direct coupling J_{ij}, and their **neighborhood similarity** $\cos(J_i^+, J_j^+)$, i.e. the cosine similarity of their positive couplings.

Together, these two properties identify **mutually exclusive (ME)** pairs: features that occur in the same contexts (high neighborhood similarity) yet rarely or never fire together (strongly negative J_{ij}).

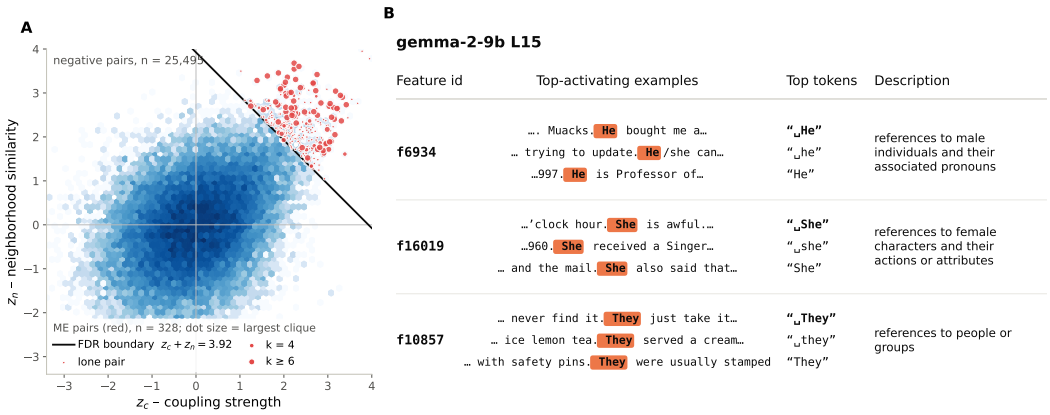

Figure 3: **Selecting and labeling ME features. (A)** Each point is a pair of features, placed by the rank-based z-scores of its negative coupling (z_c) and its neighborhood similarity (z_n). The blue 2D histogram shows the density of all negatively coupled pairs. Pairs above the solid black line $z_c + z_n = \tau^*$ are selected as ME. The threshold τ^* is chosen so that the false discovery rate, estimated against coupling-permuted null graphs, is at most 5%. Selected ME pairs are shown as red dots, sized according to the largest clique they belong to. **(B)** Example of Neuronpedia data used for each ME feature. For each of the feature's top-activating examples, we mark the token on which it fires most strongly.

Intuitively, such pairs behave like alternative values of the same type, e.g., two features encoding `East` and `West`, which appear with similar surroundings but not at the same time.

Both properties are continuous metrics, so mutual exclusivity is a matter of degree. To combine them, we convert negative coupling strength and neighborhood similarity to rank-based z-scores, z_c and z_n, separately for each (model, layer), and define **ME degree** as their sum. We then select pairs with ME degree greater than τ^* (Figure 3). The threshold τ^* is set against a null graph obtained by randomly permuting the couplings of J, so that at most 5% of the selected pairs are expected to reach their score by chance. The minimum coupling strength for pairs of features that pass this threshold is referred to as $j^* = \min_{(i,j)\in ME} |J_{ij}|$ Details are in Appendix A.

ME pairs do not always exist in isolation. There can exist collections of features, all of which are mutually exclusive. When at least three features are all pairwise ME, we call the collection an **ME clique**. If cliques overlap, we study them separately.

3.3 Feature Labeling

To interpret the features we identify, we use Neuronpedia (Lin, 2023), which provides top-activating examples, LLM-generated descriptions, and top output tokens for each feature (Figure 3).

For each feature, we compute *token purity*: the fraction of top-activating examples whose maximally activating token equals the feature's most common activating token. A purity of 1 means the feature fires maximally on exactly one token. High purity suggests a *lexical* feature that responds to a small set of tokens, while low purity suggests a *contextual* feature that fires across many different tokens. We also measure semantic similarity between pairs using the Jaccard similarity of the words in their Neuronpedia descriptions.[1]

Finally, for each maximal ME clique (including lone pairs), we ask an LLM judge to generate a description of the clique and assign a verdict of `Yes`, `Partial`, or `No`, indicating if the description provided describes all the features in the clique as alternative values of a shared type. The judge prompt is shown in Appendix F.

[1]Neuronpedia's top-activating examples capture only the extreme tail of a feature's activation distribution, so token purity may not capture the typical behavior of a feature.

3.4 STEERING

To test whether ME features play a causal role in model computation, we intervene by substituting one feature for another one in the same clique. For an ME pair with source and target features f_s, f_t and corresponding top-activating tokens s, t, we construct matched prompts x_s, x_t that differ only at position i:

$$x_s[i] = s, \qquad x_t[i] = t. \tag{2}$$

The prompts are chosen such that the model's subsequent prediction depends on this value. Importantly, we intervene at the token position itself, modifying its upstream representation and the implicated computation rather than directly steering the later output.

For example, consider an ME pair corresponding to the tokens `3` and `4`. We use the matched prompts

$$x_s = \text{"}\texttt{3 + 2 =}\text{"}, \qquad x_t = \text{"}\texttt{4 + 2 =}\text{"}. \tag{3}$$

We intervene on the representation of `3` in x_s, substituting its feature with the feature for `4`. The desired downstream effect is a shift in the model's next-token prediction from `5` toward `6`.

Let $z_s[i]$ be the activation of f_s at the intervention position, and W_s, W_t their SAE decoder directions. At layer L, we replace the source feature with the target feature at equal activation strength:

$$h_L[i] \leftarrow h_L[i] + z_s[i](W_t - W_s). \tag{4}$$

which removes the contribution of the source feature and adds the target feature with the same activation magnitude.

To test whether the effect is specific to ME partners, we repeat each swap with the target f_t replaced by one of four other partners f_p: the feature most strongly positively coupled to f_s; the feature most strongly negatively coupled to f_s that is not in an ME pair; the feature with the most similar neighborhood to f_s that is not in an ME pair; and a random feature. We also repeat each swap with *vanilla* features, chosen for each token by a standard activation-based procedure (Templeton et al., 2024; Wu et al., 2025) that picks the highest-activating feature for the token of interest. Both baselines are described in detail in Appendix I.

We measure each intervention by how much of the gap between the source and target next-token probability distributions it closes. Let $p_{\text{swap}}(\cdot \mid x_s)$ be the distribution after the intervention and d the ℓ_1 distance between distributions. The *gap closure* is

$$GC = 1 - \frac{d\big(p_{\text{swap}}(\cdot \mid x_s),\, p(\cdot \mid x_t)\big)}{d\big(p(\cdot \mid x_s),\, p(\cdot \mid x_t)\big)}. \tag{5}$$

Replacing the input token itself (evaluating x_t) gives $GC = 1$, and an intervention with no effect gives $GC = 0$. Negative values mean the intervention moved the distribution away from the target.

3.5 EXPERIMENTAL SETUP

We study four LLMs: GPT-2 Small (Radford et al., 2019), Gemma-2-2b and Gemma-2-9b (Team et al., 2024), and Llama-3.1-8b (Grattafiori et al., 2024), using publicly available residual-stream SAEs at every layer: Gemma Scope (Lieberum et al., 2024), Llama Scope (He et al., 2024), and the GPT-2 Small SAEs of Bloom et al. (2024). These span three SAE architectures and training protocols (JumpReLU, TopK, and ReLU with ℓ_1). We fit co-activation graphs on SAE activations over 2M tokens from the Pile (Gao et al., 2020), using a fixed regularization strength $\lambda = 4 \cdot 10^{-4}$ for all models; other values of λ give qualitatively similar graphs. Cliques are described and evaluated with GPT-4o-mini (OpenAI, 2024). Full details are in Appendix B.

For each layer, we select ME pairs and characterize them (Section 4.1). We (i) interpret cliques and find identities that recur across layers and models (Section 4.2), (ii) characterize their geometry in activation space (Section 4.3), and (iii) test the causal role of ME pairs by substituting one feature for the other (Section 4.4).

4 RESULTS

Across models, we find that mutual exclusivity reveals a consistent organization among SAE features. The strongest ME relationships form semantically coherent cliques, whose members represent

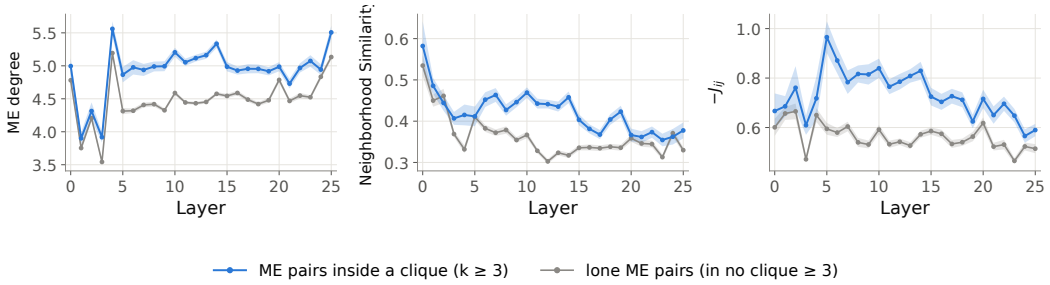

Figure 4: **Lone ME pairs versus ME pairs in cliques.** We split ME pairs into lone pairs (gray) and pairs that belong to a clique (blue). For Gemma-2-2b, we report the mean ME degree $z_c + z_n$ (left), the mean neighborhood similarity $\cos(J_i^+, J_j^+)$ (center), and the mean coupling strength $-J_{ij}$ (right) as a function of layer depth. ME pairs in cliques manifest stronger ME degree than lone pairs.

alternative values of a shared type. These cliques exhibit distinctive geometric structure, and swapping between their members causally changes downstream model predictions. We characterize each of these properties in turn.

4.1 Characterization of ME pairs and cliques

We select ME features using the procedure introduced in Section 3. Across the four models, we identify a total of 42,201 ME pairs, with the largest clique containing 10 features. GPT-2 Small yields substantially fewer ME pairs and smaller maximal cliques than the other models. We discuss these differences in Appendix J and focus primarily on the remaining three models throughout the results. The distribution of discovered pairs across layers is summarized in Appendix D.

The co-activation graphs are strongly modular. Across all models, the co-activation graphs exhibit substantial community structure. We partition each graph using the Louvain algorithm (Blondel et al., 2008) on its positive edges and find modularity roughly four to six times higher than in the baselines. High modularity means that positive relationships are concentrated within groups of features rather than distributed uniformly throughout the graph. This indicates that SAE features are organized into structured communities, motivating a closer examination of the relationships among individual features within this broader organization. Other graph statistics are in Appendix C

Pairs in cliques have higher ME degree than lone pairs. We present various statistics for ME pairs and cliques in Figure 4 for model Gemma-2-2b, and for the rest of the models in Appendix D.ME pairs that belong to a clique are less common than lone pairs, but they are more strongly mutually exclusive: their features share more similar neighborhoods and have stronger negative couplings. The stronger degree of mutual exclusivity in cliques suggests that a clique encodes a type: the shared context determines which type is relevant (e.g., digits), and mutual suppression ensures that only one of its possible values, a feature, is active at a time (e.g. 3).

4.2 Semantic interpretation of ME cliques

We provide a set of representative cliques and their features' top activating tokens in Table 1. Additional features are in Appendix G.

ME feature interpretability. We use Neuronpedia (Lin, 2023) and an LLM judge to interpret ME features. We compute token purity (how specific a feature is to a single token, as defined in Section 3.3) and description similarity of features and summarize the findings in Figure 5. Features in ME pairs are more token-pure than random features, and features in ME cliques are more token-pure than those in lone pairs (Figure 5, left). ME pairs also have more similar descriptions (Figure 5, right). Importantly, this semantic similarity is specific to the mutually exclusive relationship rather than merely to features being statistically related: strongly positively coupled pairs ($J_{ij} > j^*$) have substantially lower description similarity, despite frequently co-activating. ME features therefore

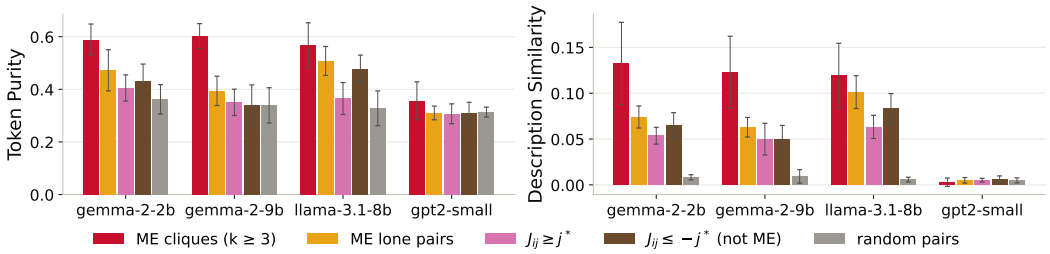

Figure 5: **Token Purity and Description Similarity between pairs of features. Left:** token purity, i.e. how specific a feature is to a single token. Features in ME cliques are more token-pure than other features across models. **Right:** Jaccard similarity between the word sets of the two features' descriptions, a coarse lower bound on semantic similarity. Pairs in ME cliques are described far more similarly than other pairs, except in GPT-2 Small. Together, these results suggest that an ME clique encodes a single attribute, with each feature specific to one of its values.

Table 1: **Representative ME cliques.** Tokens are each member's top activating token on Neuronpedia; descriptions are the clique-level LLM descriptions.

Model/layer	Top activating tokens	Clique description
Gemma-2-9B/21	`3 2 0 9 1 8 4 6 5 7`	digits
Gemma-2-9B/15	`I it She He We They`	pronouns
Llama-3.1-8B/6	`West East north`	compass directions
Gemma-2-2B/3	`June April October March`	months
Gemma-2-9B/15	`football basketball baseball United`	sports

tend to occupy similar semantic roles, whereas positively coupled features tend to represent concepts that occur together. Taken together, these results support an interpretation of an ME clique as representing a single type: high description similarity reflects the shared type, while high token purity reflects the specialization of each feature to one of its possible values. Interestingly, this mirrors a distinction between syntagmatic (compositional) and paradigmatic (mutually exclusive) relations in distributional semantics (Sahlgren, 2006): features that co-occur are related but play different roles (e.g., *New* and *York*), whereas features that substitute for one another in the same context are semantically similar (e.g., *Monday* and *Tuesday*).

Descriptions of ME cliques. We use an LLM judge to generate a description for each clique and to assess whether it describes all of the clique's features, as explained in Section 3.3. The larger a clique is, the more likely it is to receive a `YES` verdict (Figure 2). This trend is notable because larger cliques impose a stricter test: a single description must account for a greater number of distinct features. Describable cliques also have higher ME degree (Figure 15), suggesting that semantic coherence and mutual exclusivity go hand in hand: the more strongly features exclude one another, the more clearly they form a single interpretable type.

The same cliques recur across depth and across models. Given the clique-level descriptions, we can investigate if these cliques repeat across layers and if they are shared between models. Digit $(0-9)$ cliques are overwhelmingly common, having at least four members in $22/26$ layers in Gemma-2-2b, $40/42$ layer in Gemma-2-9b, and in $18/32$ layers of Llama-3.1-8b. Additionally, we find pronouns (he/she/they) and possessive pronouns (my/our/your/their) across many layers. Prepositions are also repeated, but those cliques are only shared between Gemma-2 models. See Figure 16 for a layer-wise breakdown of these examples.

4.3 Geometric Interpretation of ME Cliques

ME cliques lie in a cone. Figure 6 compares the cosine similarity between decoder directions for different classes of feature pairs. ME pairs belonging to cliques have the highest decoder similarity, even more aligned than strongly positively coupled pairs. Because high cosine similarity corresponds to small angles between directions, the features within an ME clique occupy a narrow

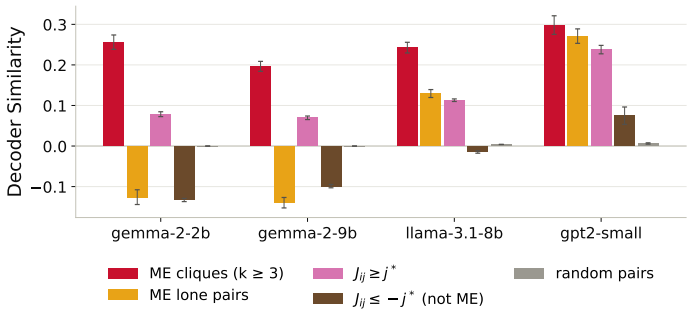

Figure 6: **Cosine similarity between decoder directions of feature pairs.** For each model, we compare ME pairs in cliques, lone ME pairs, strongly positively coupled pairs ($J_{ij} > j^*$), strongly negatively coupled pairs that are not ME ($J_{ij} < -j^*$), and random pairs. ME features in cliques have the most similar decoder directions, consistent with their lying in a shared cone.

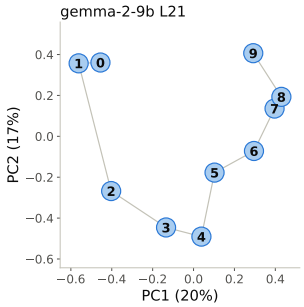

Figure 7: **2D PCA of the decoder directions of a digit clique** (Gemma-2-9b, layer 21). PC1 orders the digits; PC2 reveals an arc.

cone in activation space. By contrast, negatively coupled pairs that are not ME are anti-aligned or nearly orthogonal in all models, showing that this geometry is not a generic consequence of negative coupling.

Semantic order is encoded geometrically. An ordered clique is a clique whose values have a natural ordering, such as digits, ordinals, or months. We ask whether this semantic ordering is reflected in the geometry of the decoder directions. For each ordered clique, we compute the Spearman rank correlation between the values' semantic order and their coordinates along the first principal component. Across ordered cliques, PC1 strongly recovers the semantic ordering (Table 5), indicating that distinguishing information among clique members is organized within a low-dimensional subspace. Digit cliques provide a clear example: when projected onto their first two principal components, the decoder directions trace an arc whose positions follow the numerical ordering of the digits (Figure 7). See Appendix H for more examples.

4.4 ME FEATURES ARE CAUSAL

We next test whether ME features play a causal role in model computation. We intervene by replacing an active feature with another member of its ME clique and measure whether the model's downstream prediction shifts toward the value represented by the replacement feature. Digit cliques occur in most layers of the three larger models, so we test across depth with four prompts whose answer depends on a digit a: "$a + 1 =$", "$a + 2 =$", "$9 - a =$", and "The number is a. Once more, the number is". Figure 8 shows the results for digit cliques across layers. Additional steering results are in Appendix I. Across the three larger models, ME-based substitutions produce strong causal effects over a substantial range of layers, often closing nearly the entire gap to the target distribution. We compare these interventions with several co-activation-based baselines as well as vanilla activation-based feature selection (Appendix I). Representative interventions are shown in Figure 9.

5 CONCLUSION AND DISCUSSION

Our results suggest that feature co-activation graphs contain meaningful structure beyond the properties of individual SAE features. We focus on one underexplored relationship, mutual exclusivity: groups of features that occur in similar contexts but suppress one another. We show that these groups are semantically coherent, geometrically structured, and causally involved in model computation.

We interpret ME clique as a *type* that can take different *values*, represented by the clique's members. This view is supported by their semantic and geometric structure: features within cliques have more similar descriptions and higher token purity, larger cliques are more readily described by a single concept, and recurring examples include digits, pronouns, months, and geographic directions. ME

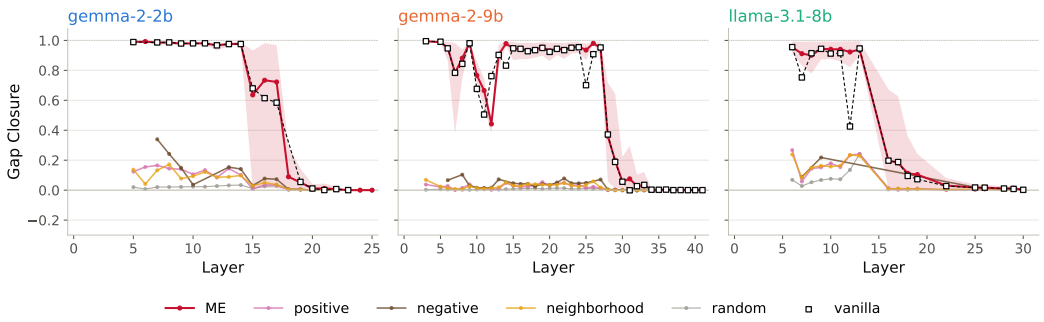

Figure 8: Replacing an active digit feature with an ME alternative produces high gap closure across early and middle layers in all three models, often approaching the effect of changing the input token itself. The effect drops sharply in later layers, after information has moved beyond the digit-token representation. ME and vanilla select mostly the same features in earlier layers, leading to ties between the methods.

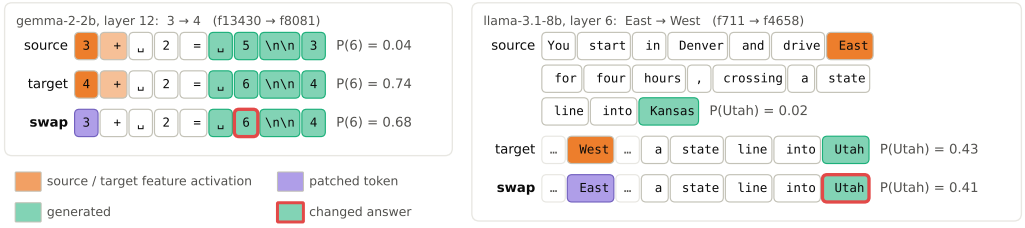

Figure 9: **Hand-picked examples of intervention experiments.** We show a pair of features for digits and compass directions. The top line for each pair is the source prompt, where the source feature is active on tokens highlighted orange. The middle shows the target prompt, which is what the intervention aims to steer towards, with the target feature activations indicated by orange. Finally, the third row shows in purple the single token which is intervened on, swapping the source feature for the target feature. Green tokens indicate tokens generated by the model after being fed the prompt. In both examples there is a key token to search for to identify if the intervention achieved its goal (red outline), and the logit probability for this token is shown on the right of each generation.

features within cliques have highly similar decoder directions, placing alternative values within a narrow cone. At the same time, the residual differences between these directions are meaningful: for ordered concepts such as digits and months, the first principal component strongly tracks the semantic ordering of the values. Thus, a clique appears to capture not only that a set of features belongs to the same type, but also relationships among the possible values of that type.

Our interventions further suggest that these structures correspond to consequential choices in model computation. Replacing an active feature with another member of its ME clique can move downstream predictions toward those produced when the corresponding value is actually present in the input, including for digits, directions, pronouns, and sports.

More broadly, these findings motivate treating the relationships among features as an object in their own right. SAE features need not be understood only as a dictionary of individually meaningful directions; the graph of how those directions interact may reveal higher-level variables, their possible values, and when choices among them become causally relevant. Mutual exclusivity is only one such relationship. Systematically studying the broader "grammar" of feature interactions may therefore provide a path from interpreting isolated representations toward understanding how they are organized into computation.

AI USE STATEMENT

In this work, we used generative AI tools for designing and providing feedback on experiments and to implement methods. Additionally, we used generative AI tools to support data analysis and creating figures. AI was used extensively to create and edit software code. AI assisted with drafting parts of the research paper, but no AI generated text was directly used in the main body of the paper. We also used generative AI to supplement relevant literature.

ETHICS STATEMENT

This work studies the structure of representations in language models using publicly available models, sparse autoencoders, and text data. Understanding these representations may support model auditing and more precise control of model behavior. However, the same intervention methods could also be used to alter model outputs for harmful purposes. Our experiments investigate controlled substitutions between represented concepts; they do not establish the safety or reliability of these interventions in deployment. Feature interpretations derived from corpus examples and automated judgments may also reflect biases in the underlying data and models.

REPRODUCIBILITY STATEMENT

We describe the construction of feature co-activation graphs, the selection of mutually exclusive pairs and cliques, and the feature-substitution intervention in Section 3. Appendix B documents the models, SAE architectures, data source, feature-filtering criteria, and graph-fitting procedure. Appendix A specifies the scoring and threshold selection procedures, Appendix F provides the LLM interpretation prompt, and Appendix I describes the steering baselines. The experiments use publicly available models and SAE checkpoints.

REFERENCES

Dana Arad, Aaron Mueller, and Yonatan Belinkov. SAEs are good for steering – if you select the right features. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10241–10259, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.519. URL https://aclanthology.org/2025.emnlp-main.519/.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds? In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=euoAoOflaN.

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008 (10):P10008, 2008.

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. Saelens. https://github.com/decoderesearch/SAELens, 2024.

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. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Isaac Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *Interpretable AI: Past, Present and Future*, 2024. URL https://openreview.net/forum?id=Wzav8fesTL.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Seyed Arshan Dalili and Mehrdad Mahdavi. Subspace-aware sparse autoencoders for effective mechanistic interpretability. *arXiv preprint arXiv:2606.06333*, 2026. URL https://arxiv.org/abs/2606.06333.

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. Toy models of superposition. *Transformer Circuits Thread*, 2022.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2405.14860.

Thomas Fel, Matthew Kowal, Mozes Jacobs, Dron Hazra, Usha Bhalla, Lee Sharkey, Lucius Bushnaq, Satchel Grant, Tal Haklay, Thomas Icard, Can Rager, Michael Pearce, Daniel Wurgaft, Aiden Swann, Fenil Doshi, Siddharth Boppana, Curt Tigges, Nick Cammarata, Thomas Serrec, Vasudev Shyam, Owen Lewis, Thomas McGrath, Jack Merullo, Ekdeep Singh Lubana, and Atticus Geiger. Structuring sparsity: Block-sparse featurizers capture visual concept manifolds. *arXiv preprint arXiv:2606.25234*, 2026a. URL https://arxiv.org/abs/2606.25234.

Thomas Fel, Binxu Wang, Michael Lepori, Matthew Kowal, Andrew Lee, Randall Balestriero, Sonia Joseph, Ekdeep Singh Lubana, Talia Konkle, Demba Ba, and Martin Wattenberg. Into the rabbit hull: From task-relevant concepts in dino to minkowski geometry. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 154538–154584, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fa68d2dfd8eb6202e7902586197f7629-Paper-Conference.pdf.

Sheridan Feucht, Tal Haklay, Usha Bhalla, Daniel Wurgaft, Can Rager, Raphaël Sarfati, Thomas McGrath, Jack Merullo, Owen Lewis, Ekdeep Singh Lubana, Atticus Geiger, and Thomas Fel. Arithmetic in the wild: Llama uses base-10 addition to reason about cyclic concepts. *arXiv preprint arXiv:2605.01148*, 2026. URL https://arxiv.org/abs/2605.01148.

Collin Francel. SMixAE: Towards unsupervised manifold discovery in language models. *arXiv preprint arXiv:2605.09224*, 2026.

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

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2406.04093.

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

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/linebreaks/index.html.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. *arXiv preprint arXiv:2410.20526*, 2024.

Xinting Huang and Michael Hahn. Decomposing representation space into interpretable subspaces with unsupervised learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SI2OZJuvPO.

Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258, 1925.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL https://openreview.net/forum?id=CqViN4dQJk.

Dhruva Karkada, Daniel James Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=XQeMPEkfdd.

Yuxiao Li, Eric J. Michaud, David D. Baek, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, 2025. doi: 10.3390/e27040344. URL https://www.mdpi.com/1099-4300/27/4/344.

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. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

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

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models, 2026. URL https://openreview.net/forum?id=V8JaldWivi.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro A. M. Mediano. From data statistics to feature geometry: How correlations shape superposition. *arXiv preprint arXiv:2603.09972*, 2026. URL https://arxiv.org/abs/2603.09972.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with JumpReLU sparse autoencoders. *arXiv preprint arXiv:2407.14435*, 2024. URL https://arxiv.org/abs/2407.14435.

Magnus Sahlgren. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Institutionen för lingvistik, 2006.

Raphaël Sarfati, Eric Bigelow, Daniel Wurgaft, Siddharth Boppana, Jack Merullo, Atticus Geiger, Owen Lewis, Tom McGrath, and Ekdeep Singh Lubana. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors. *arXiv preprint arXiv:2602.02315*, 2026. URL https://arxiv.org/abs/2602.02315.

Or David Shafran, Shaked Ronen, Omri Fahn, Shauli Ravfogel, Atticus Geiger, and Mor Geva. From directions to regions: Decomposing activations in language models via local geometry. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Tz7n7pX9cO.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeffrey Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Isaac Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Mary Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, William Saunders, Eric J Michaud, Stephen Casper, Max Tegmark, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Thomas McGrath. Open problems in mechanistic interpretability. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=91H76m9Z94.

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

Adly Templeton, 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.

Lenka Tětková, Thea Brüsch, Teresa Dorszewski, Fabian Martin Mager, Rasmus Ørtoft Aagaard, Jonathan Foldager, Tommy Sonne Alstrøm, and Lars Kai Hansen. On convex decision regions in deep network representations. *Nature Communications*, 2025. URL https://doi.org/10.1038/s41467-025-60809-y.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Axbench: Steering LLMs? even simple baselines outperform sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=K2CckZjNy0.

Jehyeok Yeon, Federico Cinus, Yifan Wu, and Luca Luceri. Graph-regularized sparse autoencoders for LLM safety steering. *arXiv preprint arXiv:2512.06655*, 2025. URL https://arxiv.org/abs/2512.06655.

A SELECTING MUTUALLY EXCLUSIVE FEATURES

We select ME pairs in two steps: we put coupling strength and neighborhood similarity on a common scale (A.1), then choose a threshold on their combined score by comparing against a null graph (A.2). Both steps are carried out independently for each (model, layer, λ) combination.

A.1 RANK-BASED z-SCORING

Coupling strength and neighborhood similarity live on different scales, so we standardize them before combining. We use a rank-based z-score instead of the usual $(x - \mu)/\sigma$, because negative couplings have a heavy-tailed distribution and neighborhood similarity is bounded in $[0, 1]$. Neither is well summarized by its mean and variance.

For a variable x, let $\hat{F}_x$ be its empirical cumulative distribution function (CDF) over all candidate pairs, i.e. the normalized rank of each observed value, and let Φ be the standard normal CDF. The rank-based z-score is

$$z_x = \Phi^{-1}\big(\hat{F}_x(x)\big). \tag{6}$$

By construction, z_x follows a standard normal distribution (mean 0, variance 1) and preserves the ordering of x.

We apply this transformation to two quantities:

- **Coupling strength** z_c, computed from $-J_{ij}$ over pairs with $J_{ij} < 0$. The sign is flipped so that larger values mean stronger mutual exclusion.
- **Neighborhood similarity** z_n, computed from $\cos(J_i^+, J_j^+)$.

The combined score $s_{ij} = z_c + z_n$ is high only for pairs that are both strongly anti-coupled and share their contexts.

A.2 CHOOSING THE THRESHOLD τ^*

Because s_{ij} is continuous, there is no natural boundary between ME pairs and the rest. We therefore choose the threshold by controlling the false discovery rate (FDR) against a null model.

Null graph. We build a null graph J' by randomly permuting the couplings of J. This moves existing couplings between pairs without creating new edges. On J' we recompute coupling strength and neighborhood similarity and obtain null scores $s'_{ij} = z'_c + z'_n$.

Threshold. For a threshold τ, the number of pairs selected in the real graph and the expected number selected by chance are

$$N_{\text{real}}(\tau) = \big|\{(i,j) : s_{ij} > \tau\}\big|, \qquad N_{\text{null}}(\tau) = \big|\{(i,j) : s'_{ij} > \tau\}\big|, \tag{7}$$

where N_{null} is averaged over 3 independent permutations. We set

$$\tau^* = \min\left\{\tau \;:\; \frac{N_{\text{null}}(\tau)}{N_{\text{real}}(\tau)} < 0.05\right\}, \tag{8}$$

so that at most 5% of the selected pairs are expected to be false detections. Equivalently, the real graph yields at least $20\times$ as many pairs as the null graph above τ^*.

A.3 COUPLING THRESHOLD FOR COMPARISON PAIRS

Given the selected ME pairs, we define j^* as the smallest coupling magnitude $|J_{ij}|$ among them. We use j^* as a matching threshold when we compare ME pairs with strongly coupled pairs.

B ADDITIONAL DETAILS OF EXPERIMENTAL SETUP

Models and SAEs. Table 2 summarizes the models and SAEs we use. All SAEs are trained on the residual stream and are available for every layer of their model. They differ in architecture and training protocol:

- **Gemma Scope** (Lieberum et al., 2024) (Gemma-2-2b/9B): JumpReLU SAEs with a learned per-feature threshold. Decoder vectors have exactly unit norm.

- **Llama Scope** (He et al., 2024) (Llama-3.1-8b): trained as TopK SAEs ($k = 50$) and released as JumpReLU SAEs with a single global threshold. Decoder vectors are not normalized, and their norms grow with depth.
- **GPT-2 Small residual SAEs** (Bloom et al., 2024) (GPT-2 Small-res-jb): ReLU SAEs trained with an ℓ_1 sparsity penalty. Decoder norms are within 10^{-3} of 1.0.

Table 2: Models and SAEs used in this work.

Model	SAE	Architecture	d_{model}	Width N	Expansion
GPT-2 Small	`GPT-2 Small-res-jb`	ReLU (ℓ_1)	768	24,576	$32\times$
Gemma-2-2b	Gemma Scope	JumpReLU	2304	16,384	$7.1\times$
Gemma-2-9b	Gemma Scope	JumpReLU	3584	16,384	$4.6\times$
Llama-3.1-8b	Llama Scope	TopK $\rightarrow$ JumpReLU	4096	32,768	$8\times$

Data. We collect SAE activations on $T = 2$M tokens from the Pile (Gao et al., 2020). The same tokens are used to fit the co-activation graphs of every model and layer.

Alive features. We call a feature *alive* if its firing rate lies between $3 \cdot 10^{-4}$ and 0.99. The lower bound corresponds to roughly 600 activations over our corpus and removes extremely rare features, whose couplings cannot be estimated reliably. The upper bound removes features that are active on almost every token. Features outside this range are excluded from the co-activation graph and from all subsequent analyses.

Fitting the Ising Model To fit the Ising model, we collect SAE feature activations over T tokens, obtaining $\{\mathbf{s}^t\}_{t=1}^T$. We restrict our analysis to alive features. For each feature, we perform ℓ_1-regularized logistic regression for each feature, depending on all other features. The couplings, up to a positive factor, for feature i are given by

$$(\hat{h}_i, \hat{J}_i) \;=\; \underset{h,\, J_i}{\arg\min} \sum_{t=1}^{T} \log\Big(1 + \exp\big(-s_i^t(h + \textstyle\sum_{j\neq i} J_{ij}\, s_j^t)\big)\Big) \;+\; \lambda_i \textstyle\sum_{j\neq i} \mathrm{sd}_j\, |J_{ij}| \quad (9)$$

where $\mathrm{sd}_j = \sqrt{p_j(1-p_j)}$ for the empirical firing rate p_j of feature j. Note that the regularization coefficient depends on the feature. This is because firing rate varies highly between features. Suppressing all couplings independently of firing rate creates the same evidence threshold for each feature's coupling. This leads to rare features having far fewer couplings, simply because the number of tokens they fire on are fewer. To account for this, the regularization strength for each feature is

$$\lambda_i = \lambda \frac{\mathrm{sd}_i}{\tilde{\mathrm{sd}}} \quad (10)$$

where $\tilde{\mathrm{sd}}$ is the median standard deviation across all features. A rarer feature has a weaker regularization.

J is symmetric by construction, so after fitting couplings independently we must symmetrize. We do this by taking the mean of two couplings $\hat{J}_{ij}$ and $\hat{J}_{ji}$, under the condition that $\hat{J}_{ij} \neq 0$ and $\hat{J}_{ji} \neq 0$. This is to ensure a coupling is only created where both independent fits select for it. $J_{ij} = 0$ if and only if i and j are conditionally independent given all other features (Besag, 1974).

We fit co-activation graphs with five regularization strengths, $\lambda \in \{1.6{\cdot}10^{-4}, 2.5{\cdot}10^{-4}, 4{\cdot}10^{-4}, 6.3{\cdot}10^{-4}, 10^{-3}\}$, and find no qualitative differences in the graph statistics we report (Appendix C). To narrow the scope, all results in the main text use $\lambda = 4 \cdot 10^{-4}$.

C Properties of co-activation graphs

We fit the Ising model under different choices of the regularization coefficient λ for all models and all layers considered in our studies. Figure 10 reports how various graph statistics change with λ. The figure is a grid of plots. Each row reports one of the following graph statistics:

- **alive/width:** fraction of alive features over total number of features (SAE) width.
- **SAE L0:** the number of SAE features that are active per token, averaged over tokens.
- **Graph degree:** average number of edges per node in the graph
- **Edge probability:** total number of edges divided by maximum possible edges, $\frac{N(N-1)}{2}$ where N is the number of features.
- **%neg:** fraction of negative edges
- **%nodes$^-$:** number of nodes with at least one negative edge
- **%frustrated-triangles:** fraction of triangles in the graph whose edges have either (+ + -) or (- - -) signs.

Below we discuss in detail the observed behavior.

Stability of the Ising model across choices of λ The main effect of λ over all the statistics is a change in magnitude, but the trend across layer remains qualitatively the same. This justifies our usage of λ as a normalization on the graph sparsity and indicates that our results should stand for any reasonable choice of λ.

Size and sparsity. Of the 16k, 24k, and 32k features across models, $95\% - 99\%$ of total features are alive for the Gemma2 models, with between $45\% - 64\%$ for Llama and gpt-2-small. Even in the sparsest graphs, we encounter very few isolated features, with the giant connected component comprising over 90% of active features. There are no isolated features at our chosen regularization strength. The average degree per feature varies across model. For $\lambda = 4 \cdot 10^{-4}$, it peaks in Gemma-2-9b with each feature having an average of around 170 edges. Llama-3.1-8b, on the other hand, has only around 50 edges per feature. This is likely because llama's SAEs are far sparser, only using 29 features/token on average compared to Gemma-2-9b's 114. This means there are fewer opportunities for a pair of features to couple together, since on average they co-fire less often. These graphs are sparse, with edge probability ranging from 0.3% to 1%.

Community structure. A common trait of these graphs is their strong community structure. We measure modularity across 5 Louvain partitions over the positive subgraph, and take the maximum resulting in Q. As a reference, we generate a random graph with the same degree distribution and compute modularity the same way. This is Q_{swap}. The difference between these two values quantifies how much more modular these graphs are than random, with large differences resulting from graphs which have communities. These communities have been studied in Bhalla et al. (2026) and have been shown to related to "concept manifolds." We do not rely on this community structure in this work and instead just point out that these graphs are sparse, yet highly structured. Features for communities and their connections are far from random. We also compute the modularity of an Erdős-Rényi graph of matched size and density. These comparisons are reported in Figure 11.

Negative couplings are not rare In these graphs, positive couplings are far more common than negative. However, a moderate number of features have a negative coupling, showing that suppressive relationships are not rare as a feature property. We do observe suppressive "hubs" - features with disproportionally many negative couplings. The top 10% of features by negative coupling count hold 46-73$\%$ of total negative couplings across models.

Increasing lambda, we notice that negative couplings are eliminated faster than positive. Stepping λ from $4 \cdot 10^{-4}$ to 10^{-3}, 34-39$\%$ of positive edges remain, while only $4 - 10\%$ of negative edges remain. This shows the estimator prioritizes positive couplings over negative ones.

D SUPPORTING MATERIAL TO CHARACTERIZATION OF ME PAIRS

Figure 12 summarizes the number of discovered ME pairs and cliques across models and layers. The number of discovered pairs and cliques in Gpt2-small is significantly lower number than the other models. Table 3 complements the figure reporting the total number of ME pairs across clique size.

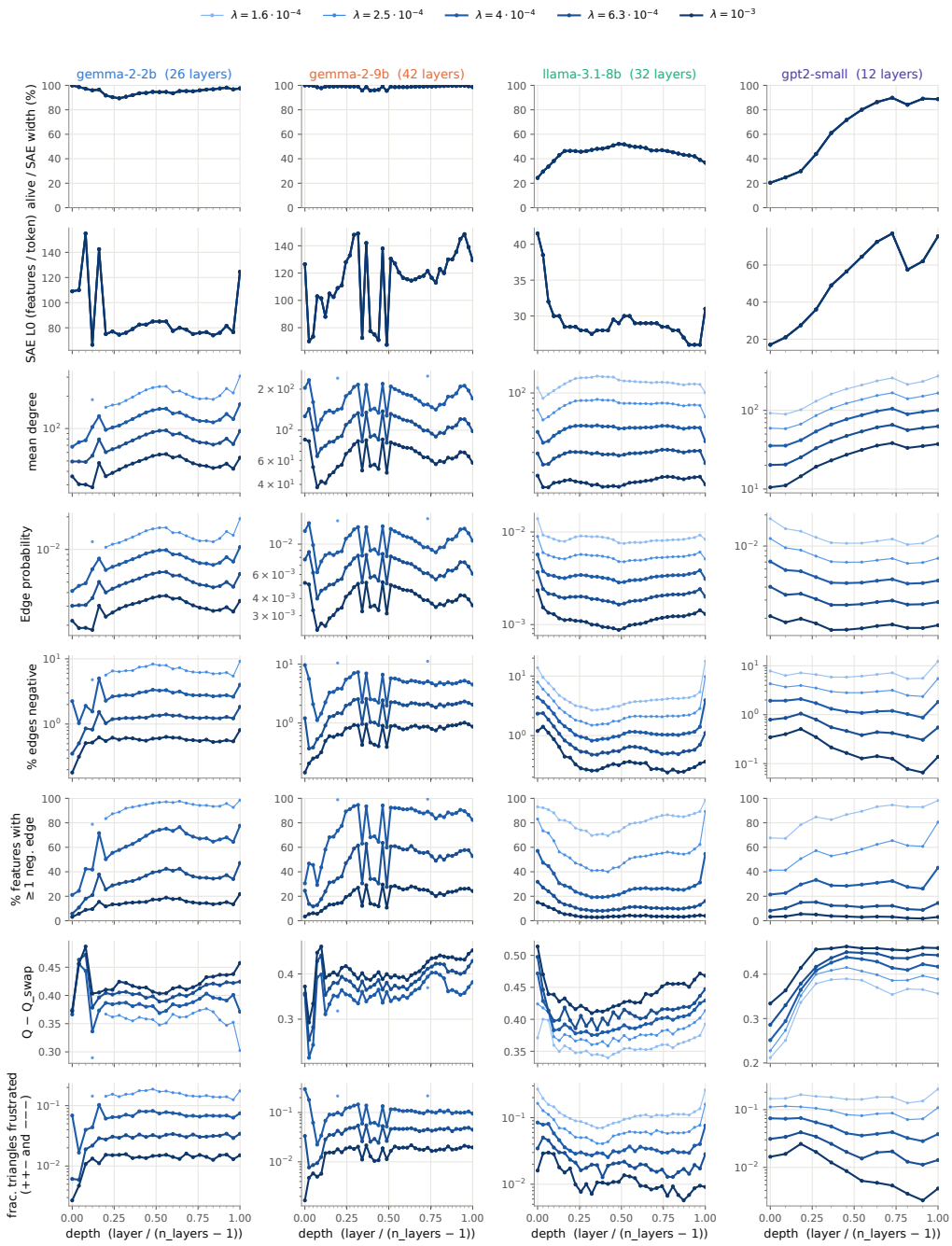

Figure 10: **Summary of graph statistics across layers.** Each row of panels reports a different graph statistic, and each columns a different model among the four studied in this work. In each panel, the solid lines of different blue shades represent the values of the statistic as a function of layer depth under different regularization strength λ. The first two statistics, measuring the graph sparsity, are not influenced by λ. In the remaining ones, λ mainly changes the magnitude, preserving the trend across layers.

Lone ME pairs versus ME pairs in cliques. Figure 13 complements the results presented in Figure 4 for Gemma-2-2b.

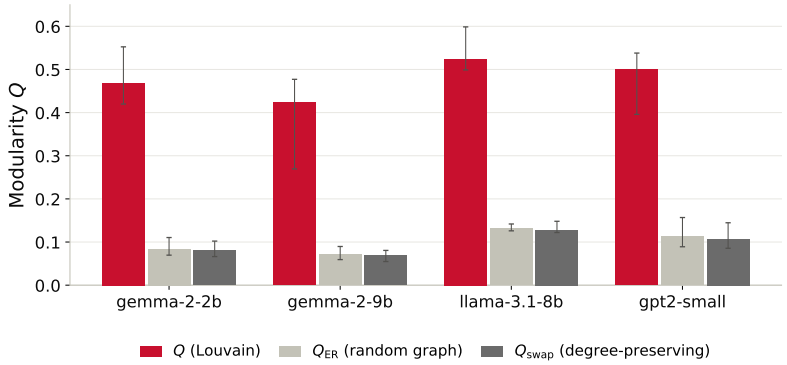

Figure 11: **Summary of modularity degree across models.** We compute the maximum modularity computed from 5 Louvain partitions of all the co-activation graphs studied in this work and report its average value for each model. We compare the observed value (Q) with two baselines, obtained by computing (1) the Erdős-Rényi graph (Q_{ER}), and (2) a degree-preserving random graph (Q_{swap}).

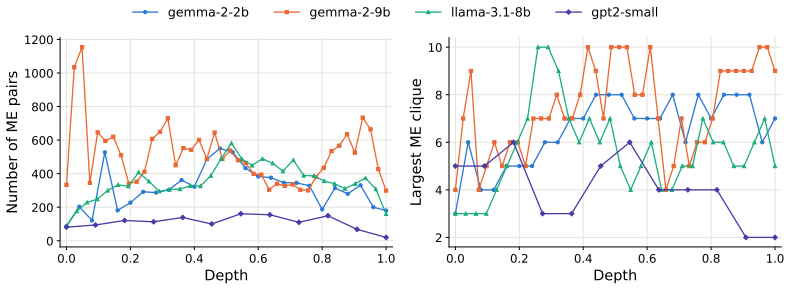

Figure 12: **Summary of discovered ME pairs across models and layers.**

ME pairs share communities. We use the Louvain partitions computed for the modularity analysis above to measure how often the two features of a pair fall in the same community. ME pairs share a community more often than strongly positively coupled pairs, and pairs within cliques more often than lone pairs (Figure 14). This is notable because the partition uses only positive edges: ME features are grouped together even though they are not directly connected, whereas positively coupled pairs are. Their co-membership is instead driven by their shared neighbors, the same property we select them on. Together, these results suggest that ME features lie within a common region of the graph, consistent with the view that they sit on the same concept manifold (Bhalla et al., 2026).

E Interpretability of ME features: additional results

Figure 15 shows how the average ME degree of clique pairs varies conditioned on the description verdict output by the LLM judge used to score the clique-level description. Figure 16 shows how often some specific ME cliques appear across models and layers for Gemma-2-2b, Gemma-2-9b and Llama-3.1-8b.

Model	Lone pairs	k = 3	k = 4	k ≥ 5
Gemma-2-2b	5,648	360	138	156
Gemma-2-9b	15,439	981	320	305
Llama-3.1-8b	8,013	708	123	70
GPT-2 Small	1,137	21	8	10

Table 3: **Number of ME pairs across clique sizes.** This breaks down the number of lone pairs and number of cliques of different sizes. The LLM judge was given every lone pair and every maximal clique and returned a verdict for each.

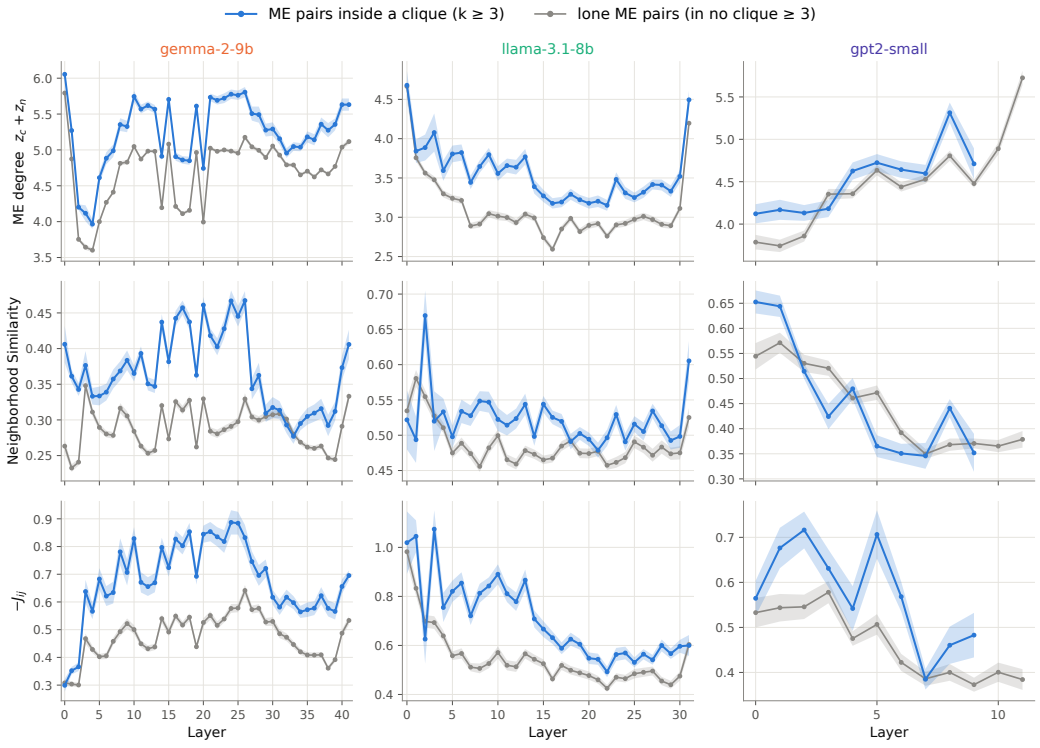

Figure 13: **Lone ME pairs versus ME pairs in cliques.** We split ME pairs into lone pairs (gray) and pairs that belong to a clique (blue). For Gemma-2-9b, Llama-3.1-8b and GPT-2 Small (columns), we report as a function of layer depth: the mean ME degree $z_c + z_n$ (first row), the mean neighborhood similarity $\cos(J_i^+, J_j^+)$ (second row), the mean coupling strength $-J_{ij}$ (third row). Pairs in cliques are fewer than lone pairs, but in most layers they have both higher neighborhood similarity and stronger negative couplings, and hence a higher ME degree. The main exception is GPT-2 Small, where lone pairs have higher neighborhood similarity in some layers.

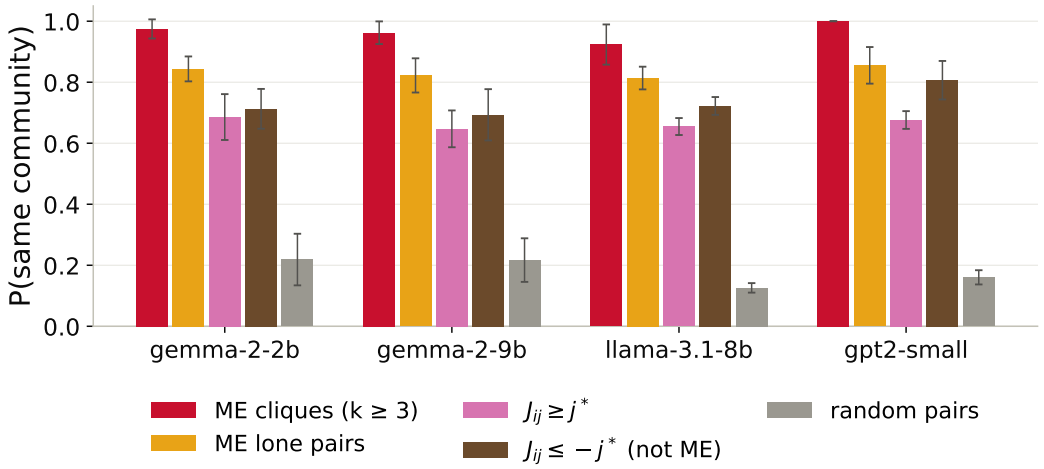

Figure 14: **ME pairs share communities in the co-activation graph.** For each model, we report the probability that the two features of a pair belong to the same Louvain community. We compare ME pairs in cliques (red), lone ME pairs (orange), strongly positively coupled pairs (pink), strongly negatively coupled pairs that are not ME (brown), and random pairs (gray). Strongly coupled pairs are those with $|J_{ij}| > j^*$ (Appendix A). ME pairs share a community more often than all baselines, and pairs in cliques more often than lone pairs.

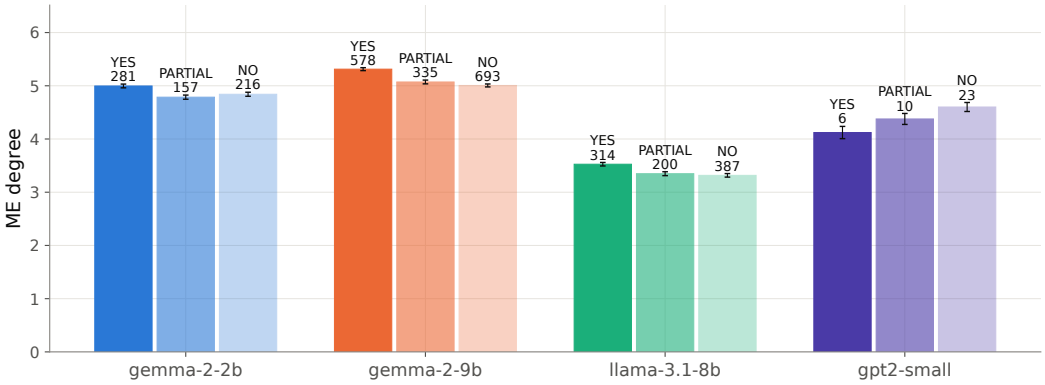

Figure 15: **Clique ME degree vs. description verdict.** Bars track the average ME degree for cliques across the three different description verdicts. This shows for the larger models that the cliques that offer a clear description have slightly higher ME degree.

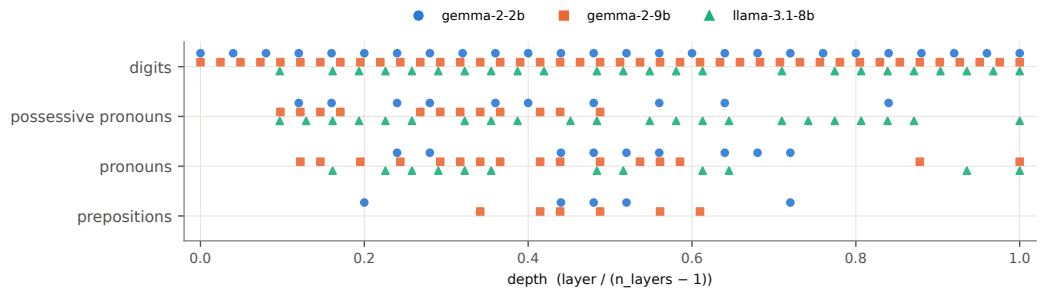

Figure 16: **Recurrence of ME cliques across layers and models.** We report four ME cliques that recur across layers in the three largest models we test. By far, the most found clique is digits, with recurrence at every layer in both gemma models and $24/32$ layers in Llama-3.1-8b. For GPT-2 Small, we find no cliques which are repeated across layers.

F LLM JUDGE PROMPT

Prompt — per maximal clique

You are reading sets of sparse-autoencoder features from a language model. The features in a set pairwise suppress each other and activate in similar contexts. The hypothesis under test is that such a set is a TYPE — a slot that holds exactly one VALUE at a time — and its members are the alternative values (the ten digits; the days of the week; singular vs. plural; a set of programming languages).
For each member you are given: the SAE feature id, its firing rate, its PEAK TOKENS (the token at the maximum activation in each Neuronpedia exemplar, counted across exemplars, with the share of exemplars), an auto-interp EXPLANATION written by another model, and the top OUTPUT tokens it promotes. Peak tokens are the more trustworthy signal; explanations are often generic ("numerical data").
Two labelled examples, in the format you will see and the format you must answer in.

Example 1 (`ex_YES`), layer 8, $k = 7$

feature	rate	peak tokens (share)	explanation	top output
5362	1.76e-03	his 0.82, His 0.05, is 0.02	the possessive pronoun "his"	/her own
5601	1.03e-03	its 0.89, Its 0.04	phrases emphasizing ownership or possession	own _own
10189	1.72e-03	their 0.93, su 0.02	the possessive pronoun "their."	themselves own
22121	1.94e-03	your 0.75, Your 0.07, Your 0.05	the word "your" in various contexts	SELF
24331	1.33e-03	her 0.89, Her 0.02	references to female subjects and their roles	/her own
24671	1.27e-03	my 0.89, may 0.02	possessive terms indicating ownership	SELF
25487	1.54e-03	our 0.85, OUR 0.02, Our 0.02	references to collective identity or community ownership	own

Answer row: id=ex_YES, layer=8, k=7, proposed type="possessive determiner (his/her/its/their/your/my/our)", mutually exclusive?=YES, distinct/duplicates="7 distinct, 0 duplicates", evidence="each member peaks > 0.75 on one possessive; a noun phrase takes one determiner", confidence=high.

Example 2 (`ex_NO`), layer 5, $k = 5$

feature	rate	peak tokens (share)	explanation	top output
7858	1.34e-03	winner 0.12, winners 0.09, submissions 0.07, win 0.05	text related to contests and competitions	submissions prizes
12666	2.71e-04	number 0.70, ⊂ 0.03, Ĭ 0.02	numerical quantities or counts	number of
14289	2.23e-04	circulation 0.19, circ 0.11, circulate 0.08	terms related to the circulation of fluids or objects	circulation
22477	3.78e-03	Ļ 0.70, – 0.02, , 0.01	words related to political figures and their actions	(byte-level artefact tokens)
23553	6.69e-03	is 0.27, has 0.12, was 0.06, have 0.05	phrases discussing actions, decisions, or events	nevertheless

Answer row: id=ex_NO, layer=5, k=5, proposed type="unrelated: contest / number / circulation / byte token / auxiliary verb", mutually exclusive?=NO, distinct/duplicates="5 distinct but not values of one slot", evidence="peaks are unrelated content words plus a formatting byte; no shared slot", confidence=high.

For every set below, fill one row of a table with these columns, in order: id, layer, k, proposed type (3–6 words), values look mutually exclusive?, distinct values / duplicates, evidence, confidence.

- id: copy the set's id exactly.

- `proposed type`: what slot the members are values of, in your words.
- `values look mutually exclusive?`: YES if the members read as alternative values of one slot (one would not expect two of them on the same token); PARTIAL if most do but some members are of a different kind; NO if the members look like a co-firing bundle, a generic/junk set, or unrelated. For a set of two ($k = 2$) the question is the same: do the two read as two values of one slot?
- `distinct values / duplicates`: how many members are clearly distinct values, how many look like near-duplicates of another member (same peak token), e.g. "9 distinct, 1 duplicate of 2302".
- `evidence`: the peak tokens that carry the reading, briefly.
- `confidence`: high / medium / low.

Do not speculate beyond the tokens shown. If the peak tokens are whitespace or formatting, say so and mark the set NO or PARTIAL. Do not invent members or ids. Output the header row once, then one row per set, and nothing else — no text before or after the table.

Sets. `<sets>` *(the clique's feature sets are substituted here, each formatted like the two worked examples above)*

G Extended Table of representative ME Cliques

Table 4: More examples of ME cliques across models. Same conventions as Table 1. GPT-2 Small yields few cliques the LLM judges as a single attribute. Those it does are shown.

Model/Layer	Feature Ids	Top activating tokens	Clique description
Gemma-2-2b/12	10229, 13430, 8081, 1185, 1889, 5327, 9331, 6594	`1`, `3`, `4`, `5`, `6`, `8`, `9`, `7`	numerical values (0-9)
Gemma-2-2b/3	3925, 9763, 12722, 7421	`June`, `April`, `October`, `March`	month names
Gemma-2-2b/11	972, 4467, 8940, 5306	`I`, `He`, `They`, `We`	personal pronouns
Gemma-2-2b/12	1861, 6910, 787, 5679, 11480, 1032	`its`, `their`, `your`, `My`, `his`, `our`	possessive pronouns (my/your/our/their)
Gemma-2-2b/12	2230, 8655, 1846	`in`, `at`, `on`	prepositions
Gemma-2-2b/13	3177, 6192, 14063	`that`, `who`, `which`	pronouns and questions
Gemma-2-2b/17	10364, 11887, 12591	`can`, `could`, `may`	modal verbs
Gemma-2-2b/18	9704, 11925, 1418	`was`, `is`, `be`	auxiliary verbs
Gemma-2-2b/11	2904, 11814, 3482	`make`, `take`, `give`	verbs related to making, giving, taking
Gemma-2-2b/15	12979, 15324, 7136	`run`, `yards`, `goals`	sports performance statistics
Gemma-2-2b/21	86, 11591, 12816	`c`, `G`, `C`	nucleotide sequences
Gemma-2-9b/21	2498, 4513, 6079, 4328, 14292, 15373, 4170, 6201, 2453, 12529	`3`, `2`, `0`, `9`, `1`, `8`, `4`, `6`, `5`, `7`	numerical digits
Gemma-2-9b/15	13399, 4933, 16019, 6934, 8159, 10857	`I`, `it`, `She`, `He`, `We`, `They`	pronouns of various persons
Gemma-2-9b/18	8413, 14050, 525, 13516, 1584, 3856, 5818	`its`, `her`, `our`, `Your`, `his`, `My`, `their`	possessive pronouns
Gemma-2-9b/24	10091, 12147, 11385	`that`, `who`, `which`	relative pronouns
Gemma-2-9b/36	1203, 1964, 1020, 15198	`would`, `will`, `may`, `should`	modal verbs

continued on next page

Table 4 continued from previous page

Model/Layer	Feature Ids	Top activating tokens	Clique description
Gemma-2-9b/6	14549, 9249, 2497	is, are, was	auxiliary verbs (is/are/was)
Gemma-2-9b/20	11301, 5319, 4815	at, on, in	prepositions
Gemma-2-9b/17	4427, 3759, 532	each, all, every	inclusivity and universality phrases
Gemma-2-9b/18	6073, 4819, 10155	week, month, year	time units
Gemma-2-9b/15	5856, 5309, 9497, 3494	football, basketball, baseball, United	sports (football/basketball/baseball)
Gemma-2-9b/0	3793, 4714, 6817, 7060	purposes, part, respect, the	references in legal texts
Llama-3.1-8b/6	4954, 2843, 13783, 19511, 9094	5, 9, 7, 6, 8	numerical values (digits 0-9)
Llama-3.1-8b/9	24305, 11503, 12242, 29889, 14717, 9276, 29752, 22921, 30651, 6294	6, 4, 8, 9, 3, 5, 7, 12, 10, 21	numerical data (0-9 digits)
Llama-3.1-8b/24	13112, 12917, 17569	first, third, second	ordinal numbers
Llama-3.1-8b/6	4658, 711, 14233	West, East, north	geographic directions (East/West/North)
Llama-3.1-8b/28	24195, 30495, 5407, 22957, 27087	would, may, should, could, will	modal verbs indicating ability or obligation
Llama-3.1-8b/10	20354, 6452, 27807	their, his, its	possessive pronouns
Llama-3.1-8b/11	30243, 23208, 11536	that, which, who	relative pronouns
Llama-3.1-8b/26	2425, 17510, 15835	year, day, week	references to time periods
Llama-3.1-8b/10	29843, 12940, 13610	what, how, why	question words
Llama-3.1-8b/12	6454, 18905, 7521	that, this, these	demonstrative pronouns
Llama-3.1-8b/19	20131, 25157, 40, 15840	ns, 5, mysql, int	SQL/database commands
Llama-3.1-8b/16	27727, 21531, 21101, 29362	␣, ␣, the, ␣	sports-related statistics
GPT-2 Small/3	18979, 8400, 13532	blood, b, protest	mentions of protests
GPT-2 Small/2	23320, 17611, 13674, 18049	windows, frame, \n, sever	phrases related to frames or framing
GPT-2 Small/2	15980, 7756, 15569, 8156, 13051, 16136	seeds, PL, Ñĭ, of, the, \n	acronyms related to organizations

H GEOMETRIC INTERPRETATION OF ME CLIQUES: ADDITIONAL RESULTS

In this appendix we report additional results about the geometric interpretation of ME features. Figure 17 shows the decoder similarity for lone pairs of ME features (yellow) and ME pairs belonging to cliques (red). Clique pairs are the ones with the most similar directions. Lone pairs (yellow), in the case of Gemma-2-2b and Gemma-2-9b, are anti-aligned with one another. In Llama-3.1-8b and GPT-2 Small, on the other hand, we see that ME pairs are generally very similar to each other, akin to ME cliques. Random pairs (gray) are orthogonal, as is expected by superposition. Positively coupled pairs are moderately similar, showing an increase with depth. The $J_{ij} < -j^*$ pairs (brown) show that both conditions of mutual exclusivity define this geometric property. Table 5 shows the rank correlation between ordered cliques and their first principal component (PC1). Figure 18 reports additional examples of two-dimensional PCA plots of ordered cliques.

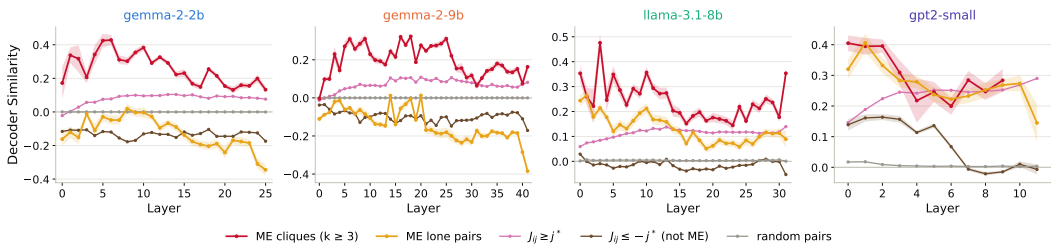

Figure 17: **Decoder similarity across depth.** ME cliques (red) are the ones with the most similar directions. ME pairs (yellow), in the case of Gemma-2-2b and Gemma-2-9b, are anti-aligned with one another. In Llama-3.1-8b and GPT-2 Small, on the other hand, we see that ME pairs are generally very similar to each other, akin to ME cliques. Random pairs (gray) are orthogonal, as is expected by superposition. Positively coupled pairs are moderately similar, showing an increase with depth. The $J_{ij} < -j^*$ pairs (brown) show that both conditions of mutual exclusivity define this geometric property. It is not merely that features in an ME clique repel each other, it is also that they have similar neighborhoods.

model	n	$\lvert\rho(\text{PC1}, \text{value})\rvert$	Totally ordered
Gemma-2-2b	58	0.69 ± 0.04	12%
Gemma-2-9b	138	0.71 ± 0.02	13%
Llama-3.1-8b	19	0.78 ± 0.05	21%

Table 5: **Rank correlation between ordered cliques and their first PC.** For an ordered clique, we take the ordering of the feature labels given by the first PC. We compare this ordering to the true ordering (i.e. for digits $0 - 1 - 2 - \cdots - 9$). We compute the rank correlation between these two orderings and find that in many cases the first PC completely orders the values (totally ordered). This demonstrates that within a clique, low-dimensional structure encodes the semantic relationships between features. Since sign of PC's is arbitrary, we consider the absolute value of this correlation.

I Further Steering Results and Details

Alternative feature baseline. To test whether the effect is specific to the ME features, we replace f_t with a variety of alternative features: (i) the most strongly *positively* coupled feature (largest J_{sp}); (ii) the most strongly negatively coupled feature that is not an ME feature of that layer (low co-activation without the neighborhood similarity); (iii) the feature with the highest neighborhood similarity $\cos(J_s^+, J_p^+)$ that is not an ME feature (similar neighborhood without exclusion); and (iv) random features. Clique members and the ME target are never used as partners. If ME features were merely strongly coupled or merely neighborhood-similar directions, one of these partners would close the gap as well.

Vanilla feature-selection baseline. We additionally compare ME features against a standard activation-based feature selection procedure Templeton et al. (2024); Wu et al. (2025). For each token u associated with an ME feature, over the same corpus of text used to fit the co-activation graphs, we find all occurrences of u and record activations of all alive SAE features. We select the feature with the highest mean activation across these occurrences, including zeros, and refer to it as the *vanilla feature* for u. We then repeat each substitution experiment using the vanilla features selected for s and t in place of the corresponding ME features. Because this procedure selects features independently for each token, the same vanilla feature can be selected for multiple tokens; when the source and target tokens select the same feature, the substitution is zero by definition. Such cases are excluded from our results.

Summary of the results. Figure 19 summarizes the results of our steering experiments.

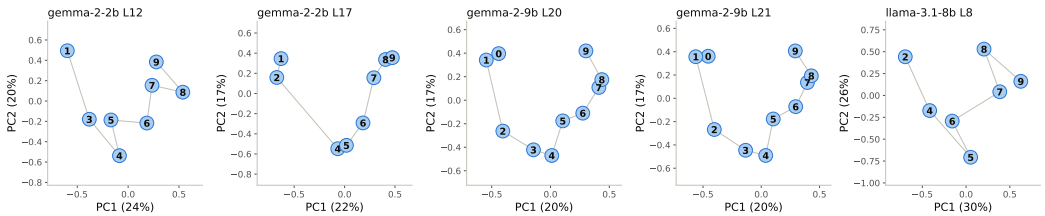

Figure 18: **Other examples of 2D PCA plots of ordered cliques.** Across models, we see that the ordering of digit cliques is present to varying degrees in the first two PC's. Generally, however, the digits seem to lie on an arc around the mean decoder direction. These projections do not capture a high amount of the variance, yet the geometric arrangement of the order information is made clear.



Figure 19: **Gap closure across cliques and models.** Each point is one substitution (source value $\rightarrow$ target value) at the clique's layer; black ticks mark medians. We swap the source feature for its ME partner (red) or for one of four controls: the most positively coupled partner (pink), the most negatively coupled non-ME feature (brown), the most neighborhood-similar non-ME feature (orange), and a norm-matched random direction (gray). Digits are shown separately for Gemma-2-2b, Gemma-2-9b and Llama-3.1-8b. Every other row pools the models in which that clique was found. ME substitutions close most of the gap for digits and compass directions, part of it for prepositions, and little of it for the remaining cliques. Wherever ME substitutions succeed, the controls stay well below them.

J GPT-2 Small discrepancies

GPT-2 Small departs from the other three models in three places, which we collect here.

Lone pairs versus cliques. In the other three models, ME pairs inside cliques have both higher neighborhood similarity and stronger negative couplings than lone pairs in almost every layer (Figures 4 and 13). In GPT-2 Small, pairs in cliques still have stronger negative couplings and higher ME degree, but their neighborhood similarity exceeds that of lone pairs in only 4 of 10 layers. GPT-2 Small yields only 39 cliques across its 12 layers, so these per-layer comparisons rest on few pairs.

Describability and clique size. In the other models, the share of ME sets that the LLM judge describes as a single type (`YES`) rises monotonically with size (Figure 2). In GPT-2 Small it stays low and does not rise monotonically: 5% for lone pairs and 14%, 12% and 20% for $k = 3$, 4 and

≥ 5, based on only 21, 8 and 10 cliques. Describable GPT-2 Small cliques also do not have higher ME degree (Figure 15). One possible explanation is the SAE: GPT-2 Small uses a ReLU SAE with a $32\times$ expansion factor, much larger than the others (Table 2). A wider dictionary may split the values of a type into narrower, less interpretable features (Chanin et al., 2024).

Decoder geometry. In both Gemma models, lone ME pairs and strongly negatively coupled non-ME pairs have anti-aligned decoder directions; in GPT-2 Small both are positively aligned (mean cosine 0.27 and 0.08), with Llama-3.1-8b in between (Figures 6 and 17). This is less surprising given how crowded GPT-2 Small's dictionary is. For every feature in every layer, we compute the cosine similarity between its decoder direction and that of its nearest neighbor, then average over features and layers. This nearest-neighbor similarity is 0.545 in GPT-2 Small, compared with 0.264, 0.230 and 0.339 for Gemma-2-2b, Gemma-2-9b and Llama-3.1-8b. Random pairs remain orthogonal in all models, so the crowding is local: GPT-2 Small features typically have a much closer neighbor, which makes aligned pairs common regardless of how they are coupled. This is consistent with its larger expansion factor, and may also reflect that it is the only ReLU SAE we study.