UNIVERSITY OF CAMBRIDGE

Ej Zhou, Suchir Salhan

Read our Extended Abstract! →

SCAN ME

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

Do independently trained monolingual small LMs across 350 languages learn similar internal features?

Although multilingual LLMs show crosslinguistic alignment, their shared tokenizers and parameters make it impossible to tell whether this similarity is truly emergent.

Goldfish (Chang et al., 2024), a family of monolingual small language models for 350 languages, provides a clean testbed needed to ask whether similar 'high-level' features emerge *even without any shared structure*.

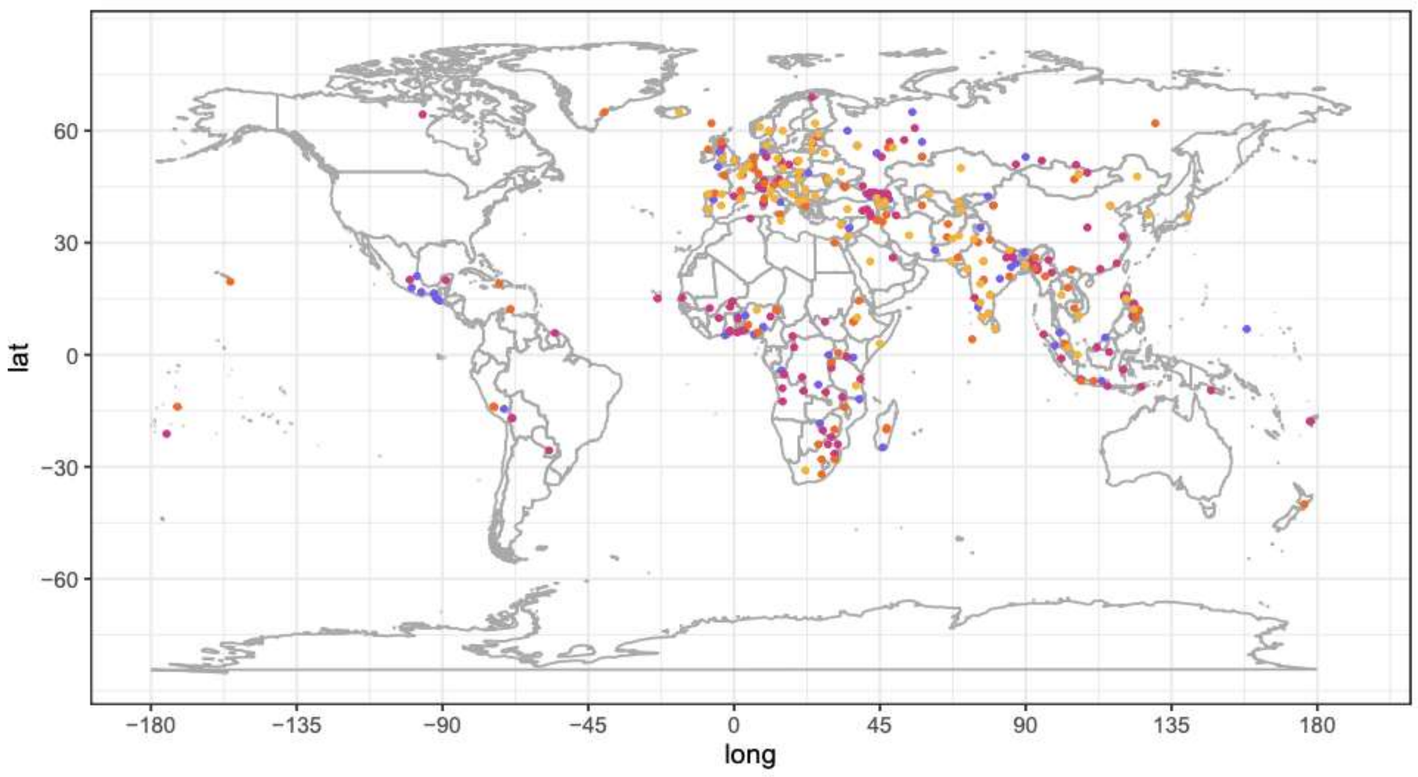
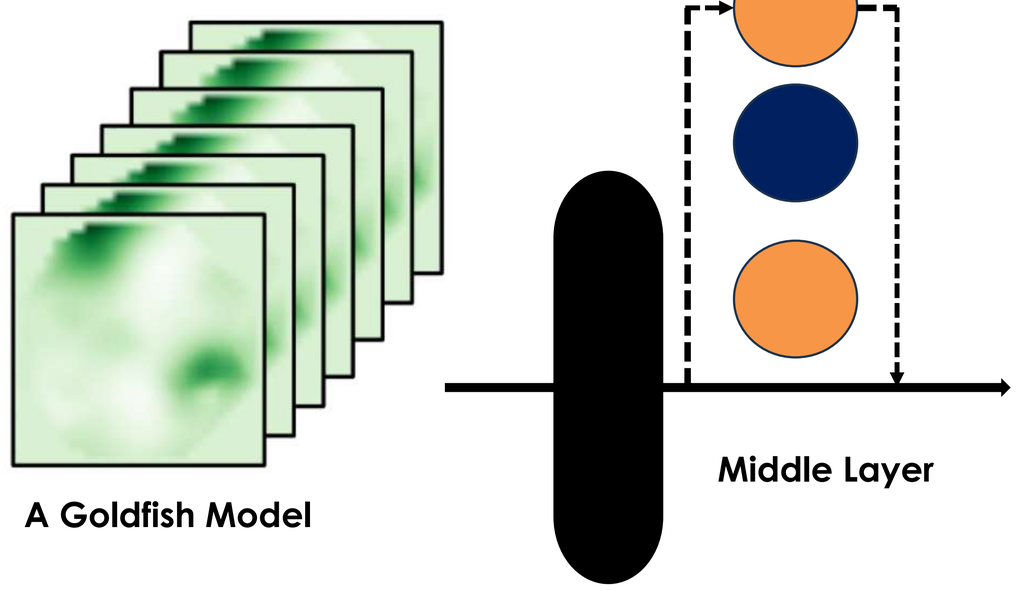


Figure: *Distribution of languages in the Goldfish Language Model Family. From Chang et al. (2024)*

= Shared, Multilingual Feature

= Language-Specific Monolingual Feature



Our methodology use Sparse Autoencoders (SAEs) to investigate shared emergent representations across small monolingual LMs.

On Thursday, scientists from Cambridge University …	en
У четвер вчені з Кембриджського університету …	ukr
Jeudi, des scientifiques de l'université de Cambridge ont déclaré …	fr

Sentences from **FLORES-200** (a parallel corpus) are passed through monolingual Goldfish models.

Method: For each language pair and each layer, we measure how similarly each *direction* in the model's hidden space reacts to the same FLORES sentences. We then find the best one-to-one matching between directions in the two languages—the matching that gives the highest total similarity. The average similarity of these matched pairs is our **alignment score**, which we compare to chance-level baselines created by shuffling. Finally, we look for directions that consistently match across many languages to identify potential universal clusters.

Takeaways: We ask whether features align above chance, which layers align most strongly, whether alignment reflects linguistic relatedness, and whether universal patterns (e.g., digits, punctuation, structural markers) appear. Strong **crosslinguistic alignment** would suggest that **monolingual small LMs independently discover shared features**—even without any shared vocabulary, parameters, or data.

{yz926,sas245}@cam.ac.uk