Correlations in the Data Lead to Semantically Rich Feature Geometry Under Superposition

Lucas Prieto , Edward Stevinson , Melih Barsbey,
Tolga Birdal*, Pedro A.M. Mediano*
Imperial College London

Abstract

Recent advances in mechanistic interpretability have shown that many features of deep learning models can be captured by dictionary learning approaches such as sparse autoencoders. However, our geometric intuition for how features arrange themselves in a representation space is still limited. "Toy-model" analyses have shown that in an idealized setting features can be arranged in local structures, such as small regular polytopes, through a phenomenon known as *superposition*. Yet these local structures have not been observed in real language models. In contrast, these models display rich structures like ordered circles for the months of the year or semantic clusters which are not predicted by current theories. In this work, we introduce Bag-of-Words Superposition (BOWS), a framework in which autoencoders with a ReLU in the decoder are trained to compress sparse, binary bag-of-words vectors drawn from Internet-scale text. This simple set-up reveals the existence of a *linear regime* of superposition, which appears in ReLU autoencoders with small latent sizes or which use weight decay. We show that this linear PCA-like superposition naturally gives rise to the same semantically rich structures observed in real language models. Code is available under https://anonymous.4open.science/r/correlations-feature-geometry-AF54.

1 Introduction

A key challenge in understanding the inner workings of deep learning (DL) models is the fact that they seem to encode features in superposition [Elhage et al., 2022]. This allows the models to represent more features than they have neurons, at the cost of allowing some interference between the features, making neurons polysemantic and harder to interpret. Elhage et al. [2022] showed that autoencoders (AEs) with a ReLU in the decoder can leverage the non-linearity to encode sparse features in superposition forming local geometries like regular polytopes. However, these local geometries have not been observed in real models. In contrast, real language models display semantically rich geometries, such as ordered circles for the months of the year [Engels et al., 2025] or semantically related features clustering together [Bricken et al., 2023]. Our lack of understanding of feature geometry has recently been highlighted as one of the key open problems in mechanistic interpretability [Sharkey et al., 2025] (further discussion of related work is provided in Appendix A).

In this work, we introduce the bag-of-words superposition framework (BOWS), which mimics the writing and reading of information from the residual stream of models like transformers. In the BOWS framework, we encode bag-of-words representations of internet text into a lower dimensional latent space using an AE with a ReLU in the decoder, similar to the one used by Elhage et al. [2022]. BOWS shows that when encoding features with realistic correlations, AEs can learn principal components (PCs) of the data in a linear form of superposition which gives rise to structure in the weights that

*Joint senior authors, equal contribution

Mechanistic Interpretability Workshop (NeurIPS 2025).

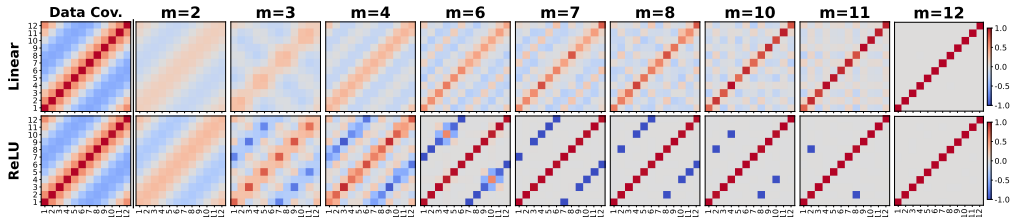


Figure 1: **Autoencoding synthetic correlated features reveals different kinds of superposition in non-linear networks.** Weight pattern inner products ($\mathbf{W}^T\mathbf{W}$) at convergence for AEs encoding $d = 12$ synthetic features with a cyclic covariance structure, varying latent size m. **Top Row (Linear AE):** Always reflects the top m principal components of the covariance. For $m = 2$, it captures the circular structure representing all 12 features. **Bottom Row (ReLU AE):** Mimics the linear AE for small m (linear regime, e.g., $m = 2, 3$), but diverges for larger m, forming antipodal pairs (non-linear "shattered" regime, e.g., $m = 4$ to $m = 11$) to better utilize the ReLU for interference reduction.

reflects the structure in the data correlations. We also find that this linear superposition is more prevalent under tight bottlenecks (latent dimension $\ll$ input size) and with the use of weight decay.

2 BOWS: Realistic data in superposition

We now describe the BOWS setup and the training procedure used throughout the paper.

Dataset. Let $\mathcal{C}$ be a corpus of English text segmented into *records* (lines or paragraphs). After word-level tokenization, we construct a vocabulary of the V most frequent words, discarding common stop-words and prepositions. This vocabulary includes words such as *sun*, *code*, and *January* which often correspond to linear features in sparse autoencoders trained on language data [Engels et al., 2025, Bricken et al., 2023]. Each record is then encoded as a binary bag-of-words vector $\mathbf{x} \in \{0, 1\}^V$ whose j-th component is 1 iff the j-th vocabulary word appears in the record.

We choose a *context size*, $c \in \mathbb{N}$. For every contiguous block of c records we take the element-wise logical OR of their individual vectors, obtaining a single sample. The resulting dataset is

$$\mathcal{D} \;=\; \{\mathbf{x}_i\}_{i=1}^{N}, \qquad \mathbf{x}_i \in \{0,1\}^V, \tag{1}$$

where N is the number of c-record chunks in the corpus.

All experiments in this paper use the WikiText-103 corpus [Merity et al., 2017]. With $V = 10{,}000$ and $c = 20$ we obtain $N = 1{,}801{,}255$ training examples. We refer to this pre-processed collection as WikiText-BOWS.

Autoencoder. We use the autoencoder setup introduced in Elhage et al. [2022] to study superposition, consisting of an encoder with weights $\mathbf{W} \in \mathbb{R}^{m\times V}$ and bias $\mathbf{b} \in \mathbb{R}^V$, where the input $\mathbf{x} \in \mathbb{R}^V$ is reconstructed using a ReLU AE with reconstruction loss:

$$\mathcal{L}_{\text{ReLU}-\text{AE}}(\mathbf{x}, \mathbf{W}, \mathbf{b}) = ||\mathbf{x} - \text{ReLU}(\mathbf{W}^T\mathbf{W}\mathbf{x} + \mathbf{b})||_2^2 \tag{2}$$

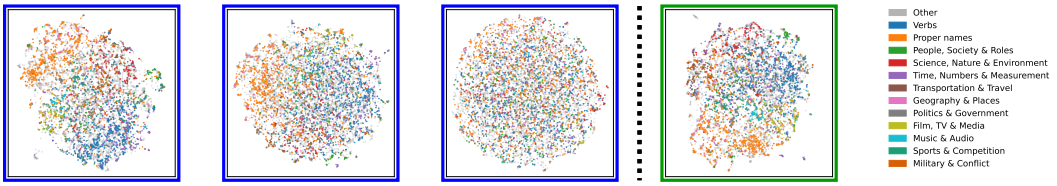
We also use a Linear AE as a baseline with loss:

$$\mathcal{L}_{\text{Linear-AE}}(\mathbf{x}, \mathbf{W}, \mathbf{b}) = ||\mathbf{x} - (\mathbf{W}^T\mathbf{W}\mathbf{x} + \mathbf{b})||_2^2 \tag{3}$$

3 A minimal example of linear and non-linear superposition

When features are i.i.d, as in Elhage et al. [2022], a Linear AE can only capture one feature per principal component, all of them represented orthogonally with no superposition. Under these i.i.d. assumptions, superposition (representing $d > m$ features) seems intrinsically linked to mechanisms involving non-linearities or specific geometric arrangements like polytopes designed to actively manage interference between non-orthogonal features.

However, features in real-world data, such as words or concepts in natural language, are rarely independent. They exhibit rich correlation patterns driven by semantic relationships and contextual co-occurrence. This correlation structure significantly lowers the *effective rank* of the feature space compared to the i.i.d. case. In this section we show how linear and non-linear AEs can leverage these correlations to represent more features than they have latent dimensions by learning PCs.



(a) $m = 200$ $wd = 0$ (b) $m = 400$ $wd = 0$ (c) $m = 600$ $wd = 0$ (d) $m = 600$ $wd = 4$ (e) Semantic categories

Figure 2: **Linear superposition appears in ReLU AEs which have small latent sizes (a) or are trained with weight decay (d), giving rise to semantic clusters.** UMAP projections of word embeddings from AEs of different latent dimensions (m) and weight decay values (wd). Points are colored by semantic category (e).

3.1 PCA as a form of linear superposition

Linear AEs with a latent dimension m are known to span the subspace defined by the top m PCs of the data, if the data is centered (or the top m singular vectors if the data is not centered) [Baldi and Hornik, 1989]. Thus, if the input data consists of correlated features $\mathbf{x} \in \{0, 1\}^V$ such that the top $m \ll V$ singular vectors explain 95% of the variance, a linear AE can attain an average R^2 score of 0.95 by capturing the top m singular vectors in a purely linear reconstruction. We argue that leveraging these correlations to encode V features in an m-dimensional space should itself be considered a fundamental form of superposition.

Figure 1 (top row, Linear) demonstrates this clearly using our BOWS setup. A Linear AE trained to reconstruct $d = 12$ features from a synthetic data distribution, generated with a cyclic covariance structure (mimicking concepts such as months or days of the week). Even with a highly compressed latent space ($m = 2$), the linear AE successfully represents all 12 features by arranging their corresponding weight vectors ($\mathbf{W}$ columns, visualized via $\mathbf{W}^T\mathbf{W}$) according to the top two principal components of the cyclic covariance matrix, which naturally form a circle. This explicitly shows that linear dimensionality reduction enables a form of superposition ($d = 12 > m = 2$) by exploiting feature correlations, without requiring any non-linearity.

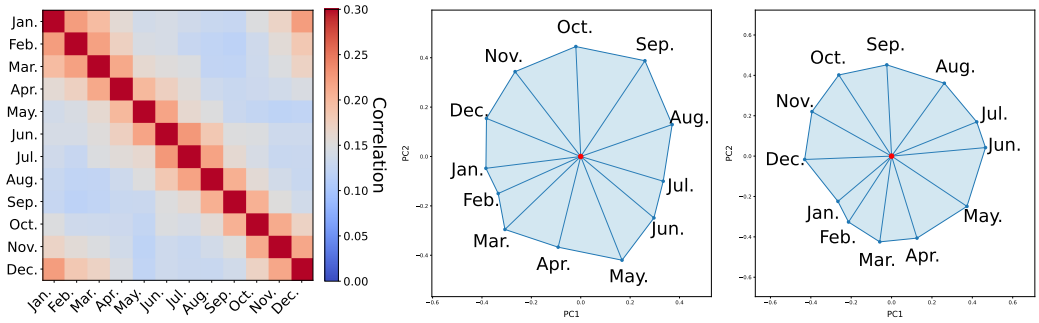
3.2 Two regimes of superposition in non-linear AEs

How does this picture change when we instead use the ReLU AE? Our key finding is that the behavior depends critically on the degree of compression (m/d).

Linear superposition and the PCA-regime. When the bottleneck is very tight ($m \ll d$), the best the ReLU AE can do is capture the large-variance components of the data. In this simple setting, the ReLU AE behaves remarkably similarly to the Linear AE. As seen in Figure 1 (bottom row, ReLU, $m = 2, 3$), the ReLU AE also recovers the circular structure dictated by the top principal components. This might indicate that the benefit of using the ReLU to form specialized structures for interference mitigation is outweighed by the need to capture the fundamental covariance structure first.

Non-linear superposition shatters the covariance structure. Figure 1 (bottom row, ReLU, $m \geq 6$) hows that as m increases, the ReLU AE abandons the circular PCA structure and instead uses the ReLU to represent features as *antipodal pairs*. This specific geometry is one of the cases studied by Elhage et al. [2022], whereby features are placed in anti-correlated pairs such that activating one feature negatively activates its antipodal partner, which is then zeroed out by the ReLU. Crucially, each antipodal pair is almost perfectly uncorrelated with every other pair. We term this non-linear superposition, where the global covariance structure is increasingly broken or "shattered" in favor of local geometries optimized for the non-linearity.

Structure disappears as features become orthogonal. As the latent size approaches the input size ($m = 12$), the AE weights converge to the identity, representing each feature orthogonally (no circular structure) for a perfect reconstruction.



(a) Data correlations (b) PCA of the data (c) PCA of the weights

Figure 3: **Circular representation of months arises from real data covariance via PCA. (a)** Empirical correlation matrix of month words in the WikiText-103 BOWS dataset, showing cyclic correlations. **(b)** Top 2 PCs of binary word occurrence data for the 12 month, reveals a circle. **(c)** PCA applied to the 12 learned encoder features (W columns) for months from a ReLU AE trained on WikiText-BOWS with ($m = 1000$), projected onto their top 2 PCs, also recovers the circular structure. This suggests the AE inherits the structure from the data via PCA-like compression.

4 Linear superposition explains feature geometry in realistic data

We now explore the pheonmenon of linear superposition in our WikiText BOWS setting and show that it replicates the circular structures for the months of the year observed in Engels et al. [2025] as well as the semantic clusters observed in Bricken et al. [2023].

In Figure 2 we show that under small latent sizes ($m = 200$) words are clustered by semantic category. This structure disappears as latent size increases ($m = 600$) but extends to larger latent sizes if we introduce weight decay Figure 2. This suggests that these kinds of structures are to be expected in the residual stream of real language models which encode many features in superposition and are trained using weight decay.

Similarly, in Figure 3 we show that the months of the year are correlated with a cyclic structure, such that taking the PCs of the binary word occurrences yields a circle and this structure is reflected in the weights of a ReLU AE trained on this data. These results suggest that this circular geometry may not be actively constructed by language models for a specific non-linear function like modular addition [Engels et al., 2025], but rather passively *inherited* from the statistical structure of the input data when subjected to dimensionality reduction.

5 Discussion

Summary of findings. In this work, we have highlighted the existence of two kinds of superposition, *linear* and *non-linear*. We argue that features in linear superposition inherit their structure from their covariance matrix, explaining previously observed, sematically rich structures in language models [Bricken et al., 2023, Engels et al., 2025]. We show this on synthetic data, where two distinct superposition regimes (linear and non-linear) arise (Section 3), as well as in realistic internet data where semantic clusters and circles also appear as a byproduct of linear superposition (Section 4).

Limitations and future work. This work provides a proof of existence for linear superposition in ReLU AEs and highlights how this can give rise to semantically rich structures akin to the ones observed in real language models. However, while the data used in this work is more realistic than that of previous studies on toy models, more work is required to understand under what circumstances linear superposition appears in real models. While our BOWS setup is designed to mimic the residual stream of a transformer, it does not model the ability of transformers to move features between token representations using the attention mechanism.

The BOWS framework, while simple, gives rise to many kinds of interesting behavior that this paper only begins to cover. Interesting avenues for future work include studying BOWS setups with untied encoder and decoder weight as well as using BOWS as a setup as an SAE evaluation setting in which we know the ground truth features which are encoded in superposition.

Acknowledgments

L. Prieto was supported by the UKRI Centre for Doctoral Training in Safe and Trusted AI [EP/S0233356/1]. T. Birdal and M. Barsbey acknowledge support from the Engineering and Physical Sciences Research Council [grant EP/X011364/1]. T. Birdal was supported by a UKRI Future Leaders Fellowship.[grant number MR/Y018818/1].

References

Carl Allen and Timothy Hospedales. Analogies explained: Towards understanding word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 223–231. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/allen19a.html.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90014-2. URL https://www.sciencedirect.com/science/article/pii/0893608089900142.

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

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

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. https://transformer-circuits.pub/2022/toy_model/index.html.

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*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Leo Gao, Tom Dupre 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*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=JYs1R9IMJr.

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

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/b78666971ceae55a8e87efb7cbfd9ad4-Paper.pdf.

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), 2025. ISSN 1099-4300. doi: 10.3390/e27040344. URL https://www.mdpi.com/1099-4300/27/4/344.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. URL http://arxiv.org/abs/1301.3781.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=UGpGkLzwpP.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162/.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

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.

Ozan Tuncer, Vitus J Leung, and Ayse K Coskun. Pacmap: Topology mapping of unstructured communication patterns onto non-contiguous allocations. In *Proceedings of the 29th ACM on International Conference on Supercomputing*, pages 37–46, 2015.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Appendix

A Related work

Superposition. Elhage et al. [2022] introduced superposition as an explanation for neuron polysemanticity. This view of DL models inspired SDL approaches like sparse autoencoders to decompose model activations into an overcomplete basis of linear features [Gurnee et al., 2023, Huben et al., 2024, Bricken et al., 2023]. This approach has successfully been scaled to frontier language models and multimodal models by Gao et al. [2025] and Templeton et al. [2024].

Feature geometry. Park et al. [2024] proposed a formalization of the LRH and proposed an inner product that preserves language structure. Park et al. [2025] studied how features with hierarchical relations are encoded in language models. Li et al. [2025] showed that language models represent integers in a helix structure to perform modular addition echoing the results from Nanda et al. [2023] on transformers trained for modular addition. Gurnee and Tegmark [2024] showed that longitude and latitude as well as a notion of time, are encoded as linear features in language models. The main results highlighted in this paper are circular structures formed by features and semantic feature clusters, described in Engels et al. [2025] and Bricken et al. [2023] respectively. These findings sparked a discussion around the potential limitations of SDL approaches and the LRH suggested by this non-linearly encoded semantic information [Sharkey et al., 2025].

Structure in word representations. Classic work on distributional semantics and word embeddings (e.g., Word2Vec [Mikolov et al., 2013], GloVe [Pennington et al., 2014]) demonstrated that training simple models on large text corpora leads to vector spaces where geometric relationships capture surprisingly sophisticated semantic and syntactic relationships. Levy and Goldberg [2014] showed that methods like Word2Vec with negative sampling implicitly factorize the Pointwise Mutual Information (PMI) matrix shifted by a constant, while others show connections to PCA or SVD on co-occurrence counts or PMI [Allen and Hospedales, 2019].

B Implementation details

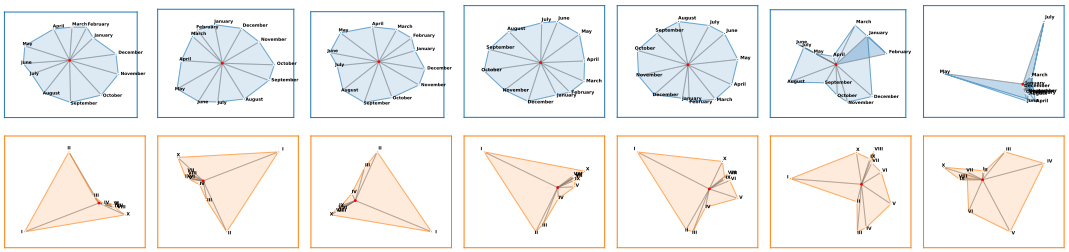
WikiText-BOWS:. All the models trained in the WikiText BOWS setup use a cosine annealing scheduler with a starting learning rate of $1e-3$ and are trained for 20 epochs with a batch size of 1024.

Synthetic "Months" dataset. Each document is a 12-bit vector $x \in \{0,1\}^{12}$ whose entries stand for the calendar months. One sample is generated as follows.

1. *Latent month angle.* Pick a discrete month $m \in \{0,\dots,11\}$ (uniformly or by cycling) and add Gaussian blur:
$$\theta = \frac{2\pi}{12}\, m \;+\; \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\theta^2).$$
2. *Embed on the unit circle.* $z = \big[\cos\theta,\ \sin\theta\big]^\top \in \mathbb{R}^2$.
3. *Project onto month directions.* Let
$$W = \Big[(\cos\tfrac{2\pi k}{12},\, \sin\tfrac{2\pi k}{12})\Big]_{k=0}^{11} \in \mathbb{R}^{12\times 2},$$
whose k-th row corresponds to month k. Compute log-odds $\ell_k = \beta\, W_k z + b$, where $b < 0$ fixes the global sparsity and $\beta > 0$ controls sharpness.
4. *Binary activations.* Draw the bits independently:
$$x_k \;\sim\; \text{Bernoulli}\big(\sigma(\ell_k)\big), \qquad \sigma(u) = \tfrac{1}{1+e^{-u}}, \qquad k = 1,\dots,12.$$

With σ_θ=0 and large β the code is nearly one-hot; decreasing β or increasing σ_θ mixes neighbouring months, producing a rank-2 correlation structure that is analytically tractable yet retains the extreme sparsity of real bag-of-words data.

UMAP plots and semantic clusters. For the UMAP plots in Figure 1 and Figure 3, the categories are created by using Gemini 2.5 Pro to split the top 4000 words into categories, with each category



$m = 200$ $m = 400$ $m = 600$ $m = 800$ $m = 1000$ $m = 2000$ $m = 4000$

Figure 5: Reconstructions at different latent-vector sizes. Top: "Months" dataset; middle: "Roman numerals"; bottom: corresponding latent size m.

inspected and refined by hand. The exact word to category mappings can be found in the code provided in the supplementary material. The UMAP plots are made with 15 neighbors, a min distance of 0.01, and a cosine metric.

C More detailed example of groups of feature geometry in BOWS

In the main paper we only show the feature structures some representative latent sizes due to space constraints. In Figure 5 we show the structures studied in Figure 6 for an extended range of latent sizes. Similarly, in Figure 6, we show UMAP plots for a more complete range of latent sizes for models with a context size of 1 record (top) and 50 records (bottom). We see that the larger context size introduces more correlations in the data, extending the prevalence of linear superposition, where-as semantic clusters disappear quickly with a context size of 1.

We also show a zoomed in version of one of the UMAP plots in Figure 4. This figure highlights the rich structure of the features beyond simple clustering of high-level classes. We see that words corresponding to sciences are clustered together, but within this high level cluster, sub-groups like words about medicine (top left), astronomy (lower left), chemistry (lower center) and biology (center) are also grouped in smaller clusters.

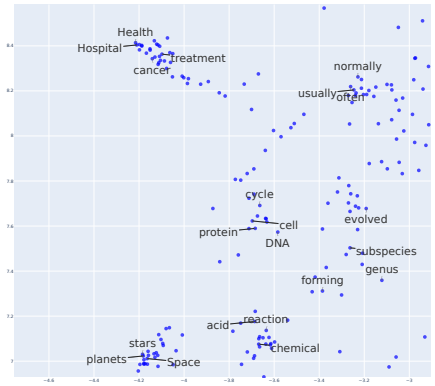


Figure 4: Zooming into the cluster for science features in the UMAP plot with a latent size of 200, we observe sub-sclusters within it. Medical features are in the top left while astronomy features are in the lower left and chemestry features are in the lower center.

C.1 Some examples beyond 2D

Beyond the 2D examples presented in the main paper, we include 2 examples showing that the days of the week and months of the year have structure beyond a 2D circle (Figure 8). This is clear in the case of the months where an ondulation in the third principal component is present beyond the 2D circular structure.

D Superposition on correlated and uncorrelated data

In Figure 7 we show the superposition patterns for the values of m missing in Figure 1, as well as a comparison with the weight patterns of AEs trained on i.i.d. data. In the i.i.d. case 12 features are drawn from a Bernoulli distribution with the same average frequencies as in the cyclic case. Figure 7 highlights that linear superposition only appears in the presence of feature correlations. In the i.i.d. case, AEs behave more like they do in Elhage et al. [2022], even when $d = 2$, the ReLU IID model uses the 2 dimensions to represent 4 features as antipodal pairs, with strong negative dot products within the pairs filtered out by the ReLU.

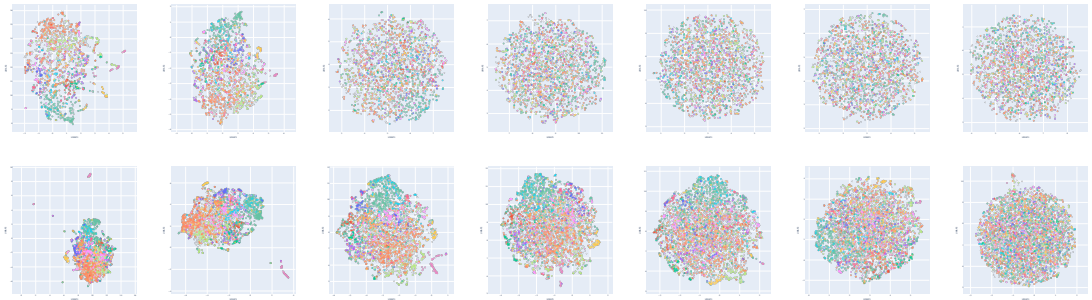


Figure 6: UMAP embeddings of features from AEs trained with context size of 1 record (top) and 50 records (bottom) across different latent sizes. The plot shows that semmantic structure remains for a larger fraction of context sizes when the contex window is larger, as it introduces additional, longer range correlations.

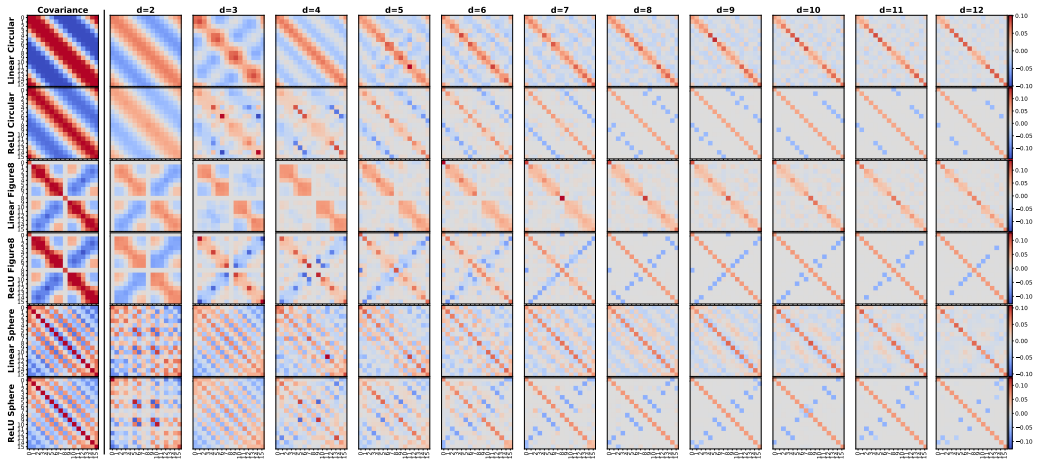
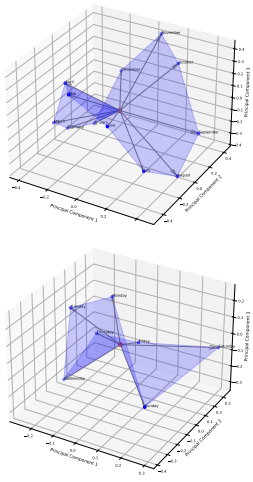


Figure 7: Extension of Figure 1 to include all values of m between 2 and 12, as well as a comparison with the weight patterns for AEs trained on data in the i.i.d. case.

E A tail of partially reconstructed features

An interesting observation is that some features seem to appear in the correct semantic cluster while the AE is only able to capture a small fraction of their variance (e.g. $R^2 < 0.3$). In Figure 9 we show that wether we filter for features with lower or higher reconstruction scores ($R^2 < 0.3$ or $R^2 > 0.3$) they still form semantic clusters. An explanation for why features with very small R^2 scores seem to have semantically meaningful representations is that, if a model is learning principal components of the data, it might project all the features, even uncommon ones, onto these principal components. This would mean that the representations of these features is only their projection onto some principal components, even if they only explain a small fraction of their variance, explaining the observed structure in poorly represented features.



(b) Days

Figure 8: 3-D PCA of the embeddings for the words and the days in a WikiText BOWS setup.

F Implications for the linear representation

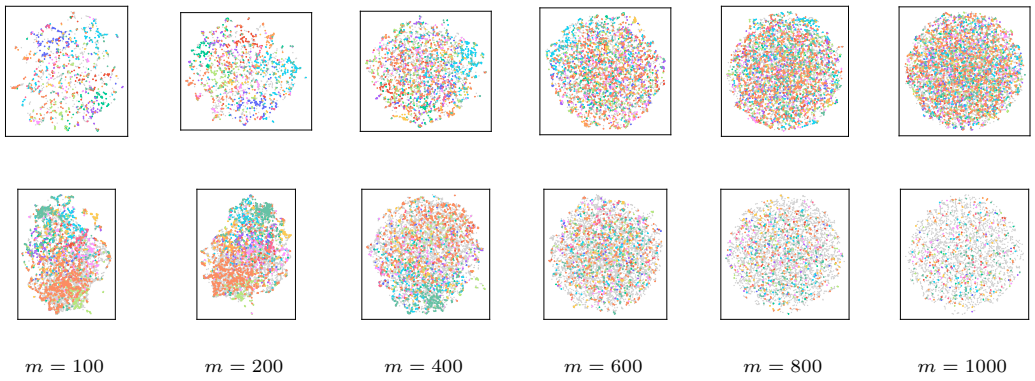


Figure 9: UMAP embeddings at different latent-vector sizes including only features with $R^2 < 0.3$ (top) and $R^2 > 0.3$ (bottom). Semantic clusters at different latent sizes are still observed in both, although this effect is combined with an increase in the number of features above the threshold in the lower one.

hypothesis

While the linear representation hypothesis (LRH) is one of the pillars of current mechanistic interpretability (MI) approaches. There is still no consensus on the correct formulation of this hypothesis. The LRH can be taken to mean that internal features of a model correspond to activations along one-dimensional directions in activation space [Engels et al., 2025]. However, the LRH can also be formalized around the mathematical notion of linearity meaning the representation of two features is the addition of their representations and scaling a feature corresponds to scaling its representation [Elhage et al., 2021].

While some works have suggested that observed feature geometry like the ordered circles formed by the months undermine the first definition [Engels et al., 2025, Sharkey et al., 2025], our results show that these structures can emerge from the compression and reconstruction of one-dimensionally linear features. This means that these structures do not necessarily undermine either formulation of the LRH.

An interesting line of research would be to explore if presence-coding features can have value-coding components. Findings like the fact that city representations in language models can be projected linearly onto a coordinates subspace [Gurnee and Tegmark, 2024], or that integers can be projected onto a helix subspace [Li et al., 2025] could be understood through this lens. In this view, city representations could have a coordinate-coding component and integers could have a size-coding component as well as sine and cosine coding components which combine to make a helix structure.

Overall, our findings show that rich feature geometry can be explained away by linear superposition recovering the structure inherent in the data, without appealing to non-linearly encoded information with a functional role in calculation. However, we believe the existence of value-coding features could be in conflict or an exception to features being mathematically linear.

G Other dimensionality reduction methods

To verify that the semantic clusters are not dependent on the choice of dimensionality reduction method, we include t-SNE [van der Maaten and Hinton, 2008] and PaCMAP [Tuncer et al., 2015] as two alternatives in fig. 10.

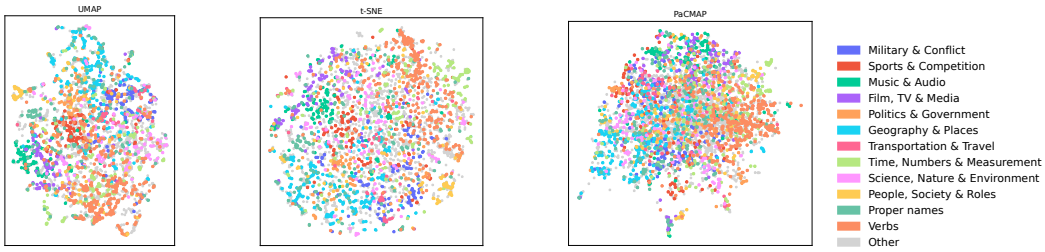


Figure 10: We show the latent representations of the top 4000 most frequent words using UMAP (left) t-SNE (middle) and PaCMAP (right) to highlight that these semantic clustering results are not dependent on the choice of dimensionality reduction technique.