




"Linguistic Universals": Emergent Shared Features in Independent Monolingual Language Models via Sparse Autoencoders

Ej Zhou*  & Suchir Salhan 

 Language Technology Lab, University of Cambridge

 Department of Computer Science & Technology, University of Cambridge

1 Introduction

Whether certain structural patterns are shared across all natural languages, despite surface-level differences, has long been a topic of debate in linguistics. In Natural Language Processing, studies have shown that multilingual language models possess semantically aligned capabilities across languages even without explicit parallel supervision (Pires et al., 2019; Conneau et al., 2020; Tang et al., 2024). This suggests that machine-learned representations capture crosslinguistic regularities, but it should be noted that this alignment is aided by shared vocabularies and parameters (Dufter and Schütze, 2020; Philippy et al., 2023).

A more fundamental question remains: can independently trained monolingual LMs – which share no parameters nor vocabulary – nonetheless converge on analogous high-level features? If so, this would suggest that certain structural principles of language emerge robustly in machine learning, even when models are trained in isolation. Goldfish (Chang et al., 2024) provides us with a suite of monolingual GPT-style models covering 350 languages. These models have identical architectures and training budgets but were each trained with strictly monolingual corpora. They thus form a controlled testbed for crosslinguistic comparison.

A key technical challenge is how to identify and compare high-level features across different models. To overcome this, we adopt sparse autoencoders (SAEs) as an analysis tool. Recent work (Cunningham et al., 2023) showed that training a single-layer SAE on a model’s activations yields a set of sparsely activating features that are far more interpretable and monosemantic than the original neuron basis. In essence, the SAE “discovers” a dictionary of latent feature directions in activation space, each corresponding to a distinct concept

or pattern in the data. Brinkmann et al. (2025) demonstrated that SAEs trained on multilingual LLMs uncover both monolingual and multilingual features. Notably, Lan et al. (2025) recently employed SAEs to compare features across different English LLMs. They hypothesized that the spaces spanned by SAE features are similar, such that one SAE space is similar to another SAE space under rotation-invariant transformations, and found high similarities for SAE feature spaces across various LLMs, providing evidence for feature space universality. We build on this approach in a novel crosslinguistic setting. Our research questions are framed as follows:

- RQ1:** Can SAE features trained on independently trained monolingual LMs be matched across languages? After matching, do they show non-trivial (above baseline) convergence (i.e., have higher alignment score)?
- RQ2:** At which model depths (layers) is feature alignment strongest across languages?
- RQ3:** Does the degree of alignment correlate systematically with linguistic relatedness (e.g., typological or genealogical distance)?
- RQ4:** Are there features that emerge universally across languages, and can they be interpreted (e.g., punctuation, numerals, structural delimiters)? How prevalent are such features?

2 Methodology

2.1 SAE Training

For each monolingual model, we collect hidden activations from each layer using held-out text sampled from the same monolingual training corpus used in Goldfish (5MB–1GB per language, depending on availability). Given these activations, we train an SAE to learn a set of latent features that can reconstruct the activations. Each SAE

*Corresponding Authors: yz926@cam.ac.uk, sas245@cam.ac.uk

is a one-hidden-layer autoencoder with tied encoder–decoder weights, a linear hidden layer, and an ℓ_1 sparsity penalty to encourage most feature units to remain off for any given input. We train separate SAEs for each language model’s each layers.

2.2 SAE Feature Activations (Data Matrix)

For each language ℓ and layer h , we construct an activation matrix $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{N \times K}$ by feeding N parallel sentences from FLORES-200 (NLLB Team, 2022) through the monolingual model and recording the activations of its K SAE features (z-scored per feature across sentences).

2.3 Pairwise Feature Matching

Given two languages (ℓ_1, ℓ_2) at layer h , we compute the $K \times K$ correlation matrix $\mathbf{C}^{(\ell_1, \ell_2, h)}$ with entries $C_{ij} = \text{corr}(\mathbf{A}_i^{(\ell_1, h)}, \mathbf{A}_j^{(\ell_2, h)})$ (Pearson over the shared FLORES sentences). We obtain a one-to-one alignment via maximum-weight bipartite matching (Hungarian algorithm) on $\mathbf{C}^{(\ell_1, \ell_2, h)}$.

2.4 Pairwise Alignment Score

For each pair (ℓ_1, ℓ_2, h) , the alignment score is the mean correlation of matched pairs:

$$\text{Align}(\ell_1, \ell_2, h) = \frac{1}{K} \sum_{(i,j) \in \mathcal{M}^{(\ell_1, \ell_2, h)}} C_{ij}.$$

We visualize the matrix of $\text{Align}(\ell_1, \ell_2, h)$ across all language pairs as a heat map.

3 Analysis

3.1 Alignment Against Baselines

To ensure the alignment is non-trivial, we compare against: (i) **random feature assignment**—shuffle columns of $\mathbf{A}^{(\ell_2, h)}$ before matching, (ii) **row-shuffled sentences**—independently permute rows of $\mathbf{A}^{(\ell_2, h)}$ (breaks sentence-level correspondence), and (iii) **within-model shuffle**—match ℓ_1 to a copy of itself with feature order shuffled. We report Δ over baseline (absolute and percentage), with 95% CIs from bootstrap over sentences.

3.2 Layer-wise Analysis

We aggregate $\text{Align}(\ell_1, \ell_2, h)$ over language pairs for each layer h to obtain layer-wise trends. We test for a peak layer via a mixed-effects model with random intercepts for language pairs and fixed effect for layer, or via paired non-parametric tests across layers.

3.3 Language-Distance Analysis

Given the Pairwise Alignment Score, a natural question—and our hypothesis—is whether more closely related (genealogically or typologically) languages share more emergent features, i.e., exhibit higher pairwise alignment score. We correlate alignment strength with linguistic distance. For each pair (ℓ_1, ℓ_2) we compute: (i) genealogical family match (binary), (ii) typological distance (e.g., URIEL/WALS features), and (iii) script match (binary). We fit $\text{Align}(\ell_1, \ell_2, h) \sim \text{distance metrics} + \text{layer}$ and report standardized coefficients. We also stratify heat maps by family/script to visualize systematic variation.

4 Universal Features

After aligning features between each pair of languages, we ask if these features are universal across all languages.

Definition. A feature cluster is *universal* at layer h if it contains aligned features from at least $p\%$ of languages (we choose $p \in \{50, 75, 90\}$).

Construction. We build a graph whose nodes are (language, feature) and whose edges connect matched pairs from Section 2.3 (weight = C_{ij}). Connected components (or communities via Louvain) define crosslinguistic clusters. For each cluster we report: coverage (fraction of languages present), mean within-cluster correlation, and stability across bootstrap resamples. Preliminary expectations are that a non-trivial fraction of the learned features – especially those capturing very general patterns – will be universal. For instance, we anticipate discovering features related to punctuation, numerals, and structural delimiters that appear in every model.

Interpretability. For each universal cluster, we list top activating n-grams/tokens per language and show cross-language trigger sets (digits, punctuation, brackets, etc.). We include exemplar sentences and activation traces for qualitative validation. We hope that uncovering such crosslinguistic universal features will shed light on whether machine-learned representations mirror long-standing hypotheses in linguistic theory, and may even provide a complementary empirical perspective to the study of linguistic universals in human languages.

References

- Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. [Large language models share representations of latent grammatical concepts across typologically diverse languages](#). *Preprint*, arXiv:2501.06346.
- Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024. [Goldfish: Monolingual language models for 350 languages](#). *Preprint*, arXiv:2408.10441.
- Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Emerging cross-lingual structure in pretrained language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. [Sparse autoencoders find highly interpretable features in language models](#). *Preprint*, arXiv:2309.08600.
- Philipp Dufter and Hinrich Schütze. 2020. [Identifying elements essential for BERT’s multilinguality](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.
- Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2025. [Quantifying feature space universality across large language models via sparse autoencoders](#). *Preprint*, arXiv:2410.06981.
- James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Searley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.
- Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. [Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.
- Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. [How multilingual is multilingual BERT?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
- Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. [Language-specific neurons: The key to multilingual capabilities in large language models](#). *Preprint*, arXiv:2402.16438.