A Mechanistic View of Intrinsic Multilingualism in Large Language Models

Lingjie Chen* **Fukang Zhu*** **Ningyu Xu** **Xuyang Ge** **Junxuan Wang**
Zhengfu He **Xipeng Qiu†**
OpenMOSS Team, School of Computer Science, Fudan University
{fkzhu21,ljchen21,nyxu22,junxuanwang21}@m.fudan.edu.cn
{zfhe19,xyge20,xpqiu}@fudan.edu.cn

Abstract

It has been conjectured that multilingual models process information in low-resource languages in an English state of mind since English takes up a large proportion of the training corpus. Recent progress in language model multilingualism provides more evidence for this hypothesis. We ask a further question: **What if the model is trained on more than one high-resource language?** By studying language models trained mostly on Chinese and English with an interpretability technique called Sparse Autoencoders, we manage to identify a three-stage process of how models think in these two languages. The model first "detokenize" inputs and both languages are aligned. The representation of these two languages then diverges and processed independently in a "conceptual stage" and is aligned again in the "retokenization stage". We name this the *Intrinsic Multilingualism*. We empirically test our hypothesis by intervening the model internal with Sparse Autoencoders trained on another language and find that the "conceptual stage" is crucial for the model to think in different languages. We also showcase a number of features detecting intriguing lingual and cultural bias in Chinese and English.

1 Introduction

Understanding the multilingual ability of language models is an important research problem. As large language models (LLMs) continue to advance and exhibit unprecedented societal impact, this becomes a more urgent issue with regard to both language model interpretability and AI fairness.

A number of language models are found to be able to achieve considerable performance in languages with a small portion of training data (Devlin et al., 2019; Conneau et al., 2020; Wendler et al., 2024). Existing literature on this problem observes the phenomenon of "pivot language" in these English-dominated LLMs, where the model internally translates the prompt to its pivot language and then translates the processed results back for generation (Shi et al., 2022; Ahuja et al., 2023; Wendler et al., 2024; Zhao et al., 2024; Tang et al., 2024). In this work, we name this type of lingual ability inherited from another language after *transferred multilingualism*.

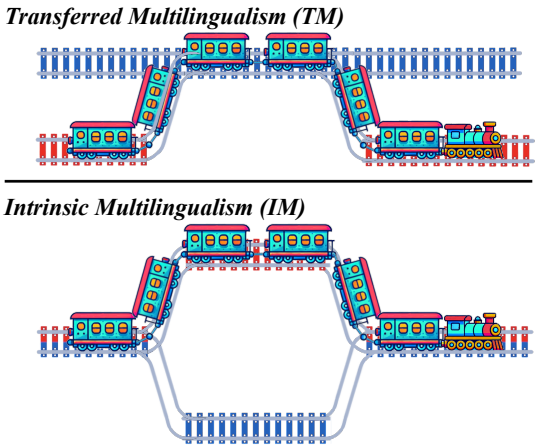

Figure 1: Two types of multilingual abilities. When a model is trained on two high-resource languages, it exhibits Intrinsic Multilingualism.

In recent years, the rise of a family of Chinese LLMs (Bai et al., 2023; Yang et al., 2023; Sun et al., 2024) provides researchers with a new possible testbed for multilingual ability. It raises a natural research question: what will be the mechanistic story of LLM multilingualism if another language appears as much as English? We find that **English and Chinese are found to be of comparable status and exhibit a clear three-phase mechanistic structure.** We call it *intrinsic multilingualism*.

We utilize Sparse Autoencoders (SAEs) to analyze an *intrinsic* zh-en Qwen-1.5-1.8B (Bai et al.,

*Equal Contribution.
†Corresponding author.

2023) and a *transferred* Phi-2 (Gunasekar et al., 2023) for their ability on (Simplified) Chinese and English. We conduct a series of exploration of the inner state of multilingual LLMs from both macroscopic and microscopic lens. These approaches lead to the same conclusion that **Transferred and Intrinsic multilingual ability are mechanistically distinct.**

We make an analogy of these two modes of multilingualism to rail tracks in Figure 1. When both languages are trained on equally, their inner processing features three stages: **Detokenization**, **Conceptual Stage**, and **Retokenization**. These two languages align at the first and the last stage and diverge at the Conceptual Stage. When English dominates, our SAE analysis leads to a conlusion agreeing with existing findings i.e. the model actually thinks in English (Wendler et al., 2024).

The main contributions of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to investigate the inner mechanism of LLMs dominated by more than one language.
- Through the analysis of SAEs' features, our experiments make an extension to existing findings in multilingualism.
- We propose a systematic method using SAE to explore LLMs' abilities in different distributions. We believe this can be generalized to more scenarios e.g. multimodal models.

2 Conceptual and Empirical Preliminaries

2.1 Dataset

Language Choice Our goal is to probe the core mechanisms of Intrinsic Multilingualism and Transferred Multilingualism, which requires comparing high-resource languages and high- and low-resource languages. Additionally, we need to select languages from different language families to avoid potential mutual interference, as mentioned in (Wendler et al., 2024). We thus choose English and Chinese as our primary experimental languages due to their significant differences and the availability of ZH/EN-based LLMs. We also included Arabic as a comparative low-resource language for Intrinsic Multilingualism experiments.

Language Datasets For our experiments, we selected ChineseWebText (Chen et al., 2023), Pile (Gao et al., 2020), and Arabic-words-dataset (Aloui et al., 2024) as the source datasets for Chinese, English, and Arabic, respectively.

For efficiency, we selected 1 billion tokens from each dataset as our ZH, EN, and ARA datasets. Additionally, we created the MIX dataset by selecting 1 billion tokens, evenly distributed between Chinese and English data. Further information about datasets can be found in Appendix A

2.2 Language Models

We choose Qwen-1.5-1.8B (Bai et al., 2023) as a testbed of Intrinsic Multilingualism. Its training dataset is multilingual, including both Chinese and English as high-resource languages. This model has 24 layers, 16 attention heads, an embedding dimension of 2048, and a vocabulary size of 152,000 tokens.

For our analysis of Transferred Multilingualism, we selected Phi-2 (Gunasekar et al., 2023) because its training corpus is primarily English, making English the only high-resource language. Phi-2 has 2.7 billion parameters and 32 layers with an embedding dimension of 2560.

2.3 Sparse Autoencoders

Though a small fraction of MLP neurons exhibit monosemanticity i.e. firing in a human-understandable pattern (Tang et al., 2024), recent progress in mechanistic interpretability shows that they may not be the right primitives to work with. This is due to both the neuron basis being privileged (Elhage et al., 2023) and the superposition hypothesis (Elhage et al., 2022b).

To this end, we follow Bricken et al. (2023) to decompose model activation on a more interpretable, overcomplete basis with Sparse Autoencoders (SAEs). Existing literature suggests SAEs are able to extract a lot of interpretable features from models across model sizes (Templeton et al., 2024) and tasks (He et al., 2024; Gandelsman et al., 2024).

Our SAEs have only one hidden dimension larger than the input dimension (i.e. $F > D$), with the training objective of reconstructing any given model activation and an L1 penalty on its hidden layer to incentivize sparsity. An SAE can

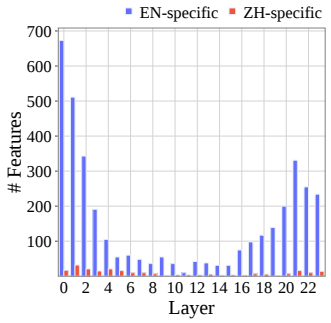

Figure 2: English / Chinese-specific features in each layer.

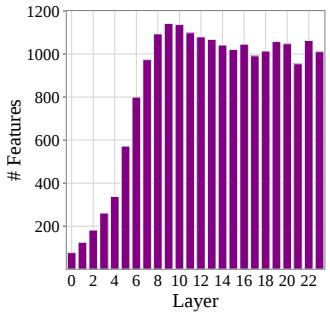

Figure 3: Features firing in both languages.

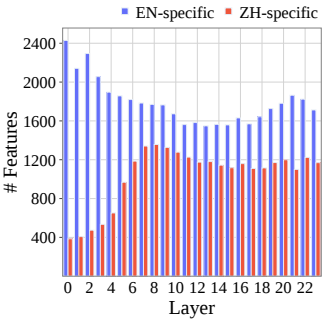

Figure 4: Features firing at English / Chinese tokens.(without specificity)

be formulated as follows:

$$f_i(\mathbf{x}) = \text{ReLU}(\mathbf{W}^{enc}_{i,\cdot} \cdot \mathbf{x} + b^{enc}_i),$$
$$\hat{x} = \sum_{i=1}^{F} f_i(\mathbf{x}) \cdot \mathbf{W}^{dec}_{\cdot,i},$$

Where $\mathbf{x} \in \mathbb{R}^D$ is the hidden activation decomposed with F features. $\mathbf{W}^{enc} \in \mathbb{R}^{F \times D}$ and $\mathbf{W}^{dec} \in \mathbb{R}^{D \times F}$ are the encoder and decoder of an SAE. $\mathbf{x}$ is linearly mapped into the feature space by the encoder and a bias b^{enc}, followed by a ReLU to ensure $f(\mathbf{x})$ (i.e. feature activations) being non-negative.

SAEs are trained to reconstruct the original activation with a linear combination of the decoder columns (i.e. features), determined by $f(\mathbf{x})$. We also set an L1 sparsity constraint on $f(\mathbf{x})$ to obtain a sparse coding of each activation.

Concretely, we train SAEs on language models in Chinese (ZH), English (EN), and a mix of both (MIX) in the residual stream after each Transformer block. We refer readers for more training details of our SAEs to Appendix B.

Intuitively, SAEs trained solely on one language only extract the model's features for this specific language. *We seek to study the commonality and divergence of these two families of SAEs to mechanistically understand LLM multilingualism.*

2.4 Transferred Multilingualism

How does it work internally when processing different languages in a model mainly trained on English? We validate existing findings (Shi et al., 2022; Wendler et al., 2024; Tang et al., 2024) through the SAE lens by analyzing features in Phi-2 MIX SAE. Our results suggest furthur evidence that **English-dominated LLMs process other languages in English.**

We are interested in the number of SAE features that activate in each language. Following Tang et al. (2024), we use the LAPE (Language Activation Probability Entropy) metric for language-specific features. By calculating the probability of a feature being activated across different languages, we can determine the entropy of the features, indicating their language specificity. In this experiment, we use 512 data samples from ChineseWebText and Pile to test feature's specificity.

The main takeaways of our validating experiment are summarized as follows:

- **Language-specific features appear at early and late layers (Figure 2):** The number of both English-specific and Chinese-specific features (i.e. only firing in English / Chinese corpus) exhibits a U-shaped trend across the 24 layers. This is consistent with neuron-level findings (Tang et al., 2024).

- **Multilingual features emerge at early-middle layers (Figure 3:** The number of multilingual features increases from the early layers to the middle layers, indicating where Transferred Multilingualism occurs and explaining the decrease in language-specific features.

- **English related feature number far exceeds Chinese's 4)** The number of features firing at EN tokens is much higher in all layers. Moreover, the ZH bar's shape is similar in Figure 3 and Figure 4), indicating that features firing at ZH tokens also fire at EN tokens, echoing with pivot language in TM.

One advantage of understanding Transferred Multilingualism is that the SAE feature basis of-

fers a better interpretability primitive. One problem with the neuron approach is that here is no guarantee that these neurons are interpretable (Elhage et al., 2022b, 2023). This poses a concern on whether the neuron approach may provide an epistemic basis for analysis. Wendler et al. (2024) utilize a tool called the logit lens, by directly sending the intermediate residual stream activation to the model unembedding. This method may provide insights into the model's internal workings. However, it usually does not work at early layers since they are too far away from the unembedding and may lead to deceptive conclusions.

3 Understanding Intrinsic Multilingualism

Different from Transferred Multilingualism, when another high-resource language comes into play, the inner mechanism of multilingualism will change thoroughly.

3.1 Conceptual Model

We first forward our conjectural conceptual model of Intrinsic Multilingualism. Intrinsic Multilingualism mainly consists of three stages:

1. **Detokenization (Layer 1-5)**: English and Chinese features are aligned, and each language's feature contains the other's linguistic information.

2. **Conceptual Stage (Layer 6-15)**: English and Chinese features are separated, and the model 'thinks' in different languages.

3. **Retokenization (Layer 16-24)**: English and Chinese features align again, allowing mutual representation.

This model is inspired by existing findings by Elhage et al. (2022a) and Ge et al. (2024c), where they find that neurons/features are at low-level in early layers, become more abstract in middle layers and are related to next-token prediction at late layers. We inherit this hypothesis and provide further supporting evidence of language models' internal hierarchical structure.

3.2 Experiments

Macro Analysis: SAE substitution SAE decomposes the model's features by learning to reconstruct the model's activation values. We use the reconstruction error between the SAE's reconstructed activation and the original activation values to quantitatively evaluate how well an SAE captures the model's features. This experiment is called **SAE substitution**.

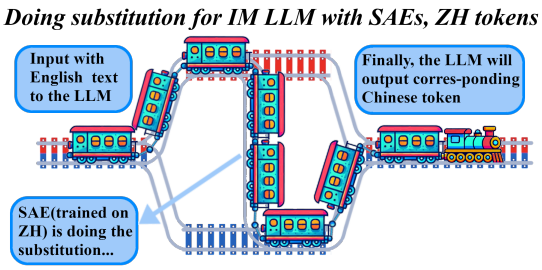

Figure 5: Concept diagram for doing SAE substitution on *IM* model.

Assuming this premise, we perform SAE substitution with different input data. For example, if the input data is in English and we conduct ZH SAE substitution, the reconstruction cross-entropy loss (ce loss) indicates how representative the model's Chinese-induced features are given English input, as shown in Figure 5.

We analyze Qwen-1.5's SAE substitution results. Figure 6's green bar shows the performance of the MIX SAE. We can see that it achieves near-perfect performance in all input settings, validating our premise that it can demonstrate the multilingual model's full competence. Additionally, the results of EN and ZH substitution in mutual languages both form U-shaped curves. This symmetry reveals that Qwen-1.5's abilities in English and Chinese are more balanced and independent. The consistent U-shape in EN and ZH SAE substitution indicates that the features of one language contain information about the other language in the early and final layers. However, the middle layers' features are more separated from the other languages, leading to comparatively poorer reconstruction results. To demonstrate the unique pattern shown by Intrinsic Multilingualism, we also carry out SAE substitution on the Arabic dataset, the result and analysis are shown in Appendix C.

This experiment provides initial proof of Intrinsic Multilingualism's inner mechanism, implying the potential drifting in the middle layers.

Micro Analysis: Activation Similarity To gain a more detailed understanding of the model's inner mechanism and the relationship between ZH and EN features, we will visualize the evolution of activations after substitution.

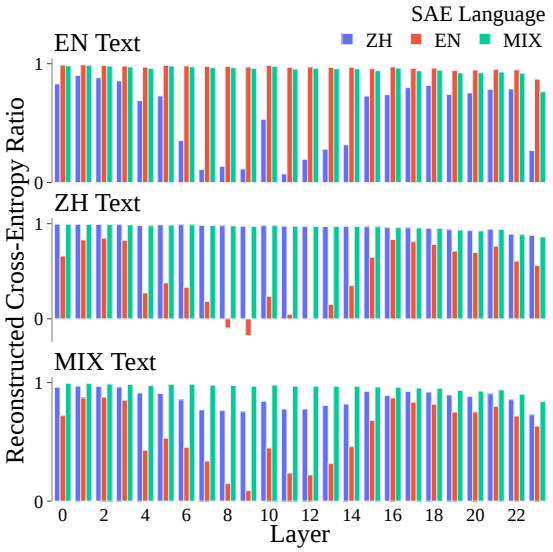

Figure 6: Qwen1.5's SAE substitution result. The formula of "Reconstructed Cross-Entropy Ratio" can be found in Appendix B

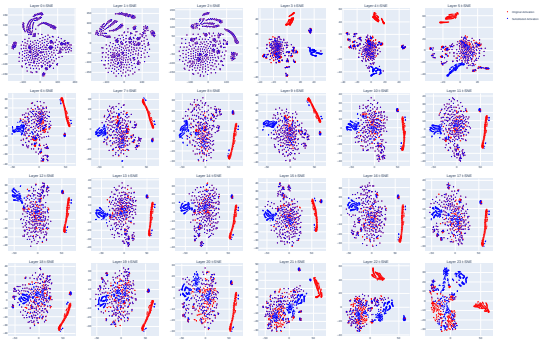

Figure 7: All layers' activation pattern after substituting Layer 3's activation value with ZH SAE's output. Red dots denote the original layers' activation, while blue dots denote the substituted activation.

When the input is English and we substitute the k^{th} activation with the ZH SAE's reconstructed activation value, we record all the following activations and compare them with the original activations. Figure 7 and Figure 8 show the results of the 3^{rd} and 8^{th} layers being substituted.

The discrepancy between these two figures lies in whether the substituted activation gradually diverges from the original activation.

From Figure 9 we can understand the relationship between the substitution's position and the final layer's activation's clustering result. We gain two key insights from this experimentation: (i) The divergence of activation starts from the 18^{th} to 20^{th} layers. (ii) If the substitution occurs in the middle layers (5^{th} to 14^{th}), the divergence will manifest

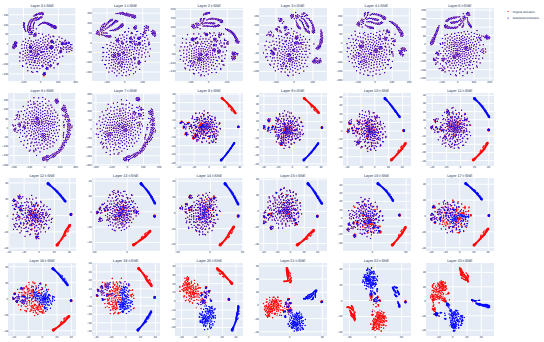

Figure 8: All layers' activation pattern after substituting Layer 3's activation value with ZH SAE's output. Red dots denote the original layers' activation, while blue dots denote the substituted activation.

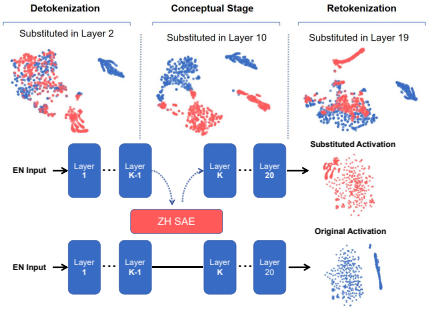

Figure 9: t-SNE results of different layers' substitution, indicating different stages of model

in the final layers. (iii) When the substitution occurs in the early or final layers, the original and reconstructed activations are very similar.

The full result of this experiment is presented in Appendix D. This result echoes the U-shape in macro analysis in that divergence caused by substitution is equivalent to the low ce score in SAE substitution.

Furthermore, this experiment validates our conceptual model, proving that the features in the early and final layers are mutually aligned while separated in the middle layers.

Finally, we will directly demonstrate the existence of the main **Conceptual Stage** within Intrinsic Multilingualism.

Causal Analysis: Substitution Inference The previous analysis focused on the intermediate activations within the model. Now, we aim to understand how different features affect the model's downstream effects. We use a substitution inference experiment. As shown in Figure 10, we substitute the intermediate activation value of a certain layer with the reconstructed activation value from

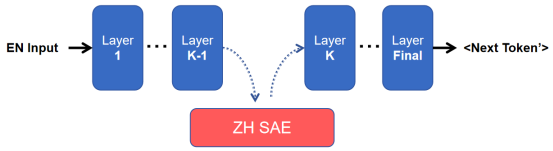

Figure 10: Demonstration of Causal Analysis. The setting is using EN input and substitute intermediate activation value with ZH SAE.

SAE. Then, we check the difference in the <next token> prediction of the model before and after substitution.

As shown in Table 1, when the substitution happens in early or final layers, the <next token> prediction remains English words and is close to the original prediction. However, if we substitute the middle layers with reconstructed activation, the final prediction will be in Chinese. This Chinese prediction can either fit the context or be a translation of the original next-token prediction. These results highlight the role of the model's middle layers.

Besides the qualitative results, we conduct a quantitative experiment to verify the previous conclusion about the uniqueness of middle stages using cloze tests (Nostalgebraist, 2020). We create a small-batch dataset of 100 samples in both languages and calculate whether the model's prediction after SAE substitution matches the correct answer or its translative pairs in ZH/EN.

Figure 11 shows the result of using EN input and ZH SAE as substitution, validating the influence that middle layers have on the final prediction. Thus we can conclude that when two high-resource languages appear within LLM, the model itself will evolve into two separated conceptual spaces, enabling the model to 'think' in different languages. And we can use different language feature substitutions to change the language that the model 'thinks with', supporting the clustering of features in middle layers.

4 Applicative Scenarios

The uniqueness of SAE lies in its ability to extract interpretable features from neurons. Thus, it provides a plausible method to probe different conceptual spaces within Intrinsic Multilingualism. We focus on features that capture the specialties of ZH and EN, divided into syntactic and semantic features. We utilize the tool (Ge et al., 2024b) to visualize the features of SAE.

4.1 Syntactic Results

English Compared to Chinese, English's verb tense and attributive clauses are both important and representative syntactic structures.

First, we analyze the verb tense for both EN SAE and ZH SAE. As shown in Figure 12, EN SAEs have features detecting different types of verb inflection with specificity. Since our custom input contains two tenses, these features only activate for one of them across all samples. On the other hand, ZH SAEs don't have similar specific features. We have shown the ZH features with the best specialty, but they still attend to other tenses.

Next, we focus on English's attributive clauses, mainly on relative pronouns for clarity. As shown in Figure 13, EN features have better specificity for these pronouns, while ZH features attend to other components in sentences. Layer-8's features attend to different relative pronouns, and features in later layers become more fine-grained. One especially interesting feature is **L13-en-13971**, which focuses on *that*'s position in attributive clauses, even if it's omitted, showing the functionality of features is far beyond character-level matching.

Chinese For distinction from English and better comprehension, we select "quantifiers" as Chinese's special syntactic structure. Most objects have their special quantifier in Chinese, while only a special set has them in English, like *a cup of tea* or *a piece of paper*.

As shown in Figure 14, ZH features are very specific to quantifiers without obvious activation for other components. Although EN features could activate towards different quantifiers, their activation is more dispersed, showing a gap in recognition.

4.2 Semantic Results

Since features are the "minimal unit of comprehension", we decided to test feature understanding about *festivals* because it is cultural-related and has clear-defined names.

As shown in Figure 15, ZH features focus more on Chinese-related festivals like *Spring Festival* and *Mid-autumn Festival*, while EN features attend more to Western festivals like *Christmas* and *Halloween*. This disparity provides a clear view of the separate conceptual spaces within the multilingual

[2]The input prompt in English is : .....(few shot) A "__" is used to read stories. Answer: ". And the input prompt in Chinese is : .....(few shot) "_"是一种通过物质或空间传递能量的扰动。答案："。

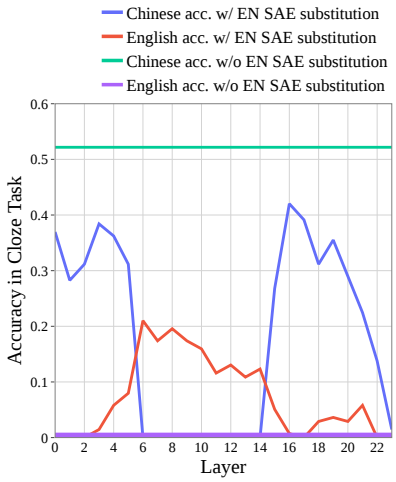

Figure 11: Result of substitution's prediction being the answer or the answer's translative pair.[1]

Layer	Input EN		Input ZH	
	EN SAE	ZH SAE	ZH SAE	EN SAE
0	book	book	书 (book)	故事 (story)
1	book	book	书 (book)	书 (book)
2	book	book	书 (book)	书 (book)
3	book	book	书 (book)	故事 (story)
4	book	书 (book)	书 (book)	文字 (writing)
5	book	书 (book)	书 (book)	文字 (writing)
6	story	书 (book)	书 (book)	story
7	book	书 (book)	书 (book)	故事 (story)
8	book	书 (book)	故事 (story)	story
⋮	⋮	⋮	⋮	⋮
14	story	阅读 (read)	故事 (story)	story
15	book	story	读 (read)	story
16	book	book	书 (book)	故事 (story)
17	book	book	书 (book)	书 (book)
⋮	⋮	⋮	⋮	⋮
21	book	book	书 (book)	书 (book)
22	book	book	读 (read)	(blank)
23	book	a	书 (book)	的 (of)

Table 1: Causal results of a piece of few-shot cloze task generated by Qwen1.5-1.8B. The corresponding English translation of each Chinese token is shown in grey within the brackets next to it.[2]

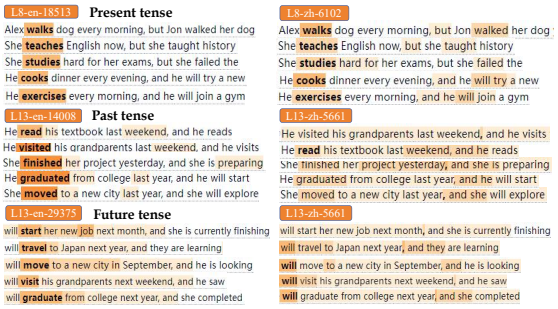

Figure 12: ZH and EN features attending to verb's different tenses

Figure 14: ZH and EN features attending to different quantifiers

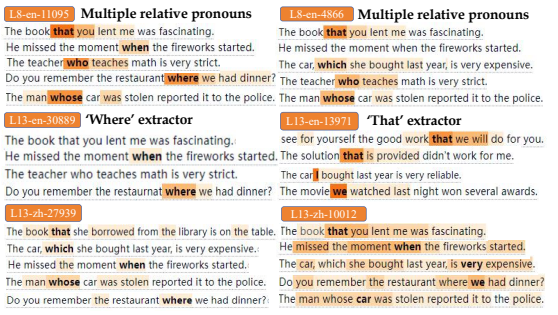

Figure 13: ZH and EN features attending to attributive clause's relative pronoun

Figure 15: Semantic difference between SAEs' features.

model and opens up more opportunities for further investigation.

5 Discussion

Intrinsic Multilingualism and Transferred Multilingualism has great potential in discovering and remedying multilingual bias researches/ But currently, we can only conduct case studies and qualitative research into the features learned by ZH and EN SAEs for the application part because we cannot train SAEs on larger models that showcase obvious language bias.

However, we believe this research approach has greater potential. Since ZH SAE and EN SAE provide a way to examine the model's conceptual space in different languages separately, we can explore certain multilingual phenomena more deeply and with finer granularity instead of testing bias simply by prompts.

Current multilingual benchmarks don't necessarily focus on bias detection: they are just common benchmarks written in other languages. Even if an LLM performs well on these benchmarks, it could be due to the "pivot language"'s outstanding ability (Wendler et al., 2024).

With the help of SAEs, we can study language bias within LLMs using SAEs and identify the source of these biases by examining the cultural or language-related features within a specific SAE. If a language's SAE doesn't meet these criteria, we can infer that it may exhibit a certain type of bias in generation. We believe SAEs provide us with a thorough and systematic method to evaluate the model's inner mechanisms and help us recognize hidden patterns that are not clearly shown in prompt-generation tests.

Furthermore, we think SAEs can also help alleviate certain unwanted behaviors of LLMs and enable us to steer them towards specific needs. Anthropic has shown how to steer LLM features in (Templeton et al., 2024). In a multilingual setting, we can discover and use more language-specific features to achieve more versatile operations and modifications on the model. This enables us to post-process potential multilingual biases within LLMs.

Future work can focus on the following directions: (i) Train separate SAEs on larger models that perform well on multilingual benchmarks but exhibit pivot language biases. SAEs can explain the origin of language bias. (ii) Study how to steer features to mitigate language bias within multilingual models. Further investigate the features' effects on different language performances. (iii) Set up a benchmark at the feature level to evaluate multilingual LLMs' overall multilingual ability, paying attention to the balance between different languages. Provide suggestions for data composition during the pretraining stage.

6 Related Work

Mechanistic Interpretability Mechanistic interpretability aims to reverse-engineer neural networks to understand their mechanisms. Cammarata et al. (2020) worked on the mechanistic interpretability of vision models, specifically InceptionV1. Due to superposition (Elhage et al., 2022b) and the polysemanticity of neurons, dictionary learning (Faruqui et al. (2015); Arora et al. (2018); Bricken et al. (2023)), previously applied to word embeddings, leverages the sparse autoencoder (SAE) (Huben et al., 2024) to discover linear combinations of features.

Similarity Analysis Methods Previous works have primarily applied static analysis to compare activations between different languages. Techniques such as Singular Value Decomposition (SVD) (Raghu et al., 2017), Canonical Correlation Analysis (CCA) (Singh et al., 2019), and PARAFAC2 (Zhao et al., 2023) have been used.

Multilingual LMs' Hypothesis of Representations As shown by Pires et al. (2019) and Singh et al. (2019), representations consist of a language-specific component, which identifies the language of the sentence, and a language-neutral component, which captures the sentence's meaning independently of the language. Many studies, such as Liang et al. (2021) and Choenni and Shutova (2020), follow this hypothesis using probing methods. Choenni and Shutova (2020) specifically focused on encoder-based LMs like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and LASER (Artetxe and Schwenk, 2019).

From the perspective of neuron architecture, Foroutan et al. (2022) applied the lottery ticket hypothesis to identify subnetworks for evaluating transferability between languages. They concluded that multilingual models contain both language-neutral and language-specific components, with the language-neutral components being more prominent in cross-lingual transfer performance.

Pivot Language Recent research has used pivot languages, such as English, to improve the capabilities of large language models (mainly GPT-3.5-turbo) (Ahuja et al., 2023). Wendler et al. (2024) used the logit lens (Nostalgebraist, 2020) to argue that multilingual models conceptually employ English-biased internal lingua franca in a semantic sense.

7 Conclusion

From macro analysis experiment (SAE substitution), we find that in the model Qwen1.5-1.8B, the reconstruction performance of a different language would get down in the middle layers and forms a U-shape curve in total. From micro analysis experiment (Activation Similarity), we gain the insight that only the substitution in the middle layers will cause an obvious final divergence between original activation and substitution activation. From causal analysis experiment (Substitution Inference), the change of <next token> prediction also take place in the middle layers. Based on all these 3 experiments' results, we propose the conceptual model that text generation for multilingual model is divided into 3 stages: detokenization, conceptual stage and retokenization. For Transferred Multilingualismmodel, the conceptual stages between different languages are partly separated. Input in different languages would go on different track

Limitations

Though we offer new insight of model multilingualism, we only study one model and one pair of languages for Transferred Multilingualism and Intrinsic Multilingualism, respectively, which restricts the scope and generality of our methods. Moreover, our narrow focus on SAEs trained in the residual streams suggest more space for improvement. For example, one can also decompose the activation of each module writing into the residual stream to understand the function of each individual module and also how these features form circuits so that one can understand the information flow inside of models, which is not included in this work.

Ethics Statement

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Manel Aloui, Hasna Chouikhi, Ghaith Chaabane, Haithem Kchaou, and Chehir Dhaouadi. 2024. 101 billion arabic words dataset. *Preprint*, arXiv:2405.01590.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2018. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

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. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. 2020. Thread: Circuits. *Distill*. Https://distill.pub/2020/circuits.

Jianghao Chen, Pu Jian, Tengxiao Xi, Dongyi Yi, Qianlong Du, Chenglin Ding, Guibo Zhu, Chengqing Zong, Jinqiao Wang, and Jiajun Zhang. 2023. Chinesewebtext: Large-scale high-quality chinese web text extracted with effective evaluation model. *Preprint*, arXiv:2311.01149.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multi-

lingual sentence encoders for typological properties. *Preprint*, arXiv:2009.12862.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. 2022a. Softmax linear units. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/solu/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. 2022b. Toy models of superposition. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2022/toy_model/index.html.

Nelson Elhage, Robert Lasenby, and Christopher Olah. 2023. Privileged bases in the transformer residual stream. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2023/privileged-basis/index.html.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. 2024. Interpreting the second-order effects of neurons in clip. *arXiv preprint arXiv:2406.04341*.

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

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Xuyang Ge, Fukang Zhu, Wentao Shu, Junxuan Wang, Zhengfu He, and Xipeng Qiu. 2024a. Automatically identifying local and global circuits with linear computation graphs. *Preprint*, arXiv:2405.13868.

Xuyang Ge, Fukang Zhu, Junxuan Wang, Wentao Shu, Lingjie Chen, and Zhengfu He. 2024b. Openmoss language model sparse autoencoders.

Xuyang Ge, Fukang Zhu, Junxuan Wang, Wentao Shu, and Zhengfu He. 2024c. Sparse dictionary learning on language models: Infrastructure, observations and agenda. `https://open-moss.com/en/language-model-SAEs/`.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Zhengfu He, Xuyang Ge, Qiong Tang, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024. Dictionary learning improves patch-free circuit discovery in mechanistic interpretability: A case study on othello-gpt. *CoRR*, abs/2402.12201.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2021. Locating language-specific information in contextualized embeddings. *Preprint*, arXiv:2109.08040.

Nostalgebraist. 2020. Interpreting gpt: the logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Xiangyang Liu, Hang Yan, Yunfan Shao, Qiong Tang, Shiduo Zhang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. 2024. Moss: An open conversational large language model. *Machine Intelligence Research*.

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. *CoRR*, abs/2402.16438.

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. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Preprint*, arXiv:2402.10588.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *CoRR*, abs/2402.18815.

Zheng Zhao, Yftah Ziser, Bonnie Webber, and Shay Cohen. 2023. A joint matrix factorization analysis of multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12764–12783, Singapore. Association for Computational Linguistics.

A Dataset Introduction

ChineseWebText (Chen et al., 2023) is a large-scale dataset containing a wide range of Chinese texts collected from the web. It includes diverse content types, such as news articles and blogs. This diversity ensures that our model is exposed to various linguistic contexts and styles, thus enabling SAEs to extract more features and ensure the completeness of our following experiments.

Pile (Gao et al., 2020) is an extensive English language dataset composed of multiple smaller datasets covering a broad spectrum of domains. Using the Pile allows us to train our model on a rich and varied dataset, improving its performance and generalization across different types of English language data.

B Sparse Autoencoder Training

We trained SAE for residual stream of all layers in Qwen1.5-1.8B (24 layers) and Phi-2 (32 layers).

B.1 Architecture

We expand the dimension of the model's intermediate activations by 16, which means that the shape of hidden units in the SAE trained for Qwen1.5-1.8B is 16×2048 and 16×2560 for Phi-2. The forward process can be illustrated by

$$\begin{aligned} n_x &= \sqrt{d}/||x||_2 \\ z &= \text{ReLU}\left(W_{\text{enc}}\left(x \cdot n_x\right) + b_{\text{enc}}\right) \\ \hat{x} &= \left(W_{\text{dec}} z\right)/n_x \end{aligned} \tag{1}$$

with $W_{\text{enc}} \in \mathbb{R}^{n \times d}$, $b_{\text{enc}} \in \mathbb{R}^n$, $W_{\text{dec}} \in \mathbb{R}^{d \times n}$ (Gao et al., 2024). The $\hat{x}$ is the reconstructed activations. The loss function is $\mathcal{L} = ||x - \hat{x}||_2^2 + \lambda||z||_1$, with λ refer to the L1 coefficient that controls the sparsity punishment.

B.2 Training Setting

Here we utilize ghost gradients to avoid "dead features", which means that the neurons are always in the state of deactivation. We use Adam as our optimizer and set 2.5e-5 as the max learning rate. Besides, the context size (sequence length) input to the model is set to 256 and λ (L1 coefficient) is set to 5e-4 which show a good balance between sparsity and reconstruction. Last, the batch size is set to 2048.

B.3 Statistics of SAE

Sparse autoencoders are not perfect feature extractors. In fact, there are two kinds of metrics to evaluate the performance of SAE. First, we use EV and $recostructed \quad cross-entropy \quad ratio$ to evaluate if the SAE could reconstruct a good $\hat{x}$.

$$\begin{aligned} EV &= 1 - \frac{||\hat{x} - x||_2^2}{\sigma^2(x)} \\ ratio &= \frac{\mathcal{L}_{\text{recons}} - \mathcal{L}_{\text{ablate}}}{\mathcal{L}_{\text{original}} - \mathcal{L}_{\text{ablate}}} \end{aligned} \tag{2}$$

with $\mathcal{L}_{\text{recons}}$, $\mathcal{L}_{\text{original}}$ and $\mathcal{L}_{\text{ablate}}$ representing the reconstruction CE loss (calculated by using substitution of SAE), the original CE loss and the ablated CE loss (calculated by setting the activation to be zero) (Ge et al., 2024a).Table 2, Table 3 and Table 4 are the statistics of the SAE trained on EN, ZH and MIX dataset on Qwen1.5-1.8B.

Layer	L0	Explained Variance	CE Score	CE Loss
0	46.37	88.98%	0.99	2.6
1	42.11	86.32%	0.99	2.61
2	42.47	84.28%	0.99	2.66
3	48.77	81.28%	0.98	2.7
4	63.21	79.09%	0.97	2.71
5	76.91	76.53%	0.99	2.71
6	82.87	84.23%	0.98	2.72
7	88.32	76.53%	0.98	2.73
8	106.36	73.51%	0.98	2.75
9	117.88	68.75%	0.97	2.81
10	121.01	68.28%	0.98	2.82
11	121.18	67.25%	0.97	2.82
12	116.87	71.11%	0.97	2.83
13	116.23	68.6%	0.97	2.88
14	109.98	69.61%	0.97	2.88
15	104.86	70.45%	0.96	2.89
16	102.78	70.94%	0.97	2.88
17	99.69	71.0%	0.96	2.91
18	97.0	71.78%	0.96	2.94
19	95.57	71.13%	0.95	2.98
20	96.79	70.7%	0.95	3.04
21	101.02	68.68%	0.95	3.12
22	104.12	66.39%	0.95	3.32
23	101.81	76.42%	0.88	3.66

Table 2: EN SAE training statistics on Qwen1.5-1.8B

C Low-resource language in Intrinsic Multilingualism

Apart from the high-resource languages in Qwen-1.5, we also selected a low-resource language to investigate how Transferred Multilingualism is demonstrated in Qwen and whether our previous conclusions still hold.

The results are shown in Figure 16. First, we analyze how ZH and EN features represent ARA features. When the input is Arabic, EN and MIX SAE achieve excellent reconstruction performance across all layers, while ZH SAE performs comparatively worse, indicating that Transferred Multilingualism exists and is transferred by EN in Qwen.

Layer	L0	Explained Variance	CE Score	CE Loss
0	19.37	92.5%	1.0	3.08
1	26.36	89.41%	0.99	3.09
2	30.11	86.92%	0.99	3.12
3	43.24	82.57%	0.99	3.16
4	60.73	79.33%	0.98	3.18
5	77.0	76.48%	0.99	3.18
6	86.32	79.25%	0.99	3.19
7	90.23	75.71%	0.98	3.21
8	106.78	77.55%	0.98	3.24
9	118.01	71.48%	0.97	3.29
10	122.23	75.69%	0.98	3.32
11	121.99	69.5%	0.97	3.33
12	117.86	75.07%	0.97	3.35
13	118.79	66.79%	0.97	3.4
14	111.81	66.97%	0.97	3.41
15	107.97	69.56%	0.97	3.42
16	104.97	71.35%	0.96	3.44
17	101.75	70.88%	0.96	3.48
18	101.59	72.04%	0.95	3.54
19	101.55	71.22%	0.93	3.55
20	103.46	71.46%	0.92	3.62
21	107.43	69.19%	0.94	3.76
22	108.93	65.97%	0.89	4.02
23	110.77	73.01%	0.87	4.16

Table 3: ZH SAE training statistics on Qwen1.5-1.8B

Layer	L0	Explained Variance	CE Score	CE Loss
0	38.07	89.85%	0.99	3.01
1	37.46	87.91%	0.99	3.02
2	41.55	84.27%	0.99	3.06
3	53.28	80.14%	0.98	3.11
4	73.26	77.14%	0.97	3.14
5	91.17	74.45%	0.98	3.12
6	98.97	83.1%	0.98	3.14
7	97.58	82.39%	0.98	3.15
8	112.89	79.26%	0.97	3.18
9	125.23	75.31%	0.97	3.23
10	129.91	68.2%	0.98	3.26
11	129.93	72.95%	0.97	3.27
12	127.76	71.31%	0.97	3.28
13	126.52	67.35%	0.97	3.31
14	118.95	68.27%	0.97	3.33
15	114.94	68.37%	0.96	3.33
16	111.64	69.83%	0.96	3.36
17	107.06	69.87%	0.95	3.4
18	104.53	70.31%	0.95	3.46
19	104.85	70.56%	0.93	3.51
20	103.29	70.5%	0.93	3.59
21	106.33	68.41%	0.94	3.7
22	99.0	66.87%	0.9	4.02
23	103.69	76.94%	0.84	4.37

Table 4: MIX SAE training statistics on Qwen1.5-1.8B

Next, we analyze how ARA SAE performs in reconstructing ZH or EN. The overall CE score is very low, indicating that features related to Arabic don't contain much information about ZH and EN. However, the reconstruction result is better in the middle layers, meaning features are more aligned with ZH and EN features. This phenomenon aligns with our Transferred Multilingualism model, where

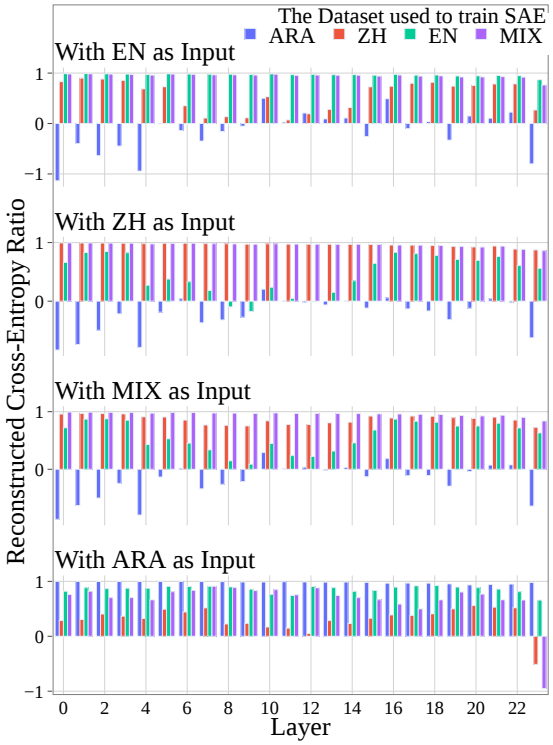

Figure 16: Qwen1.5's SAE substitution result.

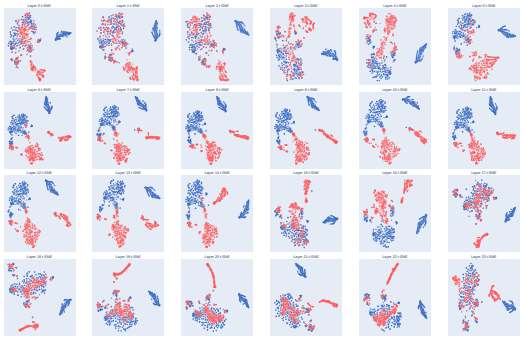

Figure 17: Full result of microscopic analysis after SAE substitution in different layers

in the middle layers, ARA features will be closer to EN features for processing, resulting in better reconstruction performance compared to early and final layers.

D Micro Analysis

Figure 17 shows the final layer's activation result comparison between original activation and activation after SAE substitution, we can see a clear divisible result when SAE substitution is applied in the middle layers.

E Cross SAE Analysis Method

Previous research using SAEs (Ge et al., 2024c,b) mostly focuses on a single SAE. We explore using multiple SAEs from one model to study the model's capabilities in different aspects. We will formulate our method, hoping it can provide insights for future research.

E.1 Input Data Choice

First, we need to decide the data used to induce SAEs from the model. The data should be related to a certain facet of the model. For studying Intrinsic Multilingualism, we choose high-resource language datasets. Additionally, we must ensure the datasets do not overlap to maintain feature purity. The training method is detailed in B.

E.2 Feature Comparison

First, we directly analyze the decoders of ZH and EN SAEs since they include the key features for the two languages. We calculate the Pearson correlation between the rows of the two decoder matrices. The resulting correlation matrix will have the shape of (ZH-feature-number, EN-feature-number). Then, we take the argmax on either dimension to discover the 'match' between the two SAEs' features. However, the results do not reveal anything meaningful; similarity across all layers is consistently low.

This feature analysis has a pivotal premise: the data used to induce SAEs should be of similar format. In our setting, Chinese and English inputs may take on totally different mechanisms, leading to intrinsic differences in their features. We tested the similarity between 10 translative pairs of the same word in ZH and EN SAEs. The average similarity is only 0.52, indicating this experimental setting isn't suitable.