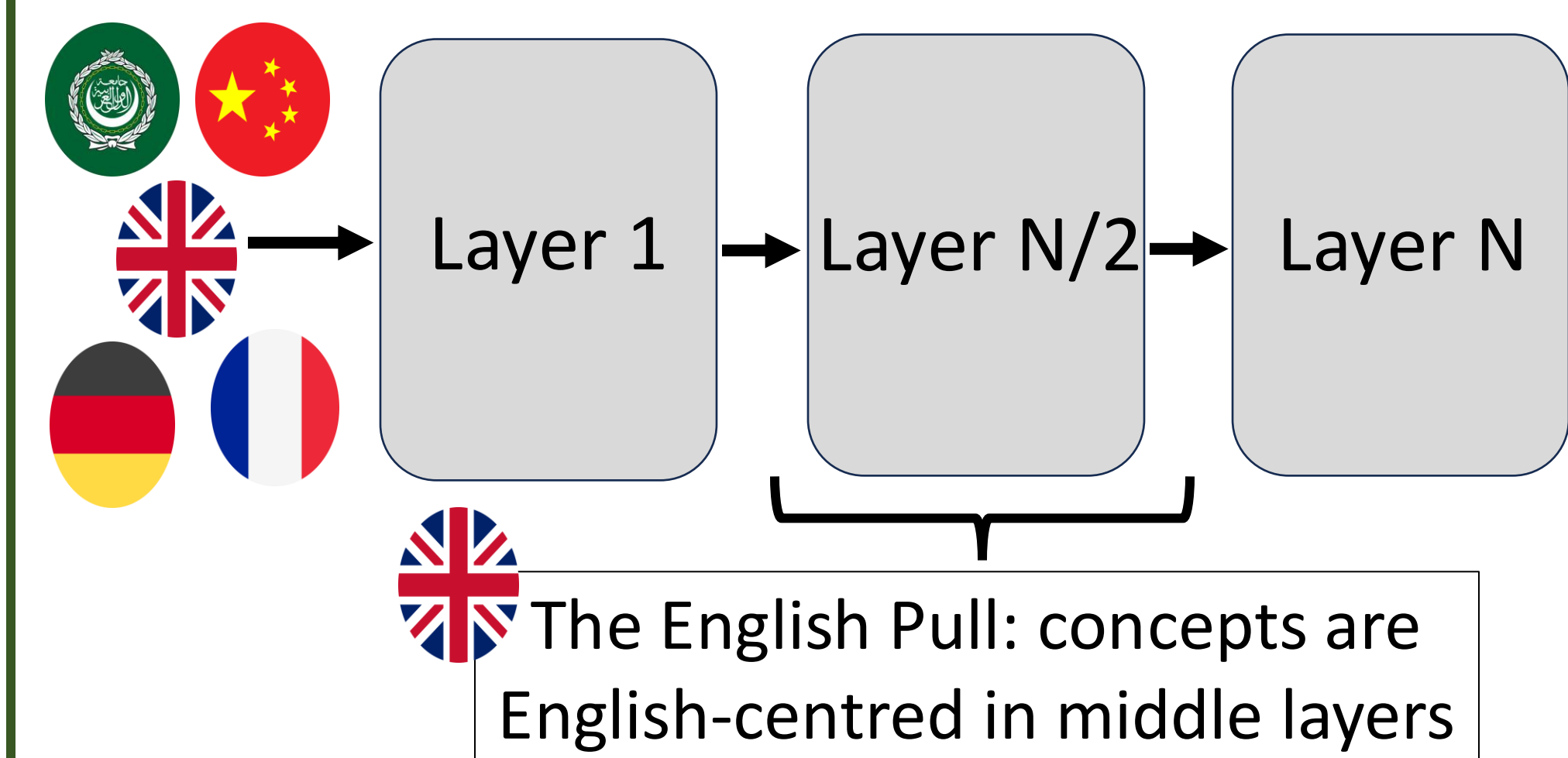


Are middle layers in LLMs English-centric?



- Are models truly **English-centric**, as prior work suggests?
- Could this **explain** the underperformance of languages like Arabic?
- Is it possible to **boost** low-resource languages without finetuning?

The Challenge: Fairness and Access in Multilingual LLMs

- LLMs power **global applications**, but research mostly focuses on English.
- Many languages, like **Arabic**, face lower accuracy and bias, risking **fairness and access**.

Recovering Activation Space Multilinguality with SAEs

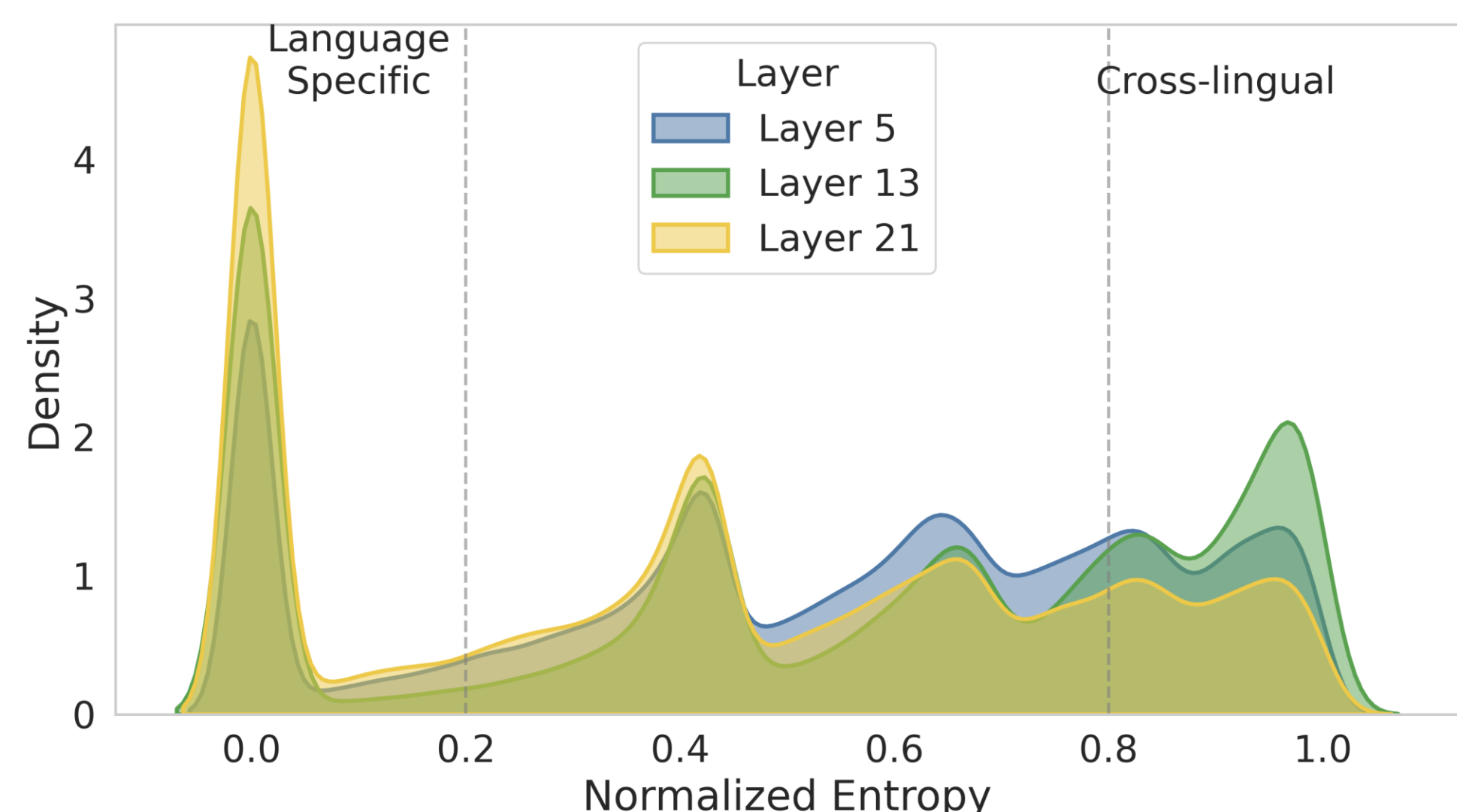
- We train **Multilingual Sparse Autoencoders (SAEs)** on the activations of layers 5, 13 and 21 of Gemma2-2B.
- We measure the **language entropy** of each SAE feature to analyze language mixing at each layer.

$$H(f) = - \sum_{l=1}^L p_l(f) \log p_l(f),$$

Where:

$$p_l(f) = \frac{A_l(f)}{\sum_{l'} A_{l'}(f)}, \quad A_l(f) = \sum_t a_{t,f}^{(l)},$$

And L is the number of languages and $a_{t,f}^{(l)}$ is the activation of feature at token t of language l .



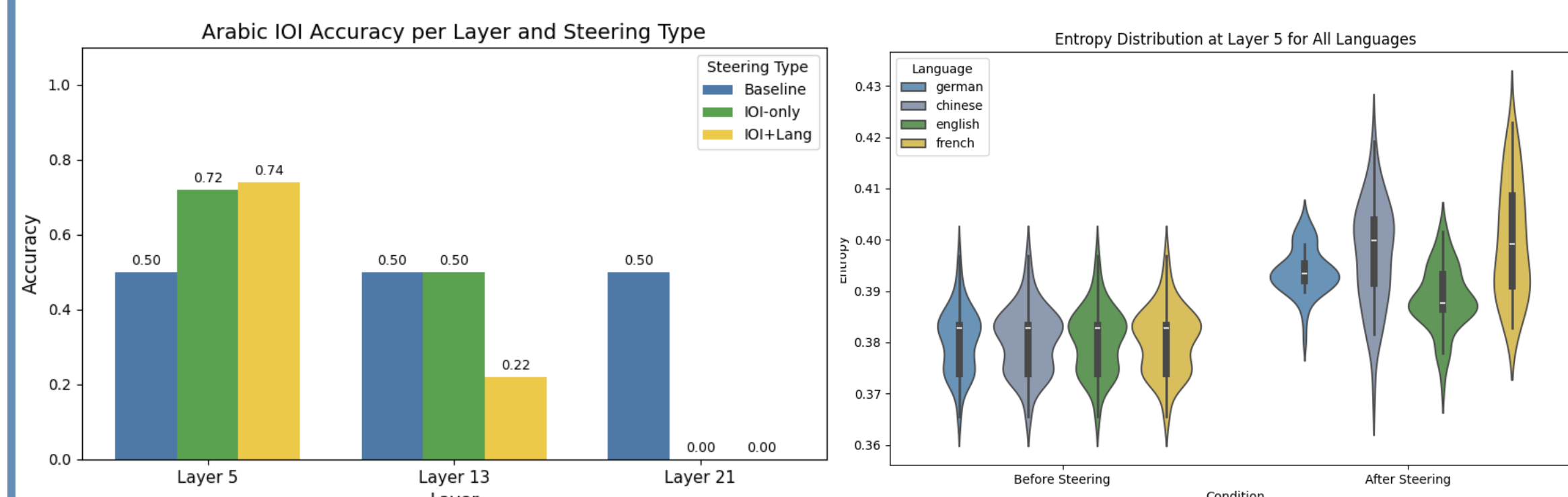
Early layers are the most multilingual, and **middle layers**, contrary to prior work, also retain many **multilingual** features and therefore **not English-centric**.

Steering Shifts Arabic Representations into Multilingual Region

- We use the IOI task as a **case study** to improve Arabic performance (starting at **50%**).
- By steering Arabic activations using IOI activations from another language (e.g., French), we boost performance.
- To do this, we isolate an **IOI-specific steering vector** by:

$$v_{\text{raw}} = v_{\text{lang}} + v_{\text{IOI}} + \epsilon$$

$$v_{\text{steer}} = v_{\text{raw}} - v_{\text{lang}}$$



Looking Ahead

- How does the model use multilingual features internally?
- How do circuits vary with training data and input language?
- Is there cross-lingual generalization from high-resource to low-resource languages? If not, can it be induced as a general capability rather than just for specific tasks?

