

Tracing Multilingual Representations in LLMs with Cross-Layer Transcoders

Abir Harrasse^{*1,2,3} Florent Draye^{*1} Punya Syon Pandey¹
Zhijing Jin^{1,2,4} Bernhard Schölkopf^{1,5}

¹Max Planck Institute for Intelligent Systems, Tübingen, Germany

²Jinesis AI Lab, University of Toronto & Vector Institute

³Mohammed VI Polytechnic University, Morocco

⁴EuroSafeAI ⁵ELLIS Institute Tübingen

^{*}Equal contribution

{aharrasse, zjin}@cs.toronto.edu, fdraye@tue.mpg.de

Abstract

Multilingual Large Language Models (LLMs) can process many languages, yet how they internally represent this diversity remains unclear. Do they form shared multilingual representations with language-specific decoding, and if so, why do multilingual performance gaps persist? To address this question, we train models on different multilingual mixtures and analyze their internal mechanisms using *Cross-Layer Transcoders* (CLTs) and *Attribution Graphs*. Our results reveal a consistent three-phase structure across models: early layers encode language-specific information, middle layers converge to highly shared multilingual representations, and final layers re-specialize for language-specific decoding.

Training models without English still produces the same shared multilingual structures, suggesting that these representations emerge independently of English dominance. We further show that decoding relies on a small set of high-frequency late-layer features that linearly read out language identity from earlier layers, and that intervening on these features allows one language to be suppressed and another substituted. Finally, through a *Model-Diffing* experiment, we show that multilingual failures arise less from missing shared representations than from weak late-layer decoding features, weak middle-layer clusters, and tokenizer fragmentation forcing early layers to specialize in word reassembly. Finetuning strengthens these features and their connections, improving language-specific decoding and providing a mechanistic explanation for multilingual performance gaps. Code, trained models, and CLT artifacts are available at: <https://github.com/abirharrasse/MultilingualCLTs>.

TRAINING SETUP

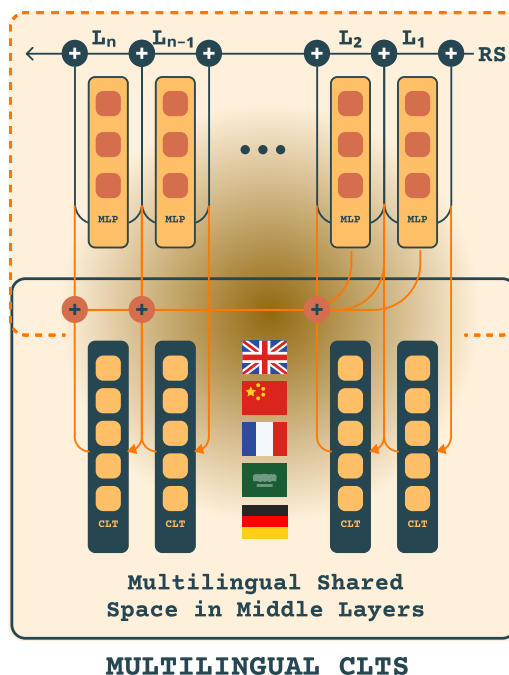
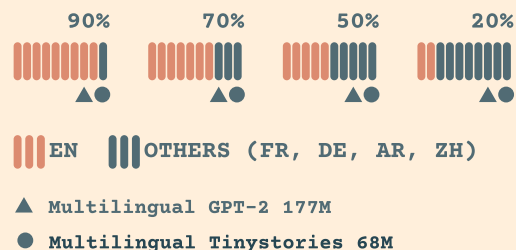


Figure 1: Cross-Layer Transcoders (CLTs) (Ameisen et al., 2025) (bottom) reveal that multilingual LLMs (middle) consistently form a shared language-agnostic space in middle layers (gradient middle region) where all languages converge, regardless of English dominance or presence in the training mixtures (top).

1 Introduction

The inclusion of diverse languages in the large-scale training data of LLMs (Team, 2024b,a;

Brown et al., 2020; Chowdhery et al., 2022) has led to remarkable multilingual capabilities (Shi et al., 2022; Conneau et al., 2020; Workshop, 2023). However, the mechanisms underlying multilingual

processing remain poorly understood. Prior investigations have yielded conflicting views of multilingual representations. Some studies suggest that LLMs process non-English inputs through English-aligned latent spaces (Wendler et al., 2024; Schut et al., 2025), while others find evidence of shared cross-lingual grammatical structures encoded in common feature directions (Brinkmann et al., 2025). Existing approaches, however, rely primarily on static probing methods that cannot capture how linguistic information is dynamically transformed across layers during computation.

To address these limitations, we derive attribution graphs from self-trained cross-layer transcoders (CLTs) (Ameisen et al., 2025), allowing us to trace feature interactions across layers and study how models transition between language-specific and shared multilingual processing modes. Our results reveal highly shared multilingual representations in middle layers together with sparse late-layer features responsible for language-specific decoding. We further show that these structures emerge even in models trained without English, suggesting that shared multilingual representations are not solely a consequence of English dominance.

Building on these findings, we conduct causal interventions on late-layer language features and show that suppressing features from one language while amplifying another allows the output language to be redirected. We further investigate multilingual failure modes through a *Model-Diffing* experiment, linking non-English underperformance to weak late-layer decoding features and tokenizer fragmentation that forces early layers to specialize in word reassembly.

These findings suggest that multilingual failures may arise not from the absence of shared multilingual computation, but from weak language-specific decoding pathways operating on top of otherwise shared internal representations. This shifts the focus of multilingual robustness from learning fully separate representations toward improving tokenization and decoding efficiency in low-resource languages.

Our contributions are as follows:

1. We show that multilingual LLMs form highly shared multilingual representations, relying on consistent internal circuits to process equivalent inputs across languages. These structures emerge even without English or a domi-

nant language.

2. We identify sparse high-frequency late-layer features responsible for language-specific decoding, and show through causal interventions that manipulating these features allows one language to be suppressed and another substituted during generation.
3. We provide a mechanistic explanation for multilingual performance gaps, linking non-English failures to weak decoding features, weak middle-layer clusters, and tokenizer fragmentation, while showing that finetuning strengthens these pathways and improves language-specific decoding.
4. We will release nine multilingual models (two sizes across four multilingual mixtures and one non-English mixture), three finetuned models, and their corresponding cross-layer transcoders (CLTs) to support future multilingual mechanistic interpretability research.

2 Related Work

2.1 Multilingual Language Models

Multilingual language models have evolved from mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) to modern LLMs such as LLaMA (Touvron et al., 2023), Gemma (Team, 2024a), and BLOOM (Workshop, 2023). Despite strong multilingual capabilities (Shi et al., 2022; Conneau et al., 2020), systematic disparities persist across languages (Pires et al., 2019; Hu et al., 2020), with English consistently outperforming low-resource languages and complex multilingual reasoning tasks (Lauscher et al., 2020; Shi et al., 2022). While English-dominant training distributions have been hypothesized as a major factor (Bender et al., 2021), the mechanisms underlying these disparities remain poorly understood.

2.2 Internal Representations and the Pivot Hypothesis

Recent studies suggest that multilingual LLMs process non-English inputs through English-aligned latent representations. Wendler et al. (2024) showed that intermediate layers often predict English tokens before target-language tokens, while Schut et al. (2025) found systematic alignment toward English-like representations regardless of input language. These findings motivated approaches based

on embedding alignment (Lample and Conneau, 2019; Artetxe et al., 2018), translation-based fine-tuning (Zhu et al., 2024), romanization strategies (Saji et al., 2025), and English chain-of-thought prompting (Wei et al., 2022; Vatsal et al., 2025).

At the same time, other work has found evidence of shared cross-lingual grammatical structures encoded in common feature directions (Brinkmann et al., 2025), suggesting partially shared multilingual representations. More broadly, these observations align with the “Platonic representation hypothesis” (Huh et al., 2024), which proposes that neural networks converge toward abstract representations beyond surface forms.

These findings are also related to prior work studying cross-lingual representation alignment and multilingual transfer in encoder-based multilingual models (Doddapaneni et al., 2025; Choenni and Shutova, 2020), which showed that multilingual representations can become partially aligned across languages despite substantial differences in scripts and token distributions.

However, existing approaches rely primarily on static probing methods, limiting their ability to explain how multilingual representations evolve across layers, how language-specific decoding emerges, and why multilingual failures persist despite shared representations. In contrast, our work studies these questions mechanistically by tracing feature interactions across layers during computation.

2.3 Mechanistic Interpretability in LLMs

Mechanistic interpretability seeks to understand transformer computations through circuit analysis (Olah et al., 2020; Elhage et al., 2021). Sparse Autoencoders (SAEs) decompose activations into interpretable features (Cunningham et al., 2023; Templeton et al., 2024), while transcoders directly reconstruct MLP outputs (Dunefsky et al., 2024). Cross-Layer Transcoders (CLTs) further extend this approach by enabling tractable cross-layer attribution graphs (Ameisen et al., 2025; Hanna et al., 2025).

Mechanistic analyses of multilingual processing remain sparse. Beyond logit-lens analyses (Wendler et al., 2024) and Anthropic’s multilingual case study (Lindsey et al., 2025), prior work has largely relied on probing or representation similarity analyses. In contrast, our work provides a mechanistic account of how multilingual representations evolve across layers, how language-specific

decoding emerges, and why multilingual failures persist despite shared representations.

3 Experimental Setup and Methods

3.1 Mixture of multilingual LLMs

We study multilingual behavior across English, German, French, Arabic, and Chinese using four GPT-2 style models (12 layers, ~ 177.6 M parameters) trained on OpenWebText + FineWeb2 (Penedo et al., 2025), four TinyStories-style models (4 layers, ~ 68.5 M parameters), and an off-the-shelf LLaMA-3.2-1B model (Meta AI, 2026). To study the effect of training distribution, we vary the English proportion from balanced (20%) to dominant (90%), with the remaining data evenly split across the other languages. We additionally train a no-English GPT-2 model using only German, French, Arabic, and Chinese to investigate whether shared multilingual representations emerge independently of English. All models use tokenizers trained on balanced multilingual data to isolate representation differences from tokenizer allocation effects. Figure 1 illustrates the training configurations.

3.2 Cross-Layer Transcoders (CLTs)

To probe internal multilingual mechanisms, we employ CLTs, which map activations between layers to reveal how semantic representations evolve during processing. A CLT consists of neurons (features) divided into encoder and decoder components.

Formally, to run a cross-layer transcoder, let $\mathbf{h}_\ell \in \mathbb{R}^{d_{\text{model}}}$ be the input to the MLP at layer ℓ for a single token position. We define

$$\mathbf{z}_\ell = \text{ReLU}(\mathbf{W}_{\text{enc}}^\ell \mathbf{h}_\ell + \mathbf{b}_{\text{enc}}^\ell) \in \mathbb{R}^{d_{\text{features}}}, \quad (1)$$

where $\mathbf{W}_{\text{enc}}^\ell \in \mathbb{R}^{d_{\text{features}} \times d_{\text{model}}}$ is the encoder weight matrix and $\mathbf{b}_{\text{enc}}^\ell \in \mathbb{R}^{d_{\text{features}}}$ is the encoder bias. The CLTs then reconstructs the MLP output at layer ℓ' as

$$\hat{\mathbf{m}}_{\ell'} = \sum_{\ell \leq \ell'} \mathbf{W}_{\text{dec}}^{\ell \rightarrow \ell'} \mathbf{z}_\ell + \mathbf{b}_{\text{dec}}^{\ell'}, \quad (2)$$

where $\mathbf{W}_{\text{dec}}^{\ell \rightarrow \ell'} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{features}}}$ is the decoder matrix from layer ℓ to ℓ' .

Following Anthropic guidelines (Ameisen et al.,

2025), the final training objective is

$$\begin{aligned} \mathcal{L} = & \underbrace{\sum_{\ell'} \|\hat{\mathbf{m}}_{\ell'} - \mathbf{m}_{\ell'}\|_2^2}_{\text{MSE reconstruction}} \\ & + \lambda_0 \underbrace{\sum_{\ell} \tanh(C(\mathbf{z}_{\ell} \odot \|\mathbf{W}_{\text{dec}}^{\ell}\|))}_{L_0 \text{ sparsity}} \\ & + \lambda_{\text{df}} \underbrace{\sum_{\ell} \text{ReLU}(\exp(\tau) - \mathbf{h}_{\ell}^{\text{pre}}) \|\mathbf{W}_{\text{dec}}^{\ell}\|}_{\text{dead feature penalty}}, \end{aligned} \quad (3)$$

where $\mathbf{W}_{\text{dec}}^{\ell}$ are concatenated decoder weights for layer ℓ , $\mathbf{h}_{\ell}^{\text{pre}}$ are pre-activation values, τ is a threshold parameter, and C is a scaling constant. λ_0 and λ_{df} control the strength of the sparsity and dead-feature regularization terms.

We train CLTs on activation vectors sampled uniformly across languages until convergence. Training details, hyperparameters, and performance metrics appear in Appendix B.

3.3 Graph attribution

Following Ameisen et al. (2025), we compute the attribution score between every feature n at layer ℓ , position k , and feature n' at layer ℓ' , position k' , as

$$a_{\ell,k,n}^{\ell',k',n'} = \sum_{\ell \leq s \leq \ell'} f_{k,n}^{\ell \rightarrow s} J_{s,k}^{\ell',k'} g_{k',n'}^{\ell'}, \quad (4)$$

where $f_{k,n}^{\ell \rightarrow s}$ denotes the vector for feature n in the decoder matrix projecting from layer ℓ to s , $J_{s,k}^{\ell',k'}$ is the Jacobian between the MLP output at (ℓ', k') and the MLP input at (ℓ, k) , computed during a forward pass where nonlinearities (normalization layers, attention computations, and MLPs) are frozen using stop-gradient operations, and $g_{k',n'}^{\ell'}$ is the corresponding encoder feature at layer ℓ' and position k' . The graph is then pruned to retain only features that cumulatively account for 80% of the effect on the final logit, and edges that cumulatively account for 95% of the edge effect on the final logit. We use the circuit-tracer library (Hanna et al., 2025).

3.4 Multilingual Score

For each CLT feature f , we compute a multilingual score using balanced multilingual samples. Let

$$p_l(f) = \frac{A_l(f)}{\sum_{l'} A_{l'}(f)}, \quad (5)$$

where $A_l(f)$ is the number of sequences in language l activating feature f . We define multilinguality as the normalized entropy

$$H(f) = - \sum_{l=1}^L p_l(f) \log p_l(f), \quad (6)$$

where low entropy indicates language-specific features and high entropy indicates multilingual features.

4 Results

We organize our results around three questions: (i) do multilingual models form shared internal representations across languages, (ii) how do they decode language identity and route information to the correct output language, and (iii) why do multilingual failures persist despite shared representations?

4.1 Emergence of Shared Multilingual Representations

Generalization Under Data Imbalance. We begin by assessing whether models trained on highly imbalanced multilingual mixtures still generalize across languages. Figure 11 reports validation cross-entropy losses for models trained with varying proportions of English tokens, including extreme settings where English accounts for 90% of the training data. Despite this imbalance, the models maintain competitive performance not only in English but also in non-English languages such as Arabic, suggesting that multilingual representations remain partially shared rather than collapsing toward dominant-language-specific circuits. Additional results are provided in Appendix C.

Language Entropy Across Layers. To study how multilingual representations evolve across layers, we analyze the language entropy of feature activations. For each feature f , we compute the multilingual score $H(f)$ defined in Section 3.4 and weight it by the feature activation rate over 4M training tokens. Figure 2 reports the resulting weighted entropy across layers for models trained on different multilingual mixtures.

Across all mixtures, we observe a consistent U-shaped pattern: entropy is low in early layers, rises sharply in middle layers, and decreases again toward the output layers. This suggests that early layers primarily encode language-specific information, middle layers converge toward highly shared multilingual representations, and late layers partially re-specialize for language-specific decoding.

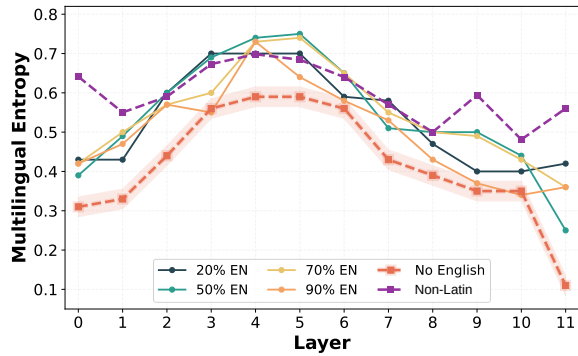


Figure 2: Rate-weighted multilingual score across layers. Middle layers consistently form a shared multilingual space while early and late layers remain language-specific. This pattern holds across all mixtures and scales, including a model trained entirely without English (dashed) and a balanced non-Latin-script model (dashed), suggesting that shared multilingual representations are not solely an artifact of English dominance or script identity.

The same layerwise structure remains stable across training mixtures, model scales, and script identity. In particular, we observe similar entropy profiles in LLaMA-3.2-1B CLTs (Figure 20), and the non-Latin-script model preserves the same middle-layer peak while sitting persistently above the mixture-specific curves, indicating that the shared multilingual structure extends beyond Latin-script languages. Similar alignment trends also emerge in independent representation-similarity analyses on larger multilingual models (Appendix G), suggesting that the observed organization is not specific to CLT attribution graphs.

In contrast, the 4-layer TinyStories models exhibit substantially flatter entropy profiles (Figure 20), suggesting that deeper models may be required for strongly separated multilingual organization to emerge.

To test whether shared multilingual representations depend on English dominance, we additionally train a GPT-2 model exclusively on French, German, Arabic, and Chinese. The resulting entropy profile exhibits the same characteristic rise-and-fall structure, with peak multilinguality in middle layers (Figure 2), indicating that shared multilingual representations can emerge even in fully non-English settings.

The same behavior appears when using the full training-distribution entropy instead of the top100 score. A comparison between both metrics is provided in Appendix E.

Multilingual Circuits Across Data Mixtures.

To complement the entropy analysis, we extract attribution graphs across data mixtures and sentence types. We study both function-word predictions (e.g., determiners and prepositions) and semantic predictions involving content words, calendar terms, and analogies. Figure 3 shows representative circuits for an English preposition and a German semantic prediction, with additional examples in Appendix H.1.

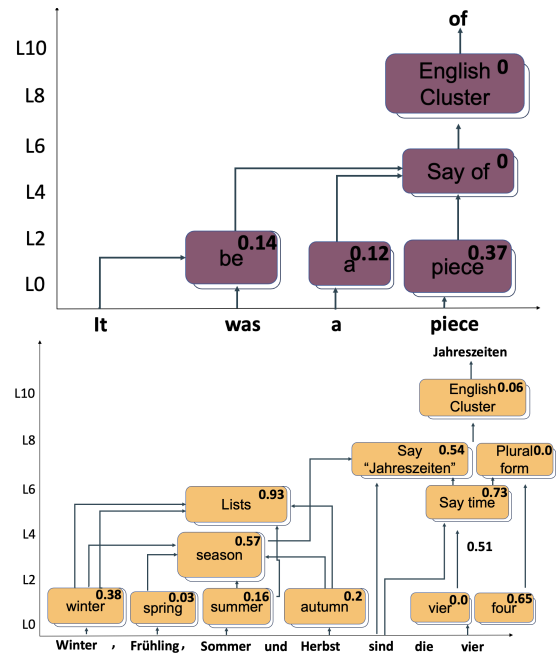


Figure 3: Representative attribution graphs across multilingual mixtures. Middle layers consistently form multilingual clusters, while early and late layers remain more language-specific. Scores at the top-right of each cluster indicate language entropy, with higher values corresponding to more multilingual clusters.

Across mixtures and sentence types, circuits consistently exhibit a three-phase organization: early and late layers remain largely language-specific, while middle layers form dense multilingual clusters. Determiner circuits remain more language-specific, consistent with observations from Schut et al. (2025). However, semantic alignment can emerge surprisingly early when tokenization is clean. In several examples, the first attention and MLP blocks already connect semantically equivalent words across languages, suggesting that multilingual semantic mapping begins from the earliest layers.

We additionally observe that the role of English varies across mixtures. In several cases from the 90% English models, late-layer English clusters are

absent despite strong multilingual alignment, indicating that semantic organization does not always require routing through English-specific representations.

Finally, multilingual failures are strongly shaped by tokenization. Even with balanced multilingual tokenizers, Arabic words are frequently fragmented into multiple subwords, forcing the first layers (up to layer 3) to focus on word reconstruction rather than higher-level semantic processing (Figures 26b and 27b).

4.2 Mechanisms of Language Decoding

A central question raised by shared multilingual representations is how models recover the correct output language during generation. Across attribution graphs, language-specific features appear predominantly in early and late layers. Early-layer features often correspond to token-level or context-specific language information rather than global language identity. Consistent with Lindsey et al. (2025), we occasionally observe early-layer features predicting the output language in constrained contexts (e.g., following quotation marks), though such features are not consistently present across prompts.

Across all models and data mixtures, language-specific decoding is mediated by sparse high-frequency late-layer features. As shown in Appendix D and Figure 18, most high-frequency CLT features are strongly language-specific. Figure 4 further shows that these late-layer language features can be linearly read from early-layer features and embedding nodes, indicating that language identity propagates hierarchically through the network before converging into sparse decoding circuits in final layers.

Direct intervention experiments confirm the causal role of these circuits. Zeroing late-layer features associated with one language while injecting activations from another language substantially redirects generation toward the target language, often shifting target-language logits from ranks around ~ 500 to the top-5 predictions. However, perfect language transfer typically requires modifying additional lower-frequency features or sweeping intervention strengths.

Activation frequency patterns further reveal both language ambiguity and dominance effects. A small number of features activate on 50–100% of tokens within each language, while inactive cases occur primarily at sequence beginnings where lan-

guage identity remains ambiguous, particularly between English, French, and German. Dominant languages also exhibit fewer high-frequency decoding features across both GPT-2 and LLaMA models, consistent with English behaving as a partially default decoding mode (Lindsey et al., 2025).

Finally, language-specific feature overlap varies systematically with training composition. Overlap between non-English and English features increases with English training proportion (77% to 84%), while non-English languages maintain stronger mutual overlap (89%) that progressively shifts toward English at higher proportions. These results suggest that training data composition shapes how multilingual representations converge and specialize during decoding (Appendix D).

To quantify this causal role more precisely, we evaluated the intervention on held-out English→French and English→German prompts using fixed hyperparameters, without per-example coefficient sweeps. We compare the sparse, high-frequency, late-layer CLT feature swap against two baselines: (i) a leave-one-out mean-difference vector $\Delta\mu$, added directly to the final-layer residual stream, and (ii) a random-feature control that preserves the positions, feature count, and activation values of the CLT swap but replaces the target-language feature indices with random CLT indices—matching sparsity and L_1 magnitude while destroying feature identity.

Averaged across both language pairs, the CLT swap reduces the median target-token rank to approximately 3% of the clean baseline while largely preserving fluency. The mean-difference baseline induces a $3.6\times$ larger logit perturbation yet severely degrades NLL and fails to improve rank, indicating that raw perturbation magnitude alone does not drive language redirection. The sparsity- and magnitude-matched random control reaches only 20%, substantially weaker than the CLT swap despite matching the intervention’s positions, activation values, and L_1 norm. Because this control isolates perturbation size and sparsity from feature identity, the gap between it and the CLT swap indicates that the effect is attributable to the identity of the intervened features rather than to the scale or sparsity of the intervention itself—evidence that the identified features causally control target-language token selection.

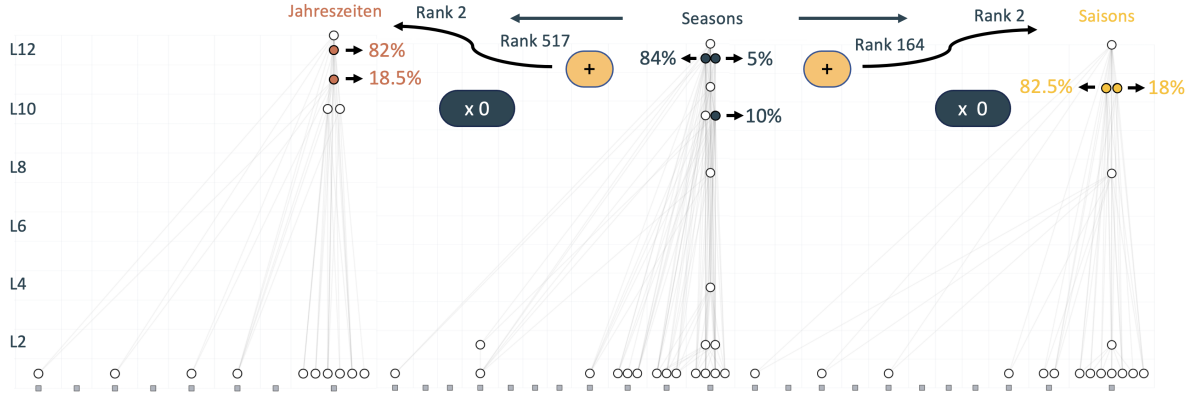


Figure 4: Attribution graphs in French, English, and German (20% model, pruned at 50%) for “Autumn, Winter, Fall, and Spring are the four”. Language features are highlighted with their activation percentages. Interventions (zeroing one language’s features, adding another’s) shift the output language.

Intervention	Rank vs. clean	NLL/token	$\ \Delta\text{logits}\ _2$
Clean	100%	1.79	—
CLT swap	3%	2.47	1,926
Mean difference ($\alpha = 1$)	150%	14.46	6,946
Random control	20%	1.75	768

Table 1: Averaged across both language pairs. The CLT swap reduces target-token rank far more effectively than magnitude-matched or naive baselines, while preserving fluency.

4.3 Failure Mechanisms in Non-English Languages

Section 4.1 showed that multilingual LLMs develop highly shared representations in their middle layers, indicating that semantically equivalent inputs are processed through partially shared internal mechanisms across languages. Despite this shared organization, multilingual models consistently perform better on English, and even models trained on balanced multilingual distributions retain a persistent English bias. This raises a central question: if multilingual models rely on shared representations, why do multilingual performance gaps persist?

To address this question, we analyze multilingual failure and repair mechanisms using a **model-diffing** approach. Intuitively, the features a model adjusts to improve performance in a language reveal what it struggles with most. Let $\mathcal{M}_0$ denote the base multilingual model and $\mathcal{M}_\ell$ a version finetuned on language ℓ . We finetune the original CLT, trained on activations from $\mathcal{M}_0$, on activations from $\mathcal{M}_\ell$ to identify which features and layers are modified during language-specific repair. We focus on low-performing languages $\ell \in \{\text{Arabic, Chinese, German}\}$ and study how these changes propagate across layers and affect

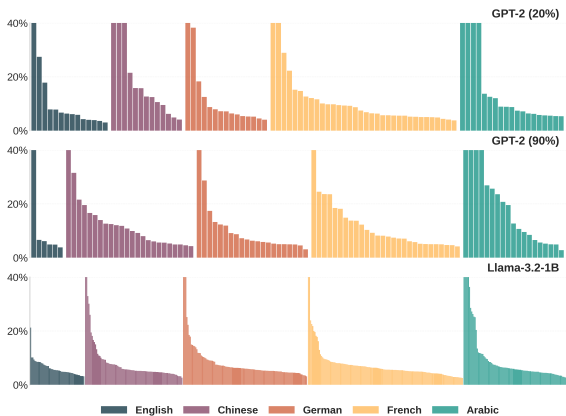


Figure 5: Frequency of token activations over their respective languages for language-specific features in the 20% model.

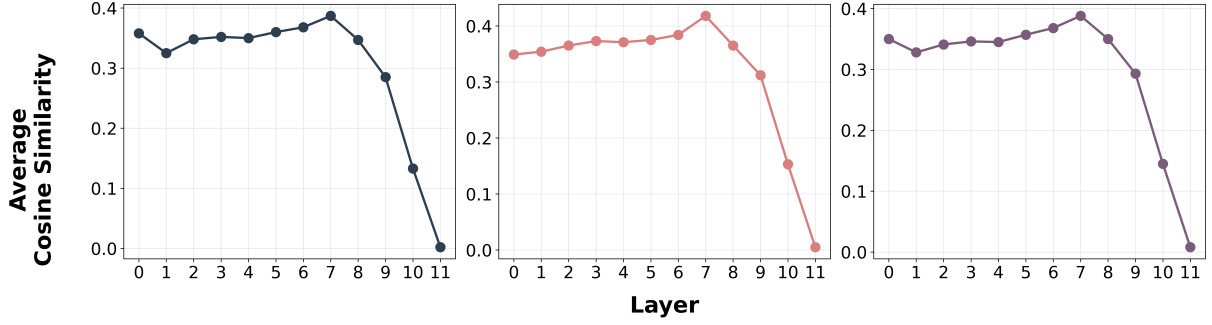


Figure 6: Cosine similarity between $\mathcal{D}_0$ and $\mathcal{D}_\ell$ across layers for Arabic (left), German (middle), and Chinese (right). The largest representation changes consistently occur in the late layers.

token-level representations.

Finetuning Primarily Repairs Late-Layer Decoding Features. For each language ℓ , we compute the cosine similarity between the feature dictionaries of the base and finetuned models across layers.

Figure 6 shows that the largest representation changes consistently occur in the later layers. This aligns with our previous observation that language-specific decoding features are concentrated near the output layers, suggesting that finetuning improves multilingual performance by strengthening late-layer decoding pathways rather than restructuring the shared multilingual representations formed in middle layers. In contrast, middle layers remain comparatively stable across all languages, indicating that the multilingual shared space itself is largely preserved during repair.

Repair Through Multilingual Feature Changes. To characterize these repairs more precisely, we partition features according to cosine similarity into high-change and low-change groups.

Figure 7 shows that high-change features are strongly concentrated in late layers across all languages, consistent with the hypothesis that multilingual failures primarily arise during language-specific decoding.

Beyond this global trend, we observe distinct behaviors across layers. In early layers, modifications primarily affect token-level language alignment, particularly for fragmented scripts such as Chinese and Arabic. These changes appear to improve the mapping between fragmented subword representations and the shared multilingual representations formed in middle layers. In late layers, however, finetuning consistently strengthens target-language features, suggesting that the model performs relatively small but targeted adjustments to

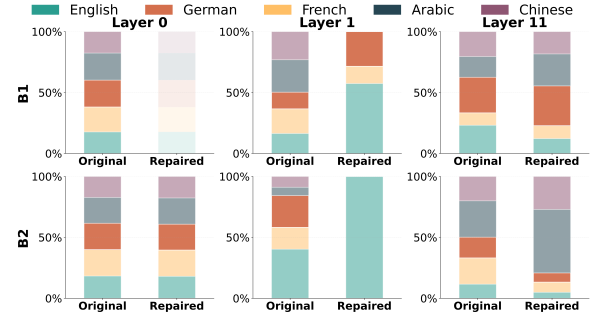


Figure 7: Language distribution comparison across layers for German. Middle layers exhibit high multilinguality and comparatively small language-distribution shifts and are therefore omitted from the figure. Results for Arabic and Chinese are presented in Appendix K.1.

align language-specific decoding with the correct output distribution while preserving stable multilingual representations.

We additionally observe cross-language effects. Finetuning German strengthens certain Arabic features, while Chinese exhibits similar interactions with French and Arabic. These interactions suggest partially overlapping multilingual subspaces where modifications in one language can partially transfer to others. Additional analyses and experimental details are provided in Appendix K.1.

Causal Impact of High-Change Features. To assess the functional role of the high-change features identified in $\mathcal{D}_\ell$, we perform a causal analysis using LogitLens. For a feature f , we compute its top aligned tokens through the decoder and unembedding matrices:

$$\mathbf{t}^*(f) = \arg \max_i [W_U W_{\text{dec}} f]_i. \quad (7)$$

We find that, after finetuning, high-change features exhibit stronger token assembly in early

layers and larger activation norms than in the base model. These results suggest that finetuning strengthens language-specific decoding pathways while simultaneously improving the model’s ability to reconstruct more coherent compositional token representations from fragmented multilingual inputs.

Additional analyses and examples of the corresponding LogitLens interventions are presented in Appendix K.2.

Circuit-Level Illustration. Finally, we analyze more than 40 representative multilingual circuits from Arabic, Chinese, and German (Appendix H.2). These examples show that activating clusters of high-change features improves retrieval of the correct answer, while sub-token fragmentation weakens connections to the most important multilingual clusters.

We further observe that, despite balanced multilingual tokenizer training, English consistently receives cleaner tokenization and stronger connectivity to late-layer decoding circuits. In contrast, languages such as Arabic are frequently fragmented into multiple subwords, forcing early layers to dedicate substantial capacity to word reconstruction before higher-level semantic processing can occur (see Appendix L for a detailed tokenizer fragmentation analysis). Together, these findings provide a mechanistic explanation for persistent multilingual performance gaps despite the emergence of shared multilingual representations.

Null control for CLT re-finetuning drift To verify that the decoder changes identified above reflect genuine language-specific repair rather than generic drift from further training, we compare against a matched null CLT, initialized from the same base checkpoint and re-finetuned on fresh activations from the *unchanged* base model, with identical token budget, batch size, and optimization schedule as the repair CLTs. Across layers 0–10, null cosine to the base CLT remains within 0.921–0.980, while repair cosine falls to 0.333–0.489 for German, 0.300–0.475 for Arabic, and 0.312–0.469 for Chinese—a gap of at least 0.488 in every language, robust to feature-selection criteria and bootstrap resampling (Appendix K.3). Repair drift therefore substantially exceeds ordinary CLT drift, and this separation persists at layer 11 despite a much smaller shared feature set.

As a further control, we verify that this drift is not driven by dead features or by degraded

reconstruction fidelity: filtering out inactive features leaves the layer-11 gap intact, and each repair CLT still reconstructs 69–83% of the target model’s squared activation energy at that layer (Appendix K.4). Repair drift therefore persists independently of feature activity thresholds and does not coincide with reconstruction collapse.

5 Conclusion

We presented a mechanistic study of multilingual processing in LLMs using Cross-Layer Transcoders. Our results show that multilingual models develop highly shared representations in their middle layers while relying on sparse late-layer features for language-specific decoding. These decoding features linearly read out language identity from earlier layers, and targeted interventions on them can reliably redirect the model’s output language.

Our analyses further suggest that multilingual performance gaps arise less from the absence of shared multilingual representations than from weaknesses in language-specific decoding pathways, tokenization fragmentation, and reduced connectivity to late-layer decoding circuits. Together, these findings provide a mechanistic account of how multilingual models transition between shared multilingual computation and language-specific generation, offering new insight into why multilingual disparities persist despite strongly shared internal representations.

Limitations

Model Scale and Architecture. Our full mechanistic analyses primarily rely on CLTs trained on GPT-2 and LLaMA-3.2-1B models, as training faithful CLTs at larger scales remains computationally challenging. However, we complement these analyses with dense-state multilingual alignment probes across 8 models spanning multiple architectures, scales, routing mechanisms, and multilingual training regimes, including BLOOM, OLMo, OLMoE, Apertus, and LLaMA models up to 8B parameters. The consistency of the observed multilingual structure across these settings suggests that our findings are not restricted to small GPT-style architectures. Nevertheless, larger frontier-scale models may develop additional mechanisms or more nuanced multilingual representations that are not fully captured by our current analyses.

Tokenization Analysis. While we identify to-

kenization as a critical factor in performance disparities, our analysis uses a single BPE-based tokenizer trained on balanced data. We do not systematically compare alternative tokenization strategies (e.g., character-level, morpheme-aware, or language-specific tokenizers) that might better preserve semantic information across scripts. The optimal tokenization approach for multilingual models remains an open question.

Reproducibility Constraints. The computational cost of training both base models and CLTs (requiring thousands of GPU-hours) may limit independent replication. While we release our models and CLTs, the specific training dynamics and their sensitivity to random seeds, hardware variations, and implementation details have not been exhaustively characterized.

References

- Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. [Circuit tracing: Revealing computational graphs in language models](#). *Transformer Circuits Thread*.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. [Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- 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.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. *arXiv preprint arXiv:2009.12862*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#). *Preprint*, arXiv:2204.02311.
- 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 ACL 2020*.
- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). *Preprint*, arXiv:1810.04805.
- Sumanth Doddapaneni, Gowtham Ramesh, Mitesh Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2025. [A primer on pretrained multilingual language models](#). *ACM Comput. Surv.*, 57(9).
- Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. [Transcoders find interpretable llm feature circuits](#). *Preprint*, arXiv:2406.11944.
- Nelson Elhage et al. 2021. [A mathematical framework for transformer circuits](#). *Transformer Circuits Thread*.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav

- Chaudhary, and Marc’Aurelio Ranzato. 2019. [The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Michael Hanna, Mateusz Piotrowski, Jack Lindsey, and Emmanuel Ameisen. 2025. circuit-tracer. <https://github.com/safety-research/circuit-tracer>. The first two authors contributed equally and are listed alphabetically.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization](#). *Preprint*, arXiv:2003.11080.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. [The platonic representation hypothesis](#). *Preprint*, arXiv:2405.07987.
- Andrej Karpathy. 2022. NanoGPT. <https://github.com/karpathy/nanoGPT>.
- Guillaume Lample and Alexis Conneau. 2019. [Cross-lingual language model pretraining](#). *Preprint*, arXiv:1901.07291.
- Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. [From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers](#). *Preprint*, arXiv:2005.00633.
- Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. [On the biology of a large language model](#). *Transformer Circuits Thread*.
- Meta AI. 2026. Llama-3.2-1b. <https://huggingface.co/meta-llama/Llama-3.2-1B>. Accessed: 2026-01-29.
- Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. 2025. [Olmoe: Open mixture-of-experts language models](#). *Preprint*, arXiv:2409.02060.
- Chris Olah, Nick Cammarata, Ludwig Schubert, et al. 2020. [Zoom in: An introduction to circuits](#). *Distill*.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. [Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language](#). In *Second Conference on Language Modeling*.
- 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.
- Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Pudupully. 2025. [Romanlens: The role of latent romanization in multilinguality in llms](#). *Preprint*, arXiv:2502.07424.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. [Do multilingual llms think in english?](#) *Preprint*, arXiv:2502.15603.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.
- Gemma Team. 2024a. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Llama Team. 2024b. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- 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*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.

- Shubham Vatsal, Harsh Dubey, and Aditi Singh. 2025. [Multilingual prompt engineering in large language models: A survey across nlp tasks](#). *Preprint*, arXiv:2505.11665.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- 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.
- BigScience Workshop. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#). *Preprint*, arXiv:2211.05100.
- Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Extrapolating large language models to non-english by aligning languages](#).

Supplementary Materials

Table of Contents

PART I: Additional Methodological Details	14
A Multilingual Model Training	14
A.1 Tokenization Strategy	14
A.2 Model Architectures and Training	14
B CLT Training	14
B.1 Training Configuration	14
B.2 Performance	15
C Additional Validation Loss Curves Under Data Imbalance	15
C.1 GPT2 Models	15
C.2 TinyStories Models	15
D Language Feature Analysis	15
D.1 Methodology and Definitions	15
D.2 Feature Overlap Results	17
PART II: Complementary Analyses	17
E Multilingual Score	17
F Shared Multilingual Space Independence from English Dominance	17
G Cross-Lingual Representation Alignment Beyond CLTs	18
PART III: Case Studies	20
H Case Studies: Multilingual Circuits	20
H.1 Multilingual Circuits: U-shaped Entropy	20
H.2 Multilingual Circuits: Failure Modes	21
H.3 Intervention Validation	24
I Extended Task Analysis: Translation and Cultural Prompts	25
I.1 Translation Prompts	25
I.2 Cultural Context Prompts	25
J Why English Performs Better: A Case Study	25
J.1 Cluster Activation Analysis	25
J.2 Sub-tokenization Effects	25
PART IV: Model Diffing and Tokenizer Analysis	26
K Model Diffing: Additional Material	26
K.1 Multilinguality of Repair Features	26
K.2 Token Assembly in Early Layers	26
K.3 Null Control for CLT Re-finetuning Drift	27
K.4 Dead Features and Reconstruction Fidelity	28
L Tokenizer Analysis: BPE Constraints on Multilingual Processing	28
L.1 Vocabulary Allocation: Cannibalization Testing	29
L.2 Morphological Coherence	29
L.3 Agglutinative Languages and Token Fragmentation	30
L.4 Mechanistic Implications	31

AI Assistance.

AI-based tools were used for limited writing refinement, proofreading, and code assistance. All scientific claims, experimental design, analyses, and conclusions were developed and verified by the authors.

A Multilingual Model Training

A.1 Tokenization Strategy

All models employ a unified tokenization approach designed to eliminate tokenization-induced biases. We train a debiased BPE tokenizer on uniformly distributed five-language data, with each language contributing exactly the same number of tokens. The resulting vocabulary contains 119,547 tokens, ensuring balanced representation across languages and writing systems.

Tokenizer Training Protocol

- **Training Data:** Equal token counts per language (20% each)
- **Vocabulary Size:** 119,547 tokens
- **Algorithm:** Byte-Pair Encoding (BPE)
- **Special Tokens:** BOS (beginning of sentence), EOS (end of sentence), PAD (padding), UNK (unknown)

A.2 Model Architectures and Training

A.2.1 Model Architecture Specifications

We train transformer models following two architectures to enable comparisons across model families and scales:

A.2.2 Training Scale and Compute Optimization

Following Chinchilla scaling laws (Hoffmann et al., 2022), we compute optimal training data sizes in Table 4.

$$N_{\text{tokens}} = 20 \times N_{\text{parameters}} \quad (8)$$

A.2.3 Training Implementation

We use the nanoGPT codebase (Karpathy, 2022) for reproducible transformer training and use the hyperparameters presented in Table 5.

Optimization Configuration

Training Infrastructure We train each model on 4×H100 80GB GPUs. Training is distributed using data parallelism with gradient synchronization, and we use mixed precision (FP16) with automatic loss scaling.

A.2.4 Training Dynamics

Figures 11, 12 and 13 show training and validation loss across English data proportions. Higher English proportions reduce English validation loss but can slightly degrade performance on other languages.

A.2.5 Training Procedure and Models Trained

Each architecture is trained on all data mixtures: 20%, 50%, 70% and 90%. Furthermore, we train another model on non-English data (25% for each of Arabic, Chinese, German and French).

For comparative purposes, we also include a pre-trained LLaMA-3.2-1B (Meta AI, 2026) model alongside our custom-trained models.

B CLTs training

B.1 Training Configuration

We report in Table 6, the hyperparameters of the CLT training. We mostly follow Anthropic guidelines (Ameisen et al., 2025). This setup is similar across Tinstories, GPT-2, and Llama-3.2-1B CLTs.

However, for the Llama-3.2-1B CLT, we use less features, as we are limited by GPU size with an expansion factor of around 4. We use torch.float32 for the training as well.

For all CLTs, we stop training when the L0 sparsity coefficient goes below 10 and Explained Variance is beyond 70%. At this point, we save the full model activations and then load them in random chunks from the dataset for training. The CLTs contain roughly $78 \times 768 \times 32 \times 768$ parameters, though for the Llama-3.2-1B CLT we use fewer features (expansion factor = 4) due to GPU memory constraints. Replacement scores are consistently in the range 0.75–0.8, which aligns with expected reconstruction performance. For circuit extraction, attribution graphs are pruned to retain 80% of cumulative node scores and 85% of cumulative edge scores using the circuit-tracing library. Training is conducted over approximately 200M tokens and follows standard CLT guidelines (Ameisen et al., 2025).

Architecture	Layers	Hidden Dim	Attention Heads	Parameters	Context Length
GPT-2	12	768	12	177.6M	1024
TinyStories	4	768	16	68.5M	512

Table 2: Model architecture specifications.

Component	GPT-2	TinyStories
Vocabulary Size	119,547	119,547
Embedding Dimension	768	768
FFN Inner Dimension	3072	3072
Attention Head Dim.	64	48
Dropout Rate	0.1	0.1
Layer Norm	Pre-LN	Pre-LN
Activation Function	GELU	GELU
Position Encoding	Learned	Learned
Context Length	1024	512

Table 3: Detailed architecture hyperparameters.

Model	Tokens	Steps	FLOPs
GPT-2	3.55B	1.73M	3.8×10^{12}
TinyStories	1.37B	670K	5.6×10^{11}

Table 4: Compute-optimal training specs.

We used a context size of 16 tokens to make online activation generation and CLT training tractable across all models, layers, and data mixtures, particularly with 24,576 latent features per layer. This supports the token- and sentence-local phenomena studied here, including feature multilinguality, language-specific decoding, token fragmentation, and short attribution paths.

The context size refers to the sequences used to generate CLT training activations. The present analysis therefore concerns mechanisms expressed within this local context. Features requiring longer dependencies—such as discourse structure, document-level language switching, long-range coreference, or topic persistence—may require CLTs trained with longer contexts.

B.2 Performance

We report the amount of dead features and the performance of the different models (TinyStories, GPT2, Llama-3.2-1B) CLTs in Figures 8, 9 and 10.

C Additional Validation Loss Curves Under Data Imbalance

C.1 GPT2-Models

In Section 4.1, we reported that the validation loss curves for the four non-English languages (Chinese, Arabic, German, French) show competitive performance. Here we show the curves.

We also provide additional results for Chinese. Figure 12 shows the validation loss curves for Chinese. The results mirror the patterns observed for other languages: despite extreme imbalance in the training data, the model maintains competitive prediction performance in Chinese, indicating that the robustness of multilingual circuits extends to typologically distinct languages.

C.2 TinyStories Models

To assess whether the observed generalization dynamics also hold at small scale, we trained models on the **TinyStories** dataset. Figure 13 reports validation loss curves for five languages (English, Arabic, German, French, Chinese). Despite the reduced data scale and synthetic nature of TinyStories, the same trend emerges: performance remains robust across all languages, confirming that the organization of multilingual circuits is a stable property that persists across both large-scale and small-scale training regimes.

D Language Feature Analysis

We analyze whether high-frequency features exhibit language specificity and quantify the extent to which language-specific features are shared across languages. This analysis uses precomputed activation statistics from CLT training with varying English proportions in the training data: 20%, 50%, 70%, and 90% English.

D.1 Methodology and Definitions

For each feature f at layer ℓ , we compute two statistics from the training data. First, the frequency $\text{freq}(f, \ell)$ represents the proportion of tokens for which the feature activates. Second, the language distribution $P(\text{lang} \mid f, \ell)$ captures the probability

Parameter	Value	Parameter	Value
Optimizer	AdamW	Learning Rate	6e-4
β_1	0.9	β_2	0.95
Weight Decay	0.1	Gradient Clipping	1.0
Warmup Steps	2000	LR Schedule	Cosine linear warmup
Batch Size	64	Gradient Accumulation	8 steps
Effective Batch Size	512		

Table 5: Training hyperparameters (horizontal layout for one-column appendix).

Parameter	Value	Parameter	Value
<i>Model Architecture</i>			
Input dimension (d_{in})	768	Latent dimension (d_{latent})	24,576
Expansion factor	32	Context size	16
<i>Training Hyperparameters</i>			
Learning rate	2×10^{-4}	Adam β_1 / β_2	0.9 / 0.999
Batch size (tokens)	1,024	LR warm-up steps	1,000
LR decay steps	3,749		
<i>Loss Coefficients</i>			
L0 coefficient	2.0	Optimal L0	10
Dead penalty coefficient	1×10^{-5}		
<i>JumpReLU Configuration</i>			
Bandwidth	1.0	Initial threshold	0.03
<i>Other</i>			
Seed	42	Precision	float32 (mixed)

Table 6: Training Configuration (horizontal layout for one-column appendix).

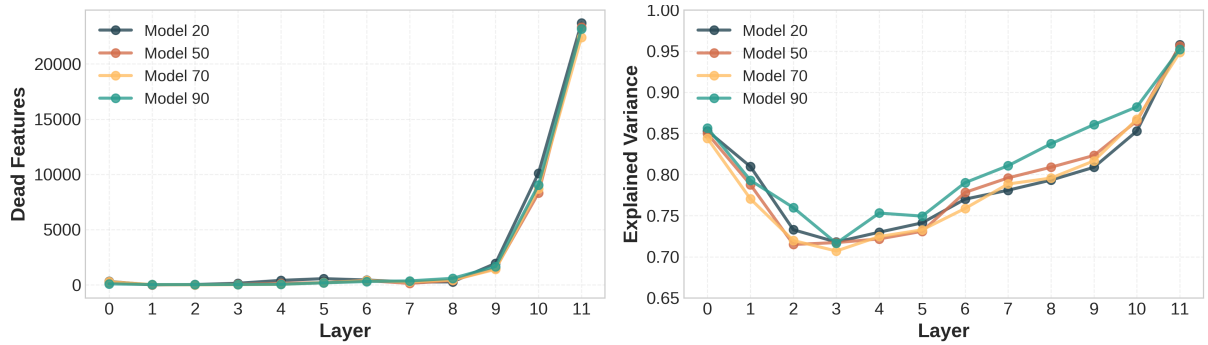


Figure 8: Dead feature count and explained variance across layers for the GPT-2 CLTs.

distribution over languages conditioned on feature activation. Specifically,

$$P(\text{lang} \mid f, \ell) = \frac{N_{\text{seq}}(f, \ell, \text{lang})}{N_{\text{seq}}(f, \ell, \text{total})}, \quad (9)$$

where $N_{\text{seq}}(f, \ell, \text{lang})$ denotes the number of sequences in language lang in which feature f activates at least once.

To label a feature as language-specific, the feature must first be high-frequency, with $\text{freq}(f, \ell) \geq 0.03$. Second, the feature must be statistically associated with a given language, requiring $P(\text{lang} \mid f, \ell) \geq 0.5$, i.e., at least 50% of the sequences on which it activates are from that language. For example, a feature f at layer ℓ is labeled French-specific if both $\text{freq}(f, \ell) \geq 0.03$ and $P(\text{French} \mid f, \ell) \geq$

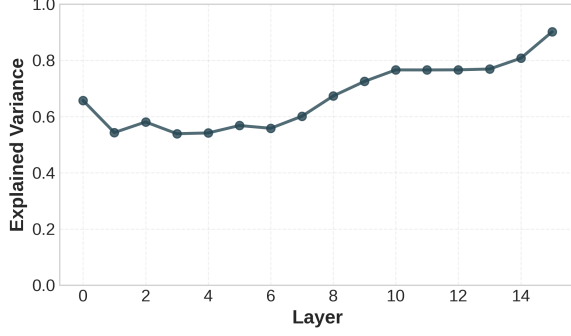


Figure 9: Explained variance over layers for the Llama-3.2-1B CLT.

0.5.

For each language pair, we compute feature overlap via set intersection. Let

$$F_{\text{lang_a}} = \{(\ell, f) : \text{freq}(f, \ell) \geq 0.03 \wedge P(\text{lang_a} \mid f, \ell) \geq 0.5\}. \quad (10)$$

denote the set of Language-A-specific features across all layers, with an analogous definition for $F_{\text{lang_b}}$. We define overlap as

$$\frac{|F_{\text{lang_a}} \cap F_{\text{lang_b}}|}{|F_{\text{lang_a}}|} \times 100\%, \quad (11)$$

D.2 Feature Overlap Results

Figure 14 shows the percentage of each non-English language’s language-specific features that overlap with English-specific features across the four training checkpoints. We observe a consistent increase in overlap as the proportion of English training data increases, rising from approximately 77% at 20% English to 84% at 90% English across all non-English languages. This trend indicates that increasing English exposure during training drives convergence toward an English-dominant feature space.

Figure 15 presents pairwise feature overlap among all non-English language pairs. We find that non-English languages overlap more strongly with each other (averaging 89% at 20% English) than they do with English (averaging 77% at 20% English), yielding a gap of approximately 12 percentage points. However, this gap progressively closes as English training proportion increases, reaching near parity at 90% English, indicating forced convergence.

Figure 16 shows the full pairwise overlap matrix for the 90% English checkpoint as a heatmap. The

visualization reveals notable asymmetries in pairwise relationships. For example, Chinese and Arabic maintain 86.6% overlap, substantially higher than the 78.1% overlap observed between French and German. All pairwise overlaps exceed 78%, confirming the existence of a broadly shared feature space. English exhibits relatively symmetric overlap with all other languages (83–84%), consistent with its role as a convergence point under high training proportions.

We add here two plots related to the language features.

E Multilingual Score

We compare the top100 multilingual score with the multilingual score computed over a large training subset (4M tokens; Figure 19). Both entropy measures show the same overall trend across layers, with the top100 entropy consistently shifted toward lower values. The main difference appears in the earliest layers. This is likely because many features in these layers correspond to single-token activations, especially in layer 0. Given that the total number of features is limited to approximately 28K and multiple languages are involved, some features that mostly activate for a single token also tend to activate for other unrelated tokens.

Figure 19 also shows the frequency of the most activated token for each feature across the top100 sentences where that feature is strongest. We observe that in layer 0, and to a lesser extent in layers 1 and 2, most activations correspond to single tokens. For this reason, we report the top100 score on the graphs.

We also report the general multilingual scores for Tinstories and Llama-3.2-1B in Figure 20.

During graph analysis, we display the multilingual score and the language distributions on a graph visual interface as shown in Figure 21.

F Shared Multilingual Space Independence from English Dominance

To validate that the shared multilingual space structures do not require English dominance, we trained GPT-2 on balanced non-English data (25% French, 25% German, 25% Arabic, 25% Chinese) and computed weighted multilingual entropy across layers. The resulting curve exhibits the identical characteristic rise-and-fall pattern observed in English-dominant models, with peak entropy in layers 5–6 (Figure 22). This demonstrates that the shared mul-

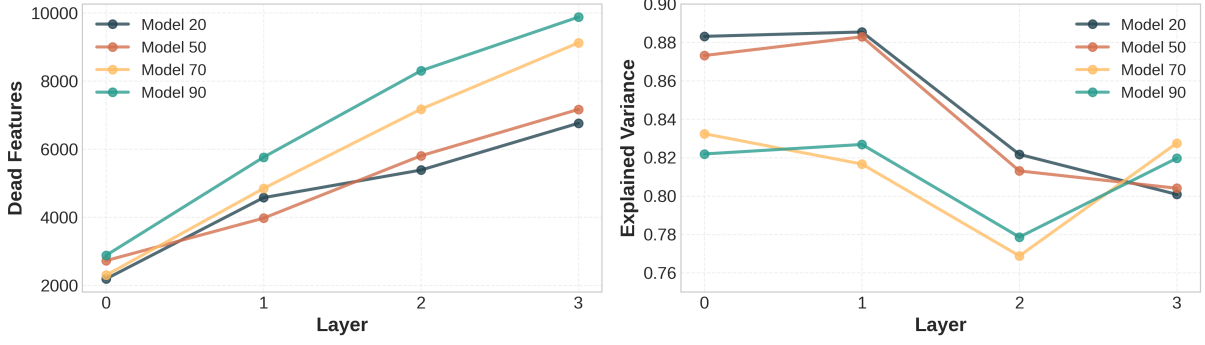


Figure 10: Dead feature count and explained variance across layers for the Tinystories CLTs.

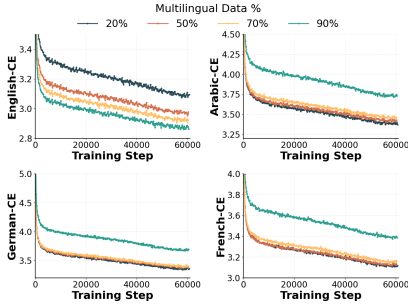


Figure 11: Validation cross-entropy loss curves for models trained on multilingual mixtures with varying proportions of English. Despite extreme imbalance (up to 90% English), the models maintain strong next-token prediction performance across languages.

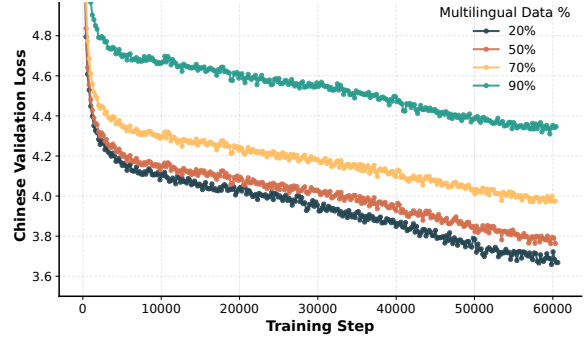


Figure 12: Validation cross-entropy loss curves for Chinese under varying proportions of English. The same robustness pattern holds as for other languages.

tilingual space emergence is a fundamental architectural property, not an artifact of English dominance.

G Cross-Lingual Representation Alignment Beyond CLTs

A potential concern is that the multilingual structures identified in our analyses may arise from artifacts specific to Cross-Layer Transcoders (CLTs) or attribution graph construction. To evaluate whether the observed multilingual organization generalizes beyond CLT-specific analyses, we introduce a complementary cross-lingual alignment probe operating directly on dense hidden-state activations without CLT decomposition.

For each layer ℓ , we compute hidden-state representations for semantically equivalent prompts translated across languages using FLORES-200 (Guzmán et al., 2019). Let $h^\ell(x_i^a)$ denote the representation at layer ℓ for sentence i in language a . We first subtract the global mean representation μ^ℓ to remove language-independent bias:

$$\tilde{h}^\ell(x_i^a) = h^\ell(x_i^a) - \mu^\ell. \quad (12)$$

We then define the alignment gap:

$$\Delta^\ell = \mathbb{E}_{i,a,b} \left[\cos(\tilde{h}^\ell(x_i^a), \tilde{h}^\ell(x_i^b)) \right] - \mathbb{E}_{i,j,a,b} \left[\cos(\tilde{h}^\ell(x_i^a), \tilde{h}^\ell(x_j^b)) \right]. \quad (13)$$

where the second term uses shuffled sentence correspondences to construct a semantic baseline. Positive Δ^ℓ values therefore indicate that semantically equivalent multilingual inputs become more aligned than unrelated cross-lingual pairs. To compare models of different depths, we additionally report

$$\tau^* = \frac{\ell^*}{L}, \quad (14)$$

where $\ell^* = \arg \max_\ell \Delta^\ell$ and L is the number of layers.

We evaluate this probe across eight models spanning multiple architectures, scales, and training regimes: GPT-2/APE, LLaMA/RoPE (Meta AI, 2026), BLOOM/ALiBi (Workshop, 2023), and OL-MoE/MoE (Muennighoff et al., 2025) models ranging from 177M to 8B parameters, including both English-dominant and massively multilingual sys-

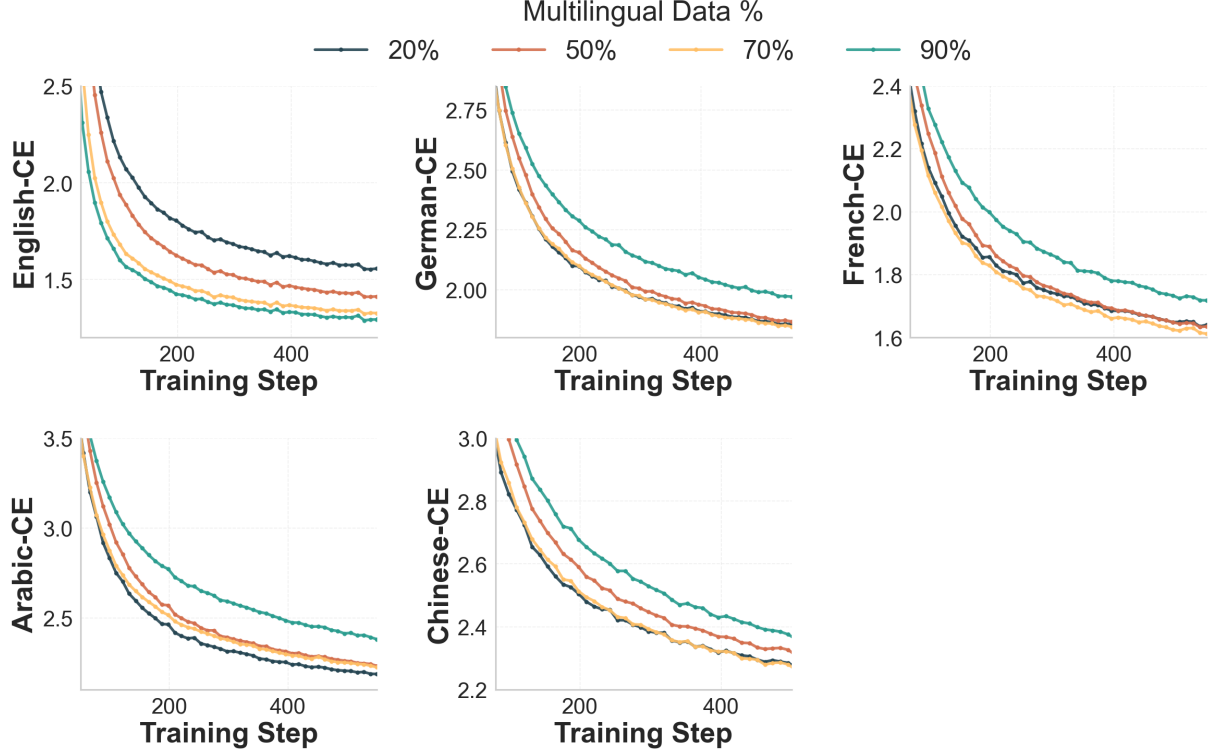


Figure 13: Validation cross-entropy loss curves for the **TinyStories-trained models** across five languages (English, Arabic, German, French, Chinese). Despite small-scale training data, models retain strong multilingual generalization under imbalance.

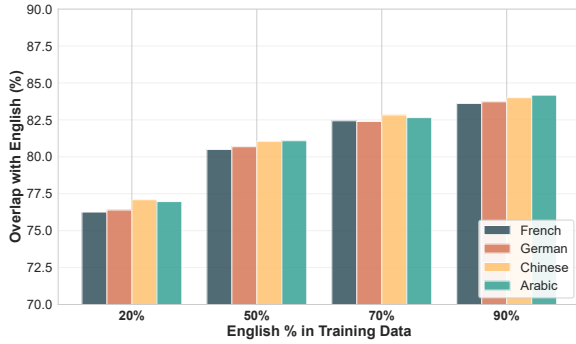


Figure 14: Overlap between non-English language-specific features and English-specific features across training checkpoints.

tems. We additionally extend the probe to 13 languages across 7 typological families and 6 writing systems: Latin (English, French, German, Turkish, Finnish, Swahili), Arabic, Devanagari (Hindi), CJK (Chinese, Japanese), Hangul (Korean), and Cyrillic (Russian, Ukrainian).

Across all models, Δ^ℓ follows a consistent inverted-U structure: alignment increases through the middle layers before partially diverging again near the final decoding layers. Peak alignment consistently occurs around $\tau^* \in [0.44, 0.67]$, closely

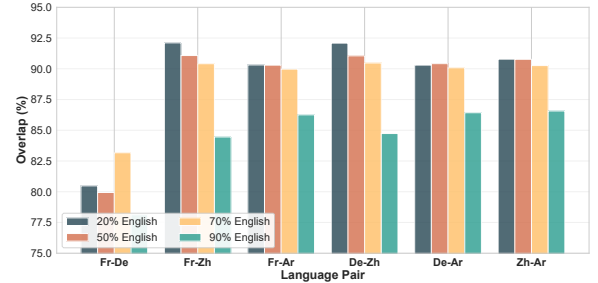


Figure 15: Pairwise overlap of language-specific features among non-English languages.

matching the middle-layer multilinguality peaks identified in the CLT entropy analyses from the main paper. Importantly, this behavior remains stable across architectures, positional encodings, routing mechanisms, and language families, including typologically distant script pairs such as Arabic–Cyrillic and Hangul–Devanagari.

We further observe that peak alignment strength scales with multilinguality. English-dominant models such as GPT-2 exhibit comparatively weak alignment gaps ($\Delta^\ell \approx 0.05$), while massively multilingual models such as Apertus-8B exhibit substantially stronger alignment ($\Delta^\ell \approx 0.52$). Sim-

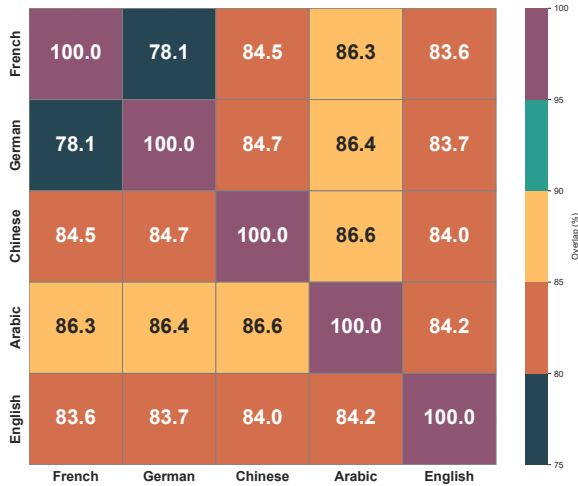


Figure 16: Complete pairwise feature overlap matrix at 90% English training.

ilarly, multilingual training dramatically accelerates alignment formation: BLOOM-7b1 develops a stable middle-layer alignment structure at approximately 2B tokens, whereas English-dominant OLMo-7B requires substantially longer training before a comparable structure emerges.

These results provide independent evidence that the multilingual structures identified in the main paper are not artifacts of CLT sparsification or attribution graphs, but reflect a broader organizational property of multilingual transformer representations. At the same time, CLTs remain essential for mechanistic analysis: while the alignment probe establishes that shared multilingual representations emerge at scale, CLTs identify the sparse features responsible for these structures, trace attribution pathways across layers, enable causal interventions that redirect output language, and characterize multilingual failure modes such as tokenizer fragmentation in Arabic.

Finally, we note that the alignment probe alone fails to recover clear inverted-U behavior in several small GPT-2 models where the CLT entropy analysis remains strongly informative. This suggests that sparse mechanistic decompositions may reveal organizational structure even when dense representation similarity metrics remain comparatively weak.

H Case Studies: Multilingual Circuits

Throughout this section, we present circuits to validate and visualize the findings discussed in the main text. Specifically, we organize our presentation as follows:

i) Multilingual validation: We show circuits across additional tasks that support the multilingual shared space finding described in Section 4.1.

ii) Failure cases and interventions: We provide examples of model failures and demonstrate that interventions on circuits—primarily those that strengthen existing internal features—can successfully guide the model to produce correct outputs. This completes Section 4.3

iii) Translation and cultural prompts case study: We further validate our results with a case study on LLaMA-3.2-1B, using translation and culturally contextualized prompts, which confirms the robustness of our findings.

All examples produce clusters of circuits, which are manually constructed from features whose auto-interpretability descriptions and top-activating examples suggest similar functions. We further validate these clusters by intervening on them and confirming that the observed effects align with their interpreted descriptions.

H.1 Multilingual Circuits: U-shaped Entropy

In this appendix, we provide extended circuit visualizations complementing the analysis of Section 4.1. For each example, we extract attribution graphs across layers and present the identified circuits. Depending on the setting, each figure contains either (i) four subfigures, corresponding to the four training data mixtures (90%, 70%, 50% or 20% English), or (ii) five subfigures, corresponding to the five languages (English, Arabic, German, French, Chinese).

As observed in our main results, early and late layers predominantly encode language-specific circuits, while middle layers form clusters with high language entropy, reflecting multilingual representations. Exceptions arise for determinants and prepositions, whose circuits remain largely language-specific.

The top-right corner of each cluster shows its entropy score, defined as the average of the entropy values of the features it contains.

We now detail the examples analyzed.

H.1.1 Preposition Sentences (English: “It was a piece”)

We begin by analyzing English preposition prediction across training mixtures. Prepositions, being primarily functional rather than semantic, offer insight into the model’s handling of syntactic struc-

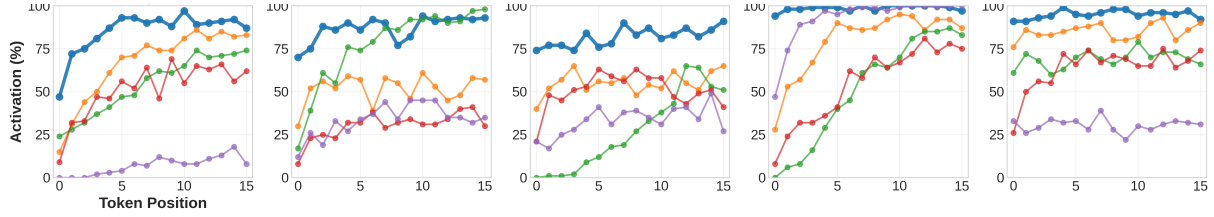


Figure 17: Activation rate over the top100 sequences of the language feature for the 20% model. We see that some language features are more active towards the end of the sequence. Highlighted is the language feature with the highest frequency.

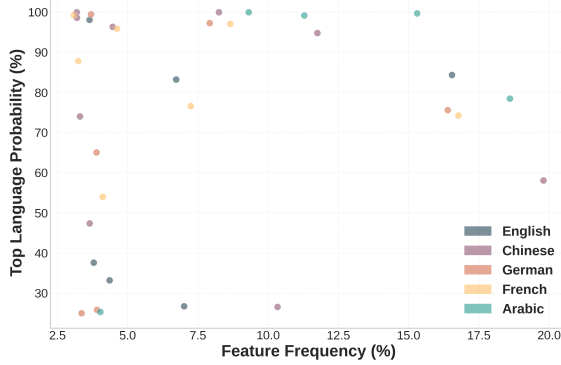


Figure 18: CLT features with the token activation frequency above 5% for the 20% model vs the probability in their top activating language. This shows that most high-frequency features are language features.

tures. In this case, we find that the model consistently relies on language-specific clusters.

H.1.2 Preposition Sentences (French: “J’ai bu une tasse”)

We next examine French preposition prediction across the training mixtures. This setting tests whether the patterns observed for English prepositions generalize to another language.

H.1.3 Content Word Sentences (English: “I prefer drinking tea to drinking”)

Next, we analyze content word prediction across languages in the 90% English mixture. Content words engage semantic and contextual representations, making them a strong testbed for cross-lingual generalization. We find robust multilingual clusters in middle layers across all five languages, even when trained under extreme imbalance.

H.1.4 Content Word Sentences: “Winter, spring, summer, and autumn are the four”

We analyze the prediction of content words in the sentence “Winter, spring, summer, and autumn are

the four” across languages in the 20% English mixture. This example probes semantic and sequential reasoning across multiple languages, highlighting whether the model reuses internal circuits for similar content words. We find that middle layers consistently form multilingual clusters across all five languages, while early and late layers remain largely language-specific.

H.1.5 Calendar Term Prediction (“Monday, Tuesday, Wednesday, Thursday”)

We study calendar term prediction across mixtures. Calendar terms primarily involve lexical memorization, providing a probe for multilingual circuit reuse. Across all mixtures, middle layers consistently form multilingual clusters, while early and late layers remain more language-specific.

H.1.6 Analogy Sentences (“the opposite of ‘men’ is ...”)

Finally, we analyze analogy-based prediction across mixtures. Unlike lexical recall, analogies engage relational reasoning. We find that the same multilingual clusters are reused across mixtures, especially in middle layers, underscoring their role in cross-lingual generalization.

H.2 Multilingual Circuits: Failure Modes

As discussed in Section 4.3, models trained on balanced data across languages still exhibit lower performance in non-English languages compared to English. To further support the findings from Section 4.3, we examine two tasks that probe distinct types of reasoning:

1. **Antonym Task:** Prompts such as “The opposite of ‘men’ is ‘’” test semantic relations.
2. **Category Task:** Prompts like “Football, cycling, baseball are all” probe conceptual grouping.

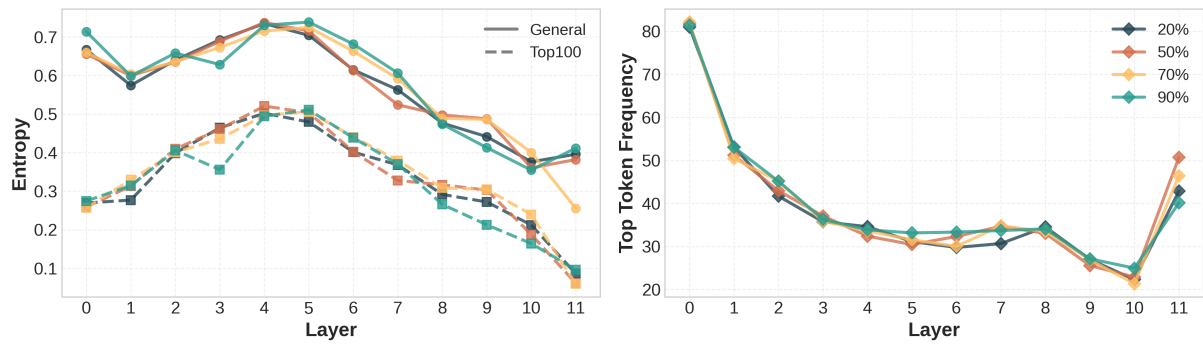


Figure 19: Comparison between the two multilingual entropy measures. One computed on the top100 activated sequences for each feature and one computed over a large training subset. On the right the frequency of activation of the most frequent token over the top100 sequences.

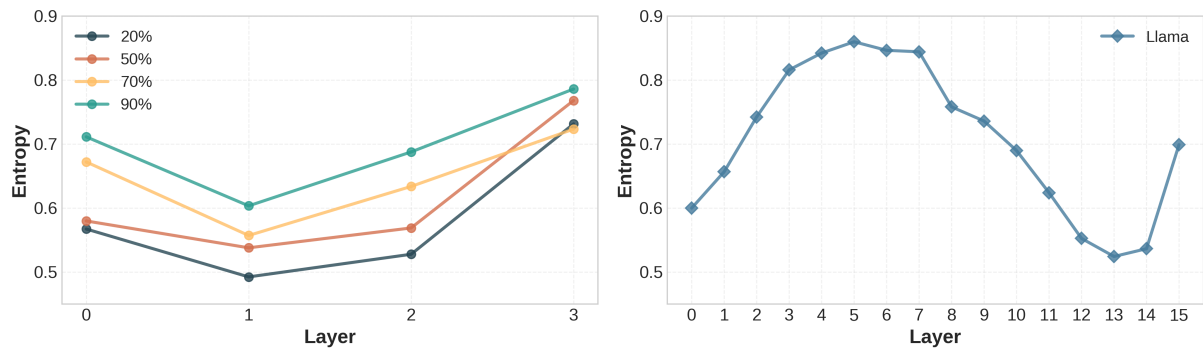


Figure 20: General multilingual score for the Tinstories model and Llama-3.2-1B. The final layer jump in the Llama-3.2-1B CLT is potentially due to training dynamics. We find that the last layer has some features activated with small values that impact slightly the reconstruction performance but have a great impact on the multilingual score.

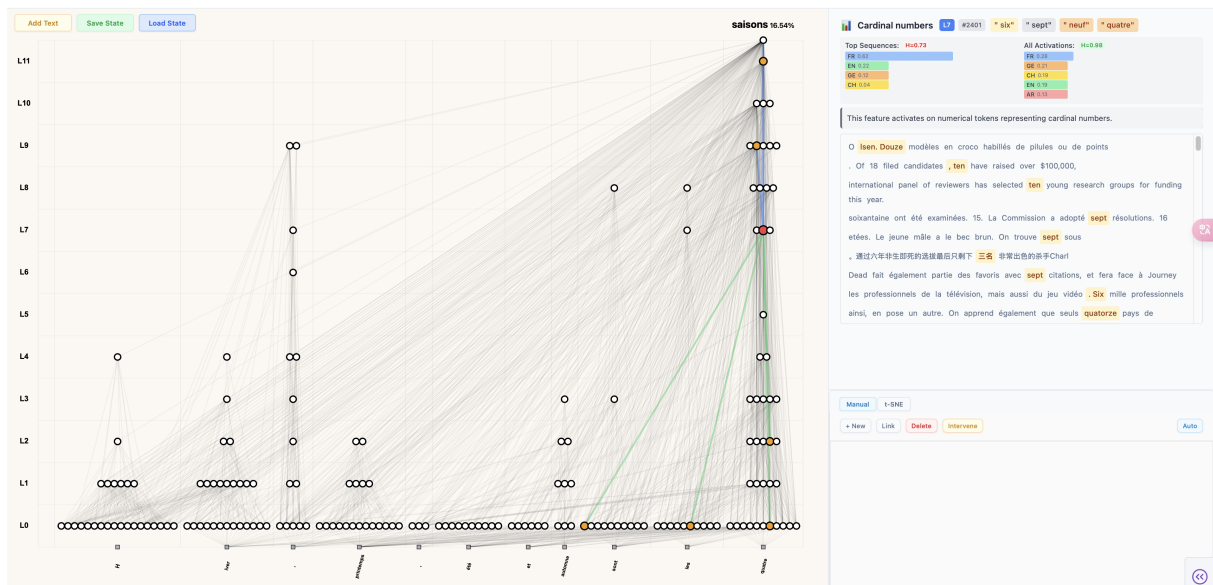


Figure 21: Graph visual interface with multilingual distributions and entropies on the top right.

We analyze five languages (English, German, French, Arabic, Chinese) across four training mixtures (20%, 50%, 70%, 90% English dominance) to uncover the mechanisms behind non-English failures.

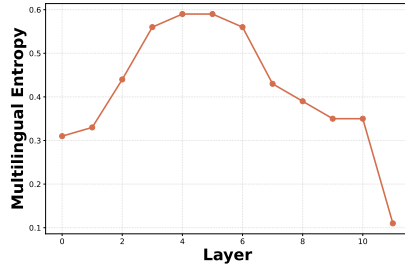


Figure 22: The Shared Multilingual Space emergence is independent of English dominance. Model trained without English (FR/DE/AR/ZH, 25% each) shows identical entropy curve to English-dominant model (20% EN + others), with characteristic rise-and-fall pattern and peak in layers 5–6.

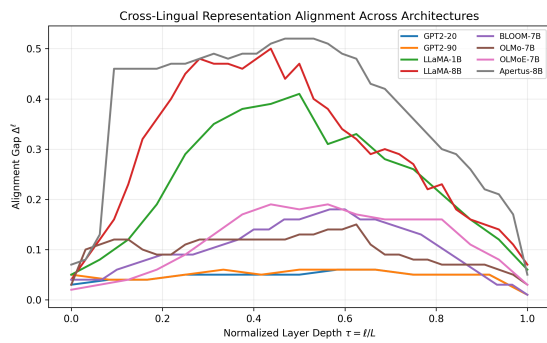


Figure 23: Cross-lingual alignment gap Δ^ℓ across normalized layer depth $\tau = \ell/L$ for models spanning multiple architectures, scales, routing mechanisms, and multilingual training regimes. All models exhibit a consistent inverted-U structure, with alignment peaking in middle layers before partially diverging near the decoding layers. The effect remains stable across GPT-2/APE, LLaMA/RoPE, BLOOM/ALiBi, and OLMoE/MoE architectures.

H.2.1 Antonym Task

Circuits driving antonym prediction are consistent across languages and mixtures, but their completeness depends on mixture balance. In Arabic under the 90% English mixture, the Men&Women cluster, present in most other languages, is missing but appears from the 50% mixture onward (Figure 30). Intervention experiments (Figure 35) confirm its causal role: adding this missing cluster restores correct Arabic predictions (Appendix H.3).

90% English Mixture Figure 31 shows circuit graphs for the antonym prompt “The opposite of “men” is “” across all five languages.

The circuits reveal why English achieves superior performance: languages that fail to predict the correct antonym consistently lack the Men&Woman cluster. Arabic shows particularly sparse connec-

tivity, missing most multilingual clusters found in other languages due to limited circuit sharing with English.

70% English Mixture Figure 32 shows identical patterns: the Men&Women cluster remains absent where models fail. Only English sufficiently activates the reasoning circuit at this mixture level.

50% English Mixture Figure 33 validates previous patterns while showing increased presence of Men&Women clusters in non-English languages, reflecting the shift toward more uniform language distribution.

20% English Mixture

20% English Mixture Figure 34 shows more Men&Women clusters and improved performance, especially for Latin-script languages. Even at this balanced mixture, non-Latin script languages still demonstrate weaker circuit formation.

The circuit patterns suggest two competing mechanisms: a reasoning circuit that correctly predicts Women and a copying circuit that simply reproduces Men from the input tokens.

To validate these hypotheses, we conduct four investigations:

1. Intervention experiments: scale up reasoning circuit activation (Men&Women cluster) while scaling down copying circuit activation (Men cluster)
2. Comparative analysis across the three remaining mixtures to identify general trends
3. Activation value analysis of reasoning and copying circuits to validate insufficient activation hypothesis
4. Investigation of why English, German, and French demonstrate superior performance across all mixtures

H.2.2 Category Completion Task

We analyze the Category Completion task using the 20% mixture to understand performance in the balanced setting. Arabic performs poorly due to tokenization fragmentation, which forces early layers to reconstruct words rather than activate answer-related clusters. Intervention experiments (Figure 36) show that scaling these clusters restores correct predictions, emphasizing the importance of activation strength.

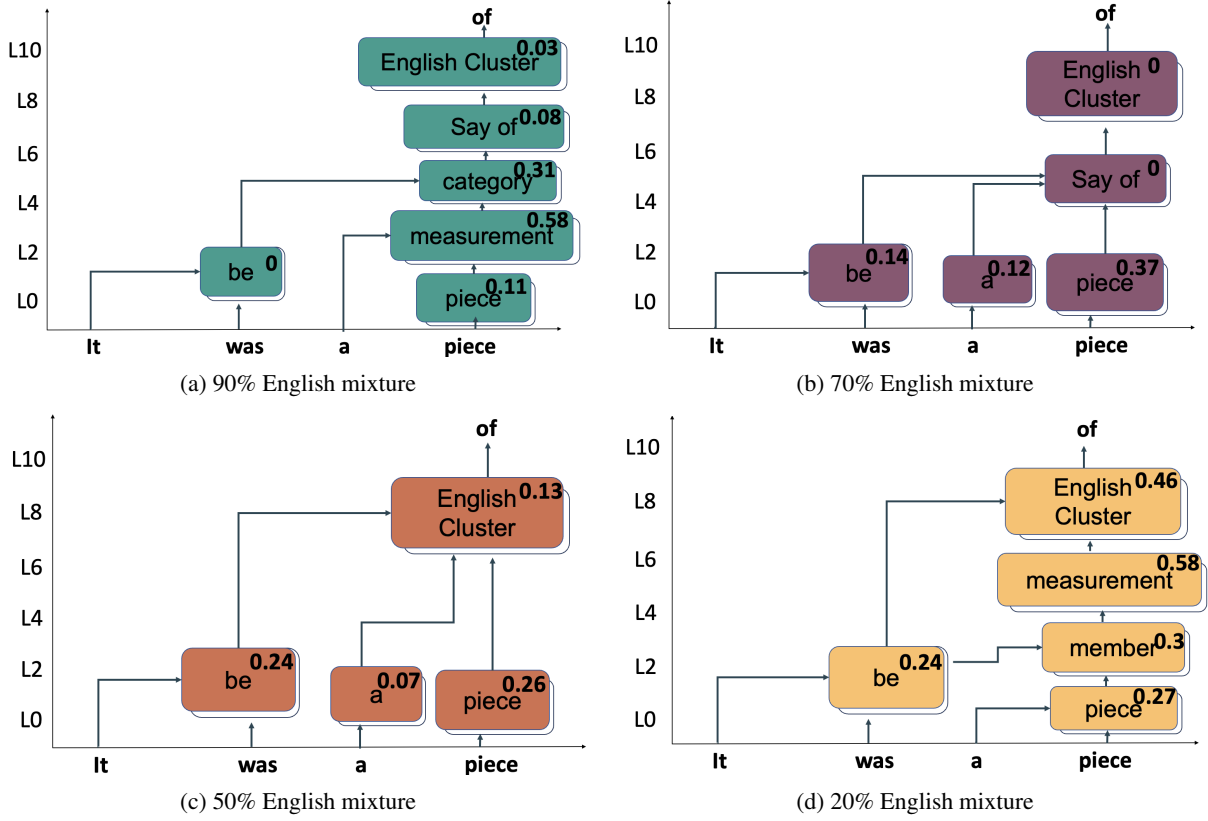


Figure 24: Circuits for English preposition prediction (“It was a piece”) across training mixtures.

We present the circuits of the Category Completion Task presented in Figures 28.

We also present the intervention results for Arabic in Figure 36, the only language where the model failed.

Comparatively, English consistently triggers stronger responses in task-specific circuits (Men&Women for Antonym, Say Sports for Category Completion). This difference narrows under balanced mixtures but remains linked to sub-tokenization: languages with fragmented tokens show weaker edge strengths from embeddings to target clusters, partially explaining their reduced accuracy (Appendix J).

We further validate these mechanisms through more complex tasks (translation and cultural prompts on Llama-3.2-1B), tokenizer analysis, and quantitative validation across the attribution graphs (see Appendix I).

H.3 Intervention Validation

We define an **intervention on a cluster of circuits** as steering the residual stream using a vector constructed from the cluster’s features. Specifically, let $f_1, f_2, \dots, f_n$ denote the features in a cluster, and let $v(f_i)$ be the decoded vector of feature f_i . The

intervention vector v_{int} is then given by

$$v_{\text{int}} = \alpha \sum_{i=1}^n v(f_i), \quad (15)$$

where α is a scaling factor determined via sweeping to optimize the effect of the intervention. The residual stream is updated by adding v_{int} at the relevant layer.

Figures 35, 38 and 36 demonstrates successful interventions that validate the competing circuits hypothesis.

In all the examples where the model fails to get the right answer, scaling up the reasoning circuit while scaling down the copying circuit consistently produces correct predictions. We sweep positive coefficients for the reasoning circuit (1 to 30) and negative coefficients for the copying circuit (-30 to -1) to determine optimal steering values.

Figures 34, 33, 32, and 31 further validate our hypothesis. In particular, most of the examples where the model fails to predict the correct answer show either an absence of the reasoning circuit or a low level of its activation, and this pattern becomes less pronounced as we move towards the 20% mixture. Taken together with the intervention

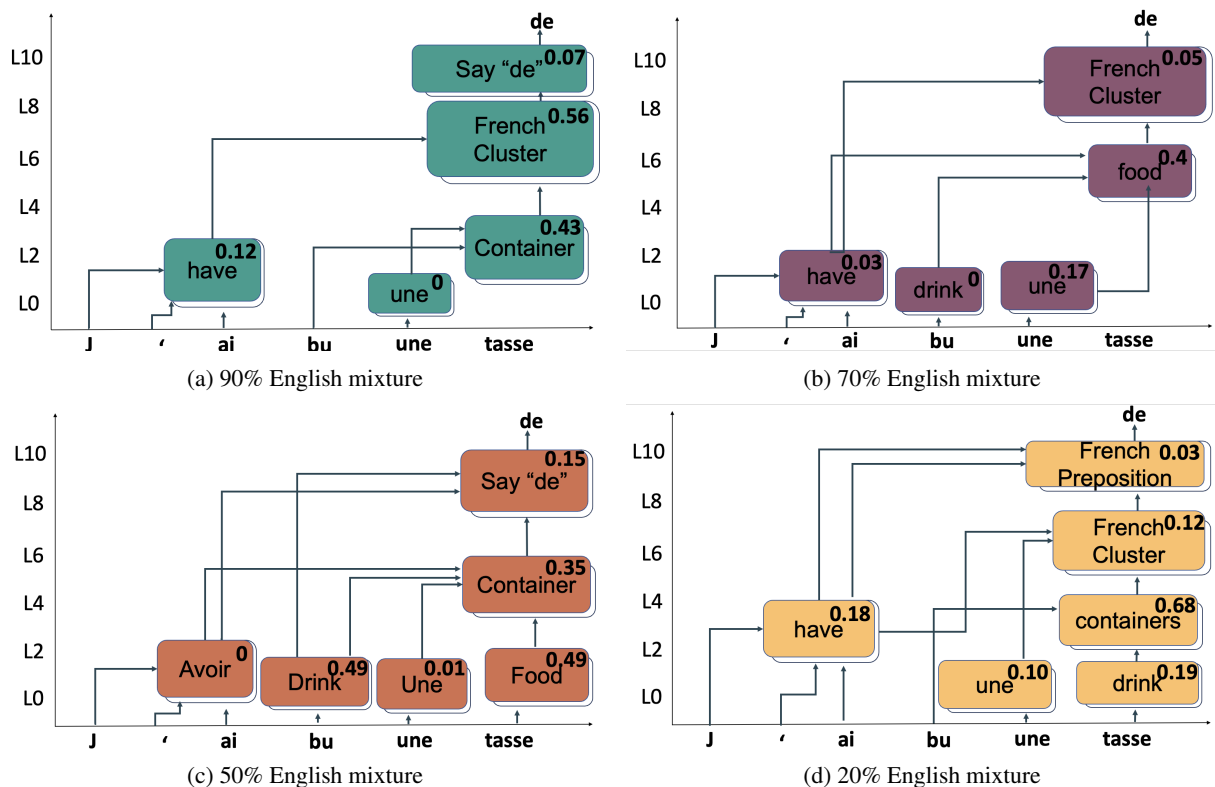


Figure 25: Circuits for French preposition prediction (“J’ai bu une tasse”) across training mixtures.

results, this supports our hypothesis about circuit competition.

I Extended Task Analysis: Translation and Cultural Prompts

We tested whether the shared multilingual space mechanisms generalize beyond simple lexical tasks (antonym, category completion) to more complex, context-dependent tasks on LLaMA-3.2-1B (Meta AI, 2026). Attribution analysis across all complex task circuits reveals the consistent three-phase structure: early layers extract language-specific features, middle layers form language-neutral semantic representations, and late layers activate language-specific output clusters.

I.1 Translation Prompts

Translation tasks require cross-lingual semantic mapping without explicit language labels:

- “English: Flower, Français: ____”
- “English: Flower, Arabic: ____”

Circuits maintain the same rise-and-fall pattern across layers, despite task complexity.

I.2 Cultural Context Prompts

Models complete: “The best dish is...” and generate language-appropriate food answers.

The same multilinguality pattern emerges across these circuits (see Figure 40).

J Why English Performs Better: A Case Study

We present a case study examining the 20% mixture to understand performance disparities in balanced training.

J.1 Cluster Activation Analysis

We measure activation strength of task-relevant clusters across languages. For the Antonym task, we track the Man&Woman cluster; for Category Completion, the Say Sports cluster.

Figure 41 shows activation patterns. English consistently produces higher activations than other languages. This disparity decreases approaching the 20% mixture, suggesting balanced training mitigates but does not eliminate the advantage.

J.2 Sub-tokenization Effects

We quantify sub-tokenization impact by measuring edge strength from token embeddings to target

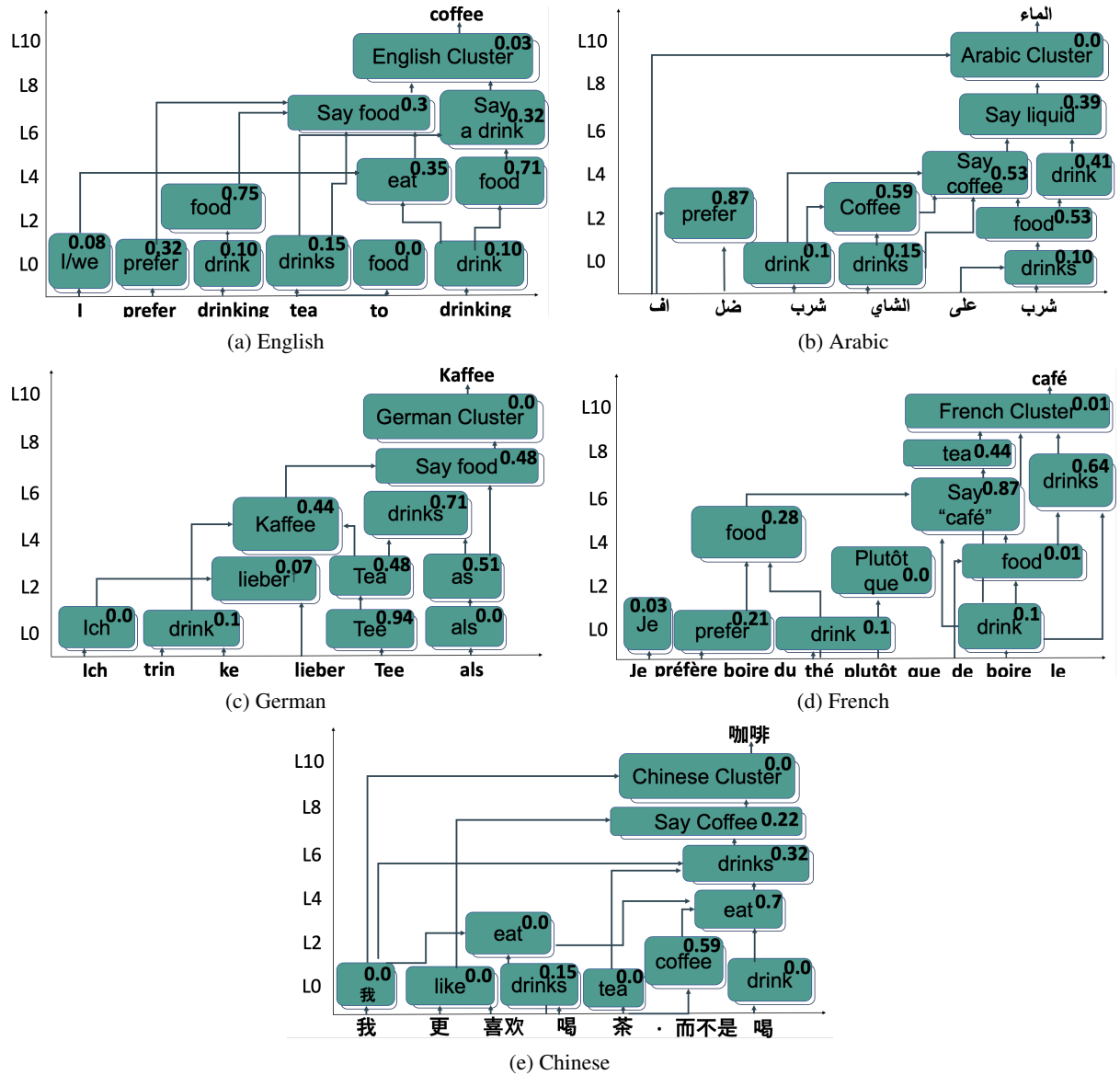


Figure 26: Circuits for content word prediction (“I prefer drinking tea to drinking”) across languages in the 90% English mixture.

clusters. Edge strength is the dot product between the residual stream after token embedding and the cluster’s input direction.

Figure 42 presents results across languages with varying sub-tokenization rates. Highly sub-tokenized languages (e.g., Arabic) show significantly weaker edge strengths. Fragmented tokens fail to properly activate downstream circuits—semantic information distributes across multiple sub-tokens rather than concentrating in a single embedding.

K Model Diffing: Additional Material

K.1 Multilinguality of Repair Features

As mentioned in Section 4.3, we estimated the distribution of languages for features in each of our buckets. We evaluate on 10000 tokens and we report the scores in Figure 43

K.2 Token Assembly in Early Layers

Token assembly refers to the process by which early layers reconstruct sub-tokenized inputs (particularly fragmented words from morphologically rich languages like Arabic) into coherent higher-level representations before semantic processing. Logit Lens results, as described in Section 4.3, are shown in Figure 44. As observed, finetuning improves

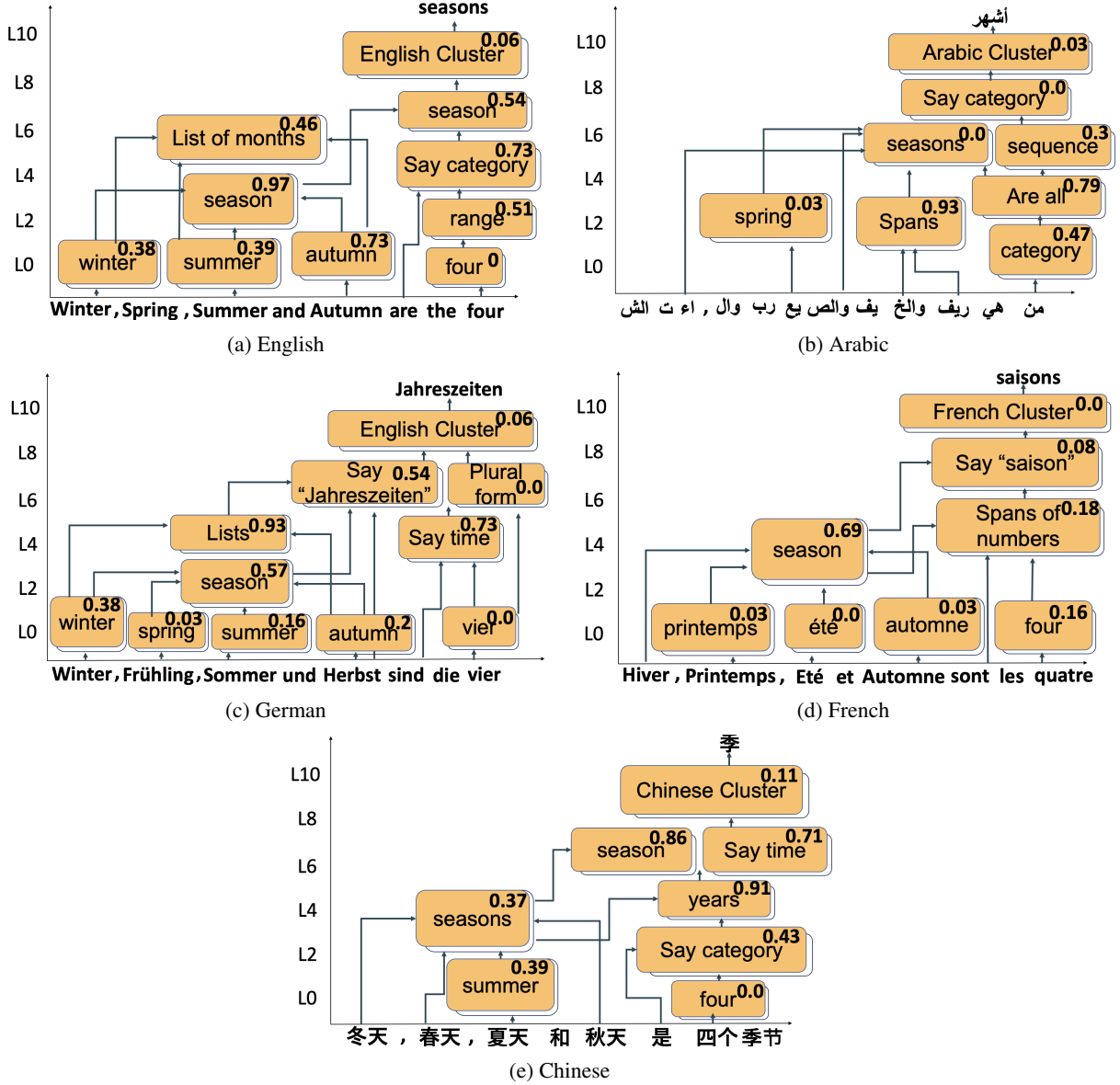


Figure 27: Circuits for content word prediction (“Winter, spring, summer, and autumn are the four”) across five languages in the 20% English mixture.

word assembly in Arabic and German. Counter-intuitively, it increases fragmentation in Chinese, likely due to the comparatively high-rank updates required for Chinese. This suggests that the model relies more on memorization than on structured rules or heuristics for Chinese, as learning this language consumes a larger portion of the model’s capacity, leaving less capacity available for other languages.

K.3 Null Control for CLT Re-finetuning Drift

We evaluate all CLTs on the same 500,000 held-out positions (100,000 each in Arabic, Chinese, German, English, and French). Each repair CLT is evaluated on activations from its corresponding

repaired model, and null data are drawn from a held-out corpus. For each language, we compute index-aligned decoder cosine similarity to the base CLT, restricted to features active in the base, null, and repair CLTs.

The smallest null-minus-repair gaps are 0.488, 0.502, and 0.508 for German, Arabic, and Chinese, respectively, with minimum lower bounds of 0.485, 0.499, and 0.505 over 2,000 bootstrap resamples. Repair drift exceeds null drift for at least 99.4% of shared active features in every language, indicating that the divergence is not an artifact of a small subset of features. For German, this separation is robust to feature-selection criteria: the smallest gap is 0.339 using all features, 0.474 under pairwise

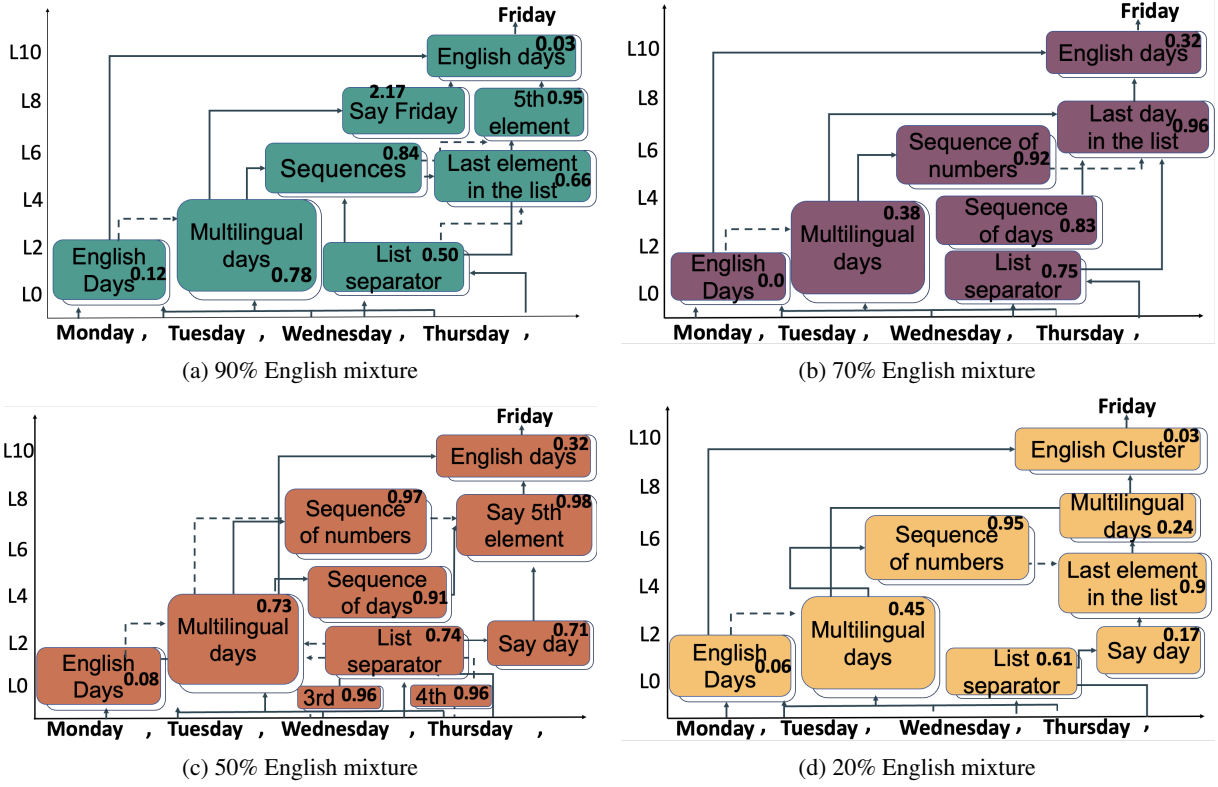


Figure 28: Circuits for calendar term prediction (“Monday, Tuesday, Wednesday, Thursday”) across training mixtures.

activity filtering, and 0.488 under the three-way mask.

Layer 11 has substantially fewer shared features (287–345 per three-way intersection), but null and repair CLTs remain clearly separated (Table 7). We interpret the exact layer-11 magnitudes cautiously given the smaller feature count, but the broader localization is well supported by the more stable layers 9–10.

K.4 Dead Features and Reconstruction Fidelity

We defined a feature as dead when it never exceeded its JumpReLU threshold, and removed features dead in either member of each base–repair pair. Final-layer cosine remains low after this filtering: 0.124 for German (643 retained features), 0.132 for Arabic (507 retained features), and 0.171 for Chinese (412 retained features).

For German, requiring at least 1, 5, 10, and 50 firings raises layer-11 cosine from 0.124 to 0.384, 0.448, and 0.504, respectively. Layers 9–10 remain more stable under the same filtering:

- Layer 9: 0.350, 0.352, 0.349, 0.351
- Layer 10: 0.287, 0.314, 0.315, 0.342

Arabic and Chinese show the same pattern. Rare low-activity features affect the precise layer-11 magnitude, whereas localization across layers 9–10 is robust.

Using each CLT with its matched fine-tuned model, we also computed

$$1 - \text{NMSE} = 1 - \frac{\sum \|y - \hat{y}\|_2^2}{\sum \|y\|_2^2}. \quad (16)$$

At layer 11, this score is 0.827 for German, 0.761 for Arabic, and 0.690 for Chinese: the CLTs reconstruct 69–83% of the squared target-activation energy.

Thus, repair drift substantially exceeds ordinary CLT drift, persists after removing dead features, and does not coincide with reconstruction collapse. We qualify the exact layer-11 effect while retaining the robust late-layer localization.

L Tokenizer Analysis: BPE Constraints on Multilingual Processing

Despite employing a debiased multilingual BPE tokenizer trained on balanced data, morphologically rich languages exhibit severe sub-tokenization compared to other languages. We isolate the source of

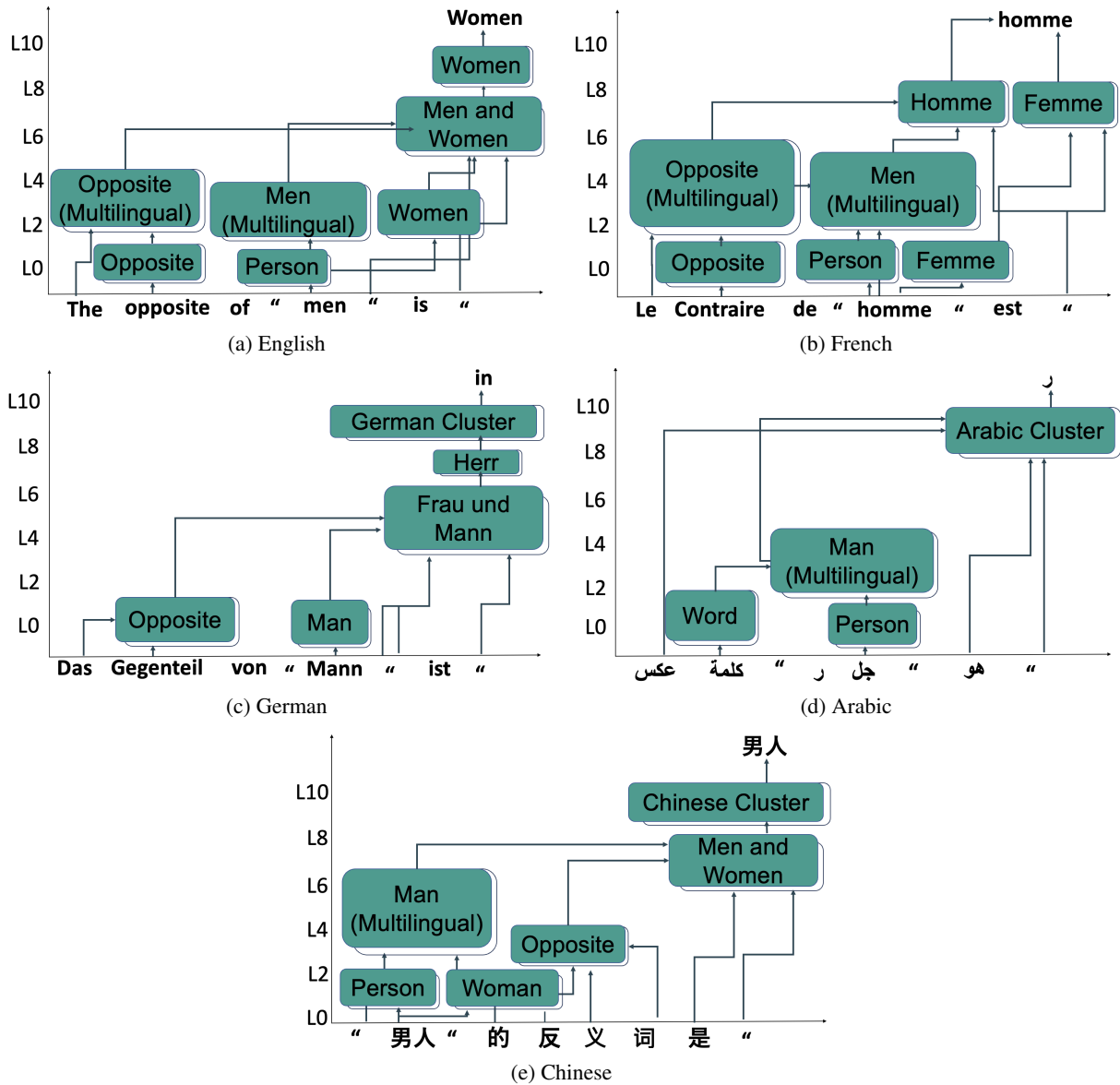


Figure 29: Circuits for analogy-based prediction (“the opposite of ‘men’ is ...”) across languages in the 90% English mixture.

this disparity through three complementary analyses.

L.1 Vocabulary Allocation: Cannibalization Testing

To rule out allocation bias, we computed fertility ratios (tokens per word) for language-specific tokenizers:

- Arabic: 1.06
- Chinese: 1.03
- English: 0.99
- French: 0.96

- German: 1.00

All ratios cluster near 1.0, confirming that the de-biased tokenizer allocates vocabulary fairly across languages. The disparity in fragmentation therefore stems from linguistic properties, not allocation bias.

L.2 Morphological Coherence

Token-morpheme coherence (how well token boundaries align with morpheme boundaries) diverges dramatically:

Concrete examples illustrate this fragmentation:

- *al-radar* (radar) fragments into [*al-r*, *ad*, *ar*]: prefix, middle, and suffix separated

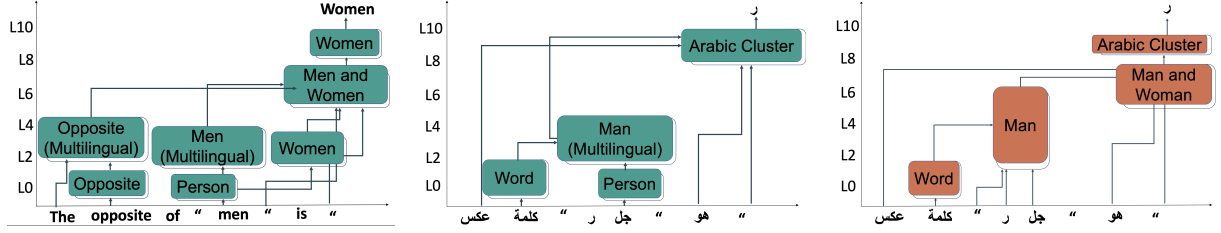


Figure 30: Circuit structures across language mixtures. Green: 90% English mixture; orange: 50% English mixture. Key clusters missing in Arabic at 90% emerge by 50%, indicating mixture-dependent circuit formation. Additional examples appear in Appendix H.2.

Language	Null	Repair	Gap (95% CI)
German	0.497	0.237	0.260 [0.229, 0.291]
Arabic	0.558	0.229	0.329 [0.291, 0.367]
Chinese	0.524	0.238	0.286 [0.247, 0.325]

Table 7: Layer-11 decoder cosine similarity to the base CLT, null vs. repair, on the three-way feature intersection.

Metric	Arabic	English
Token Fertility	1.97	1.53
Token-Morpheme Coherence	0.10	0.42
Meaningless Fragments (%)	89	58

Table 8: Arabic exhibits 4× lower morphological coherence than English despite fair vocabulary allocation. Meaningless fragments represent tokens that carry no semantic content.

- *tafaqad* (inspect) fragments into *[taf, qad]*: broken across morpheme boundaries with subtokens having no meaning.

English maintains coherence: “reading” → [“read”, “ing”], preserving morpheme structure.

L.3 Agglutinative Languages and Token Fragmentation

We conducted an additional experiment on a balanced Arabic–Chinese–Korean–Japanese GPT-2 model. This adds Korean and Japanese, two non-Latin-script languages that are predominantly agglutinative. Arabic provides a comparison with the strongest fragmentation pattern in the original experiments, while Chinese is a descriptive lexical comparison because its Jieba units are words rather than morphemes.

We evaluated 2,000 held-out FineWeb-2 documents per language and identified linguistic units using CAMEL Tools for Arabic, Kiwi for Korean, UniDic/fugashi for Japanese, and Jieba for Chinese. We distinguish:

- **Within-unit transitions:** the target BPE token begins strictly inside an analyzer-derived unit.
- **Boundary-aligned transitions:** the target begins at a boundary between analyzer-derived units.

We sampled 10,000 transitions per condition and language. Using LogitLens directly on the model’s residual stream, we define token-recovery gain as:

$$G_{i,l} = z_{i,l}(t_i) - z_{i,\text{emb}}(t_i) \quad (17)$$

and report:

$$\Delta_l = \mathbb{E}[G_{i,l} \mid \text{within}] - \mathbb{E}[G_{i,l} \mid \text{boundary}] \quad (18)$$

A negative value means that within-unit tokens gain less target-token logit than boundary-aligned tokens.

Tokenization statistics

Token-recovery results The intervals use 2,000 document-level bootstrap replicates. Every contrast direction was reproduced with an independent sampling seed. Outcome-blind matching on embedding-layer logit, token frequency, token length, and context position retained the early Arabic result, the Korean result in layers 1–5, and the opposite Japanese result.

Korean shows the same early-layer pattern as Arabic: BPE tokens that begin inside morphological units are recovered more slowly than tokens

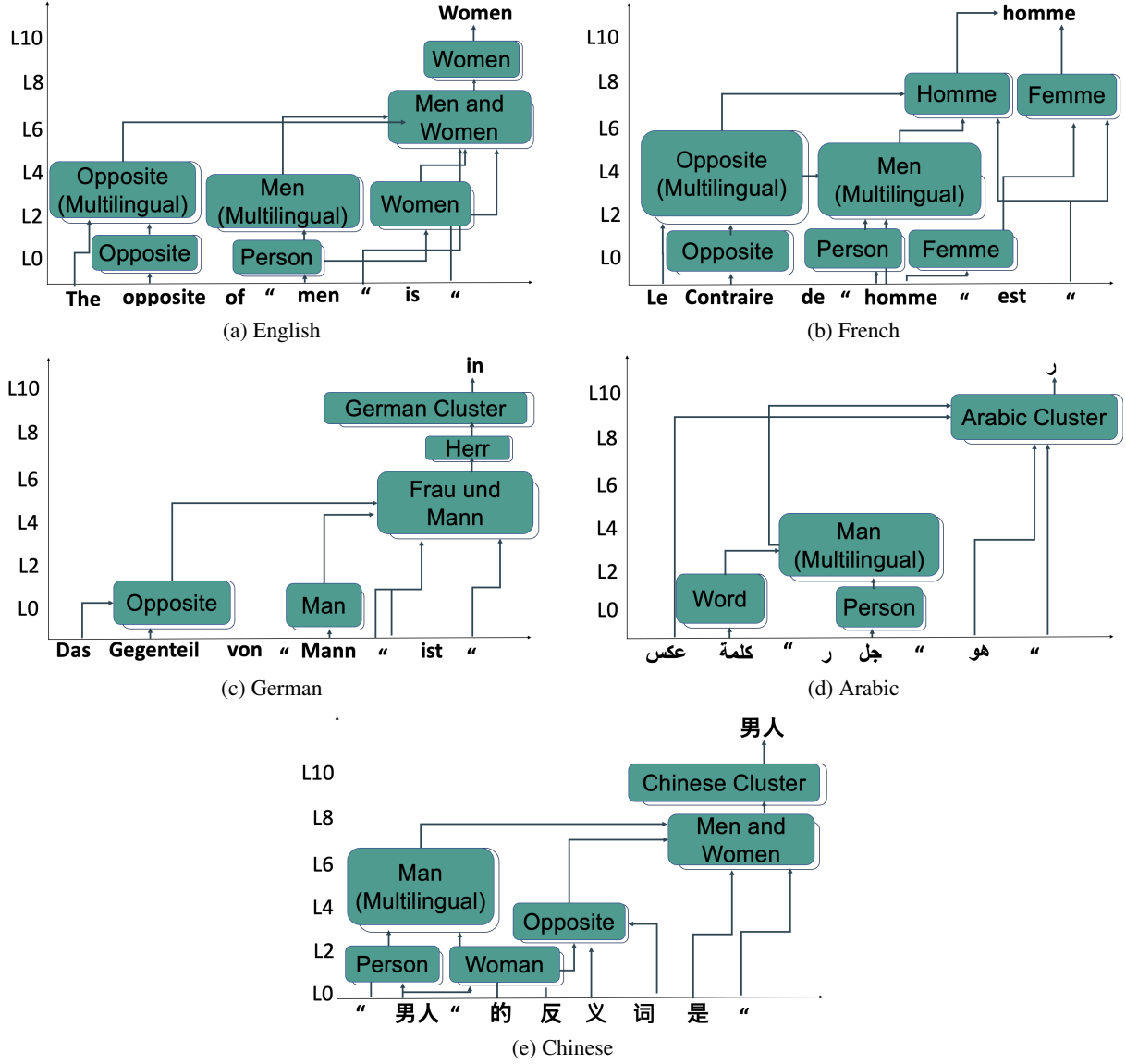


Figure 31: Circuits identified across five languages in the 90% English mixture.

aligned with unit boundaries. Japanese shows a different pattern, indicating that token recovery depends not only on fragmentation or morphology, but also on how token boundaries interact with the structure of each language.

These results extend our token-recovery analysis beyond the original five languages and clarify the role of boundary placement. We will add the full layerwise results and sensitivity analyses to the appendix.

L.4 Mechanistic Implications

This tokenization constraint directly impacts early-layer processing. Fragmented tokens lack semantic content, forcing early layers (0–3) to first reconstruct word boundaries before processing meaning. This reconstruction cost weakens the initial activa-

tion patterns that propagate through multilingual circuits.

Critically, despite this early-layer handicap, Arabic still develops identical shared multilingual representations in middle layers (5–6), demonstrating that the shared multilingual space mechanism is robust to tokenization disadvantage. However, the initial signal degradation explains why Arabic exhibits weaker circuit activation and lower downstream performance despite mechanistic equivalence.

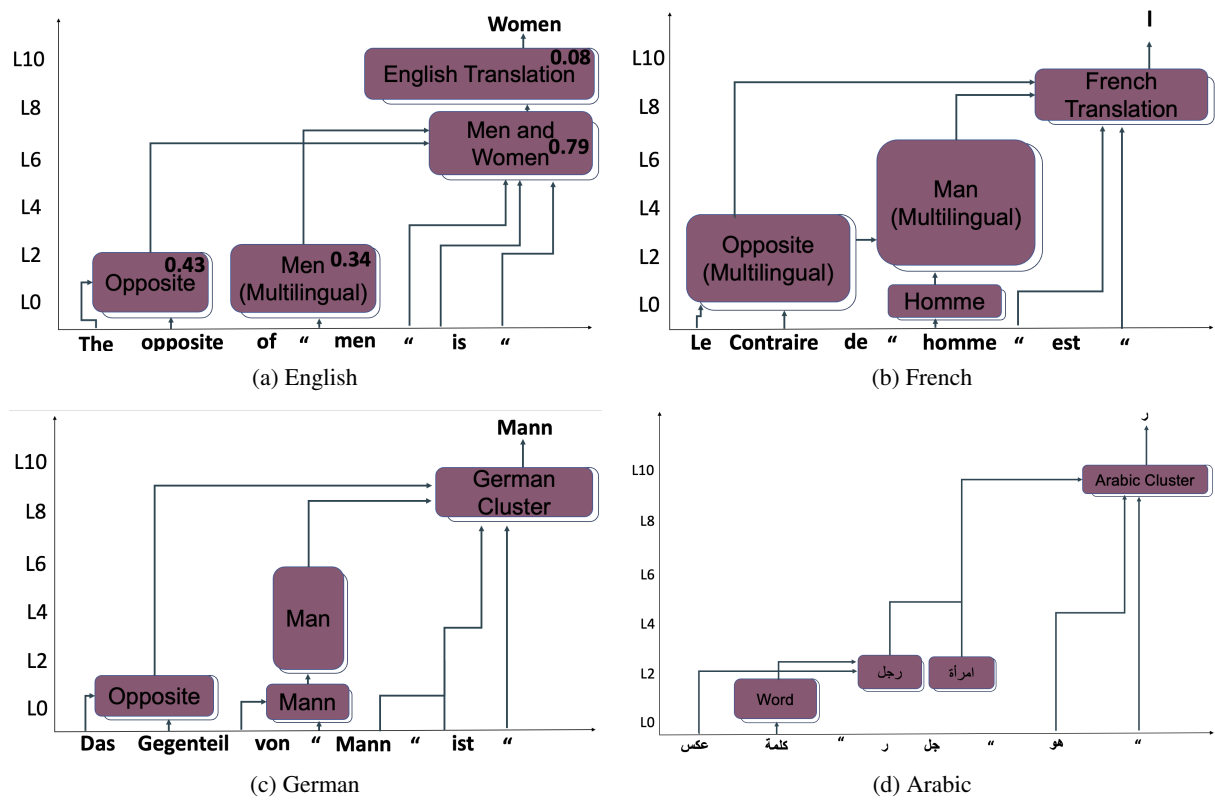


Figure 32: Circuits identified across four languages in the 70% English mixture.

Language	BPE tokens per unit	BPE boundaries inside a unit
Arabic	1.149	18.8%
Korean	1.164	17.8%
Japanese	1.120	13.1%
Chinese (descriptive)	1.239	18.2%

Table 9: Tokenization statistics.

Language	Early (Δ), L0–3	Middle (Δ), L5–8	Late (Δ), L9–11
Arabic	-2.164 [-2.526, -1.767]	-2.966 [-3.365, -2.525]	-0.851 [-1.298, -0.374]
Korean	-0.890 [-1.215, -0.568]	-0.393 [-0.719, -0.077]	0.538 [0.172, 0.903]
Japanese	1.251 [0.626, 1.835]	2.159 [1.531, 2.746]	3.408 [2.763, 4.005]
Chinese (descriptive)	0.543 [0.121, 0.929]	1.805 [1.366, 2.247]	2.696 [2.186, 3.146]

Table 10: Token-recovery results by layer band.

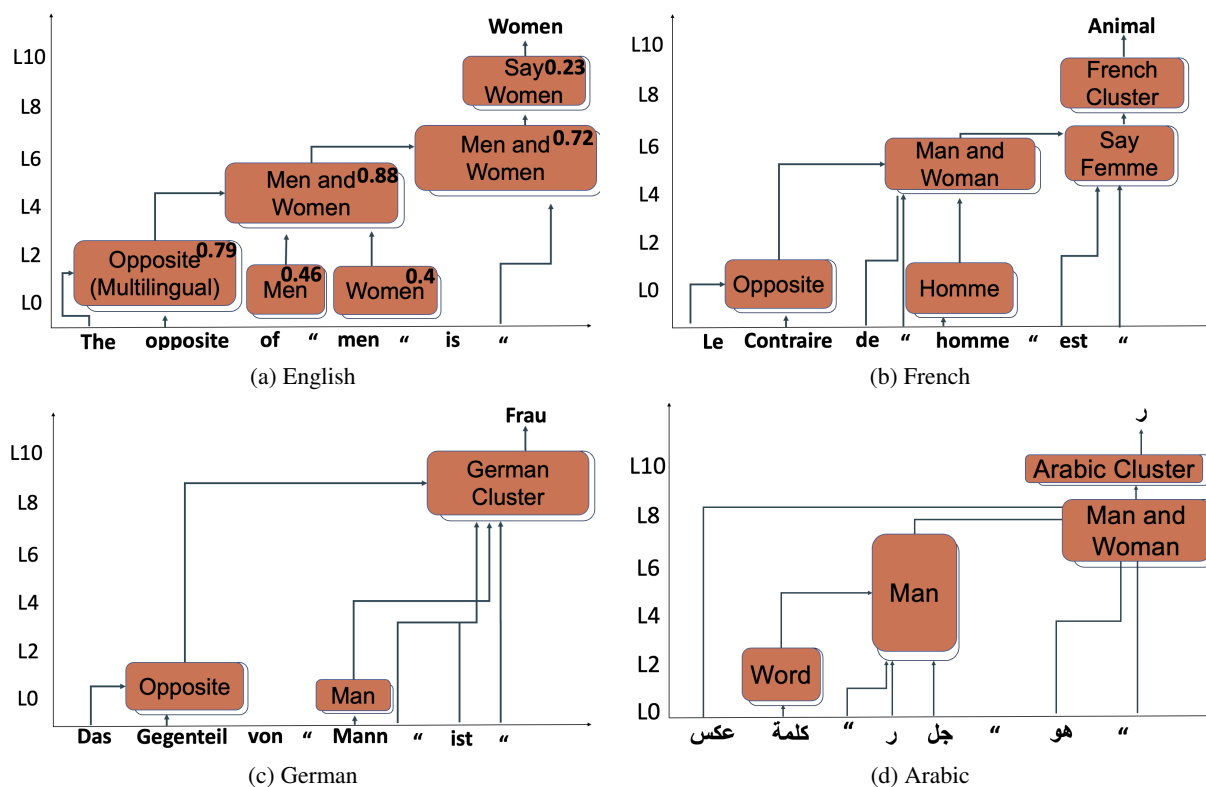


Figure 33: Circuits identified across four languages in the 50% English mixture.

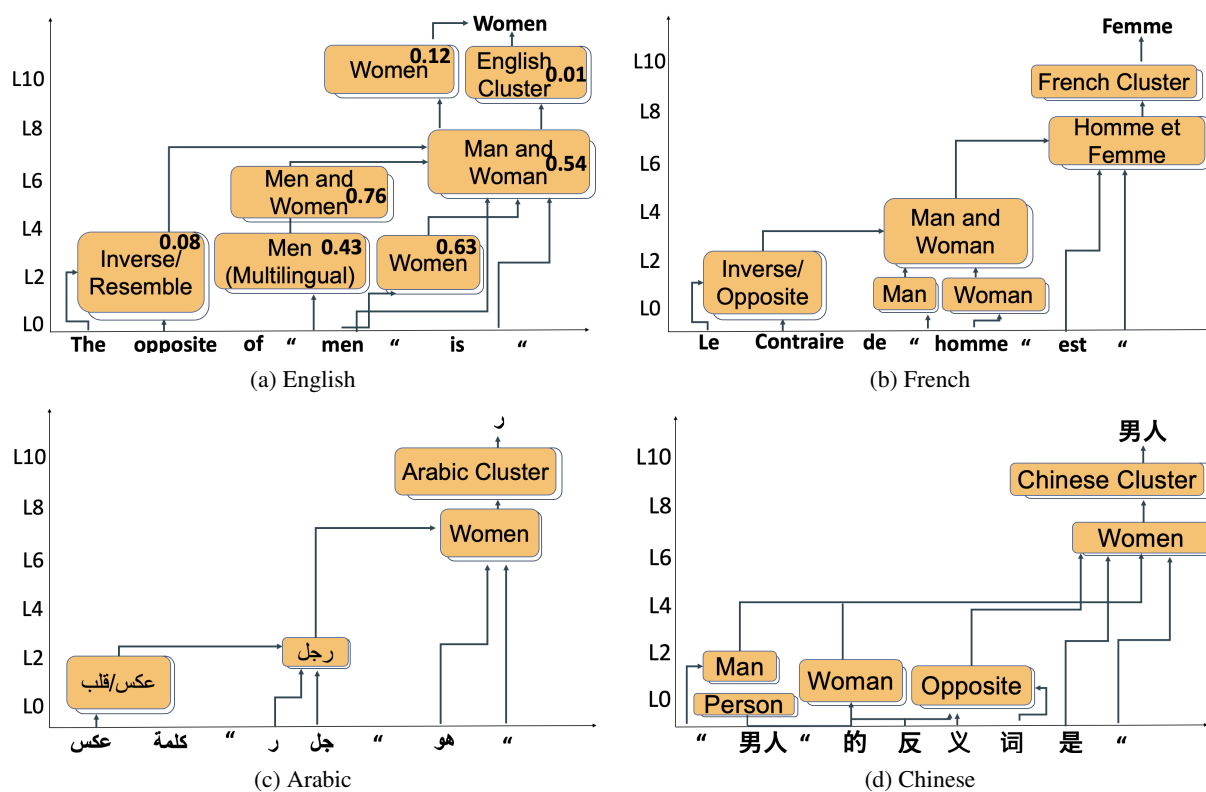


Figure 34: Circuits identified across four languages in the 20% English mixture.

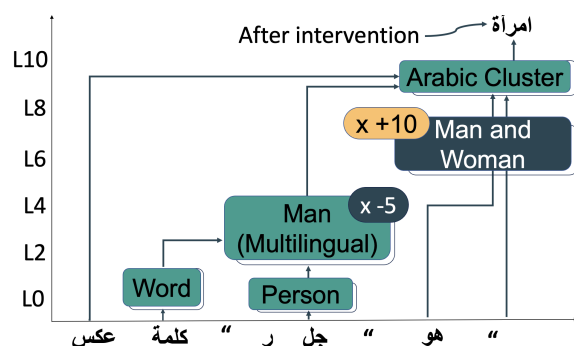


Figure 35: Adding the missing Men&Women cluster to the Arabic circuit restores performance.

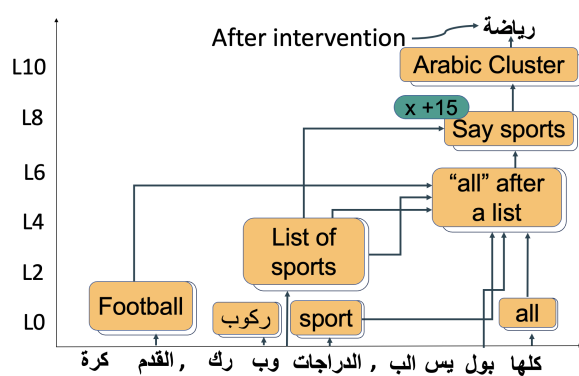


Figure 36: Scaling up the “Say Sports” cluster makes the model predict the right answer.

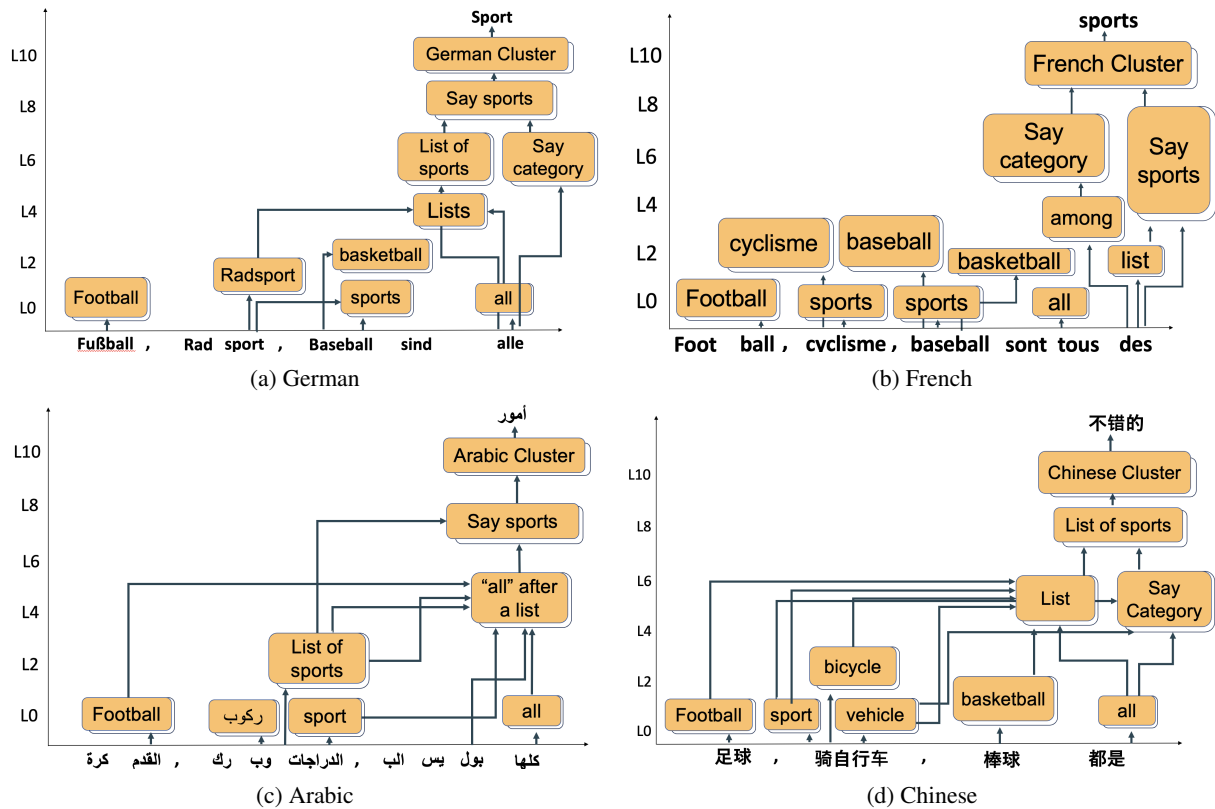


Figure 37: Circuits for category completion across languages in the 20% English mixture.

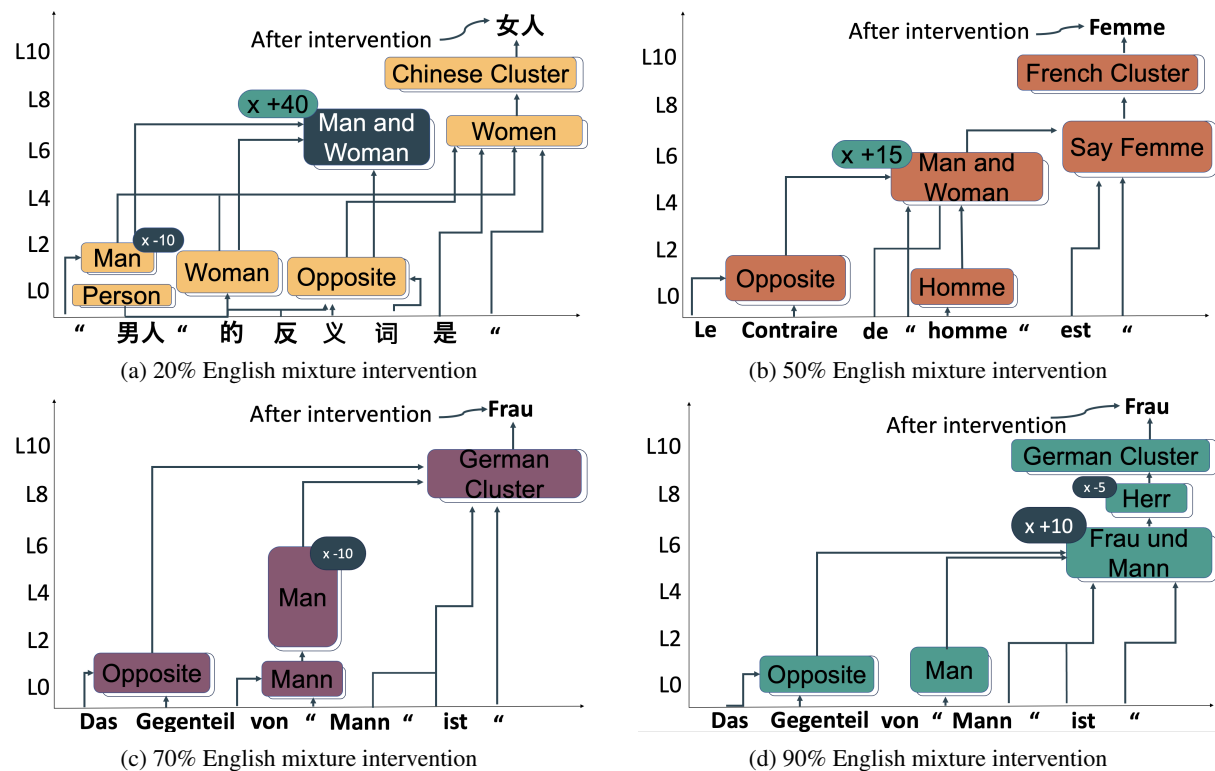


Figure 38: Examples of interventions where appropriate scaling shifts the model prediction toward the target language response.

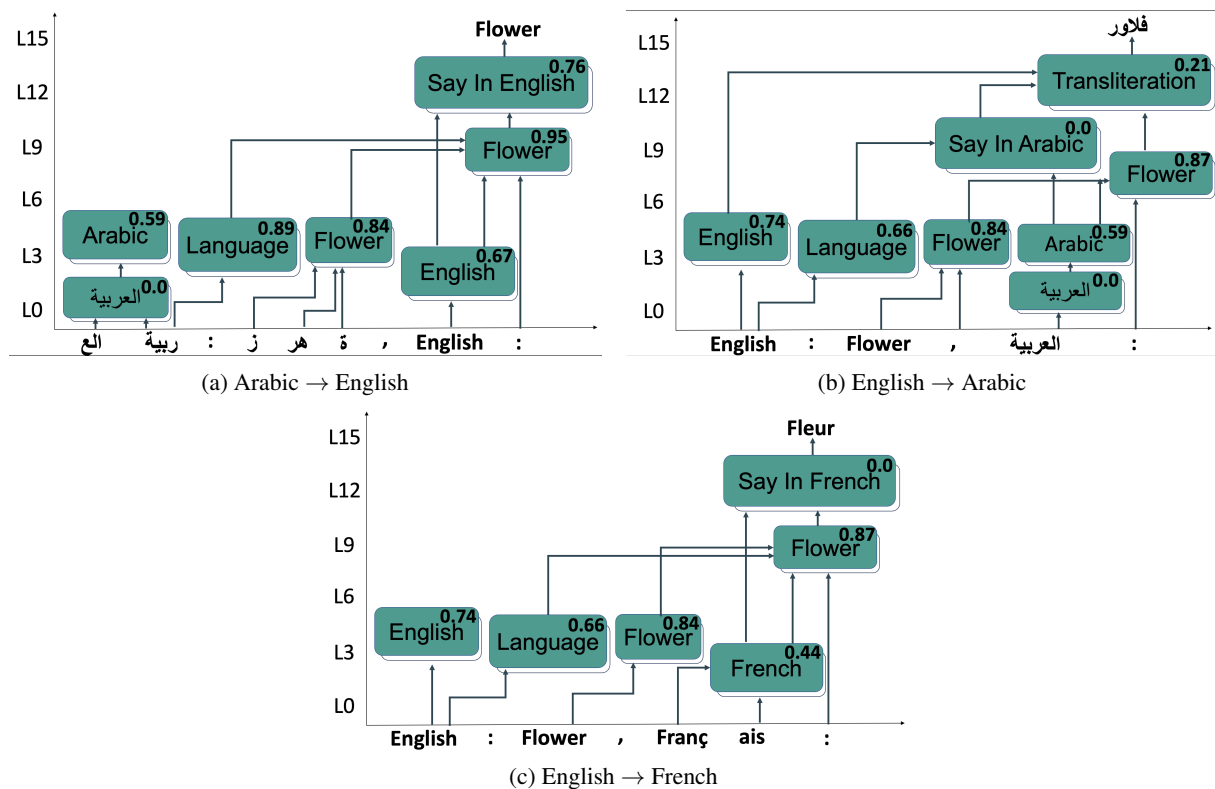


Figure 39: Circuits identified for multilingual translation prompts.

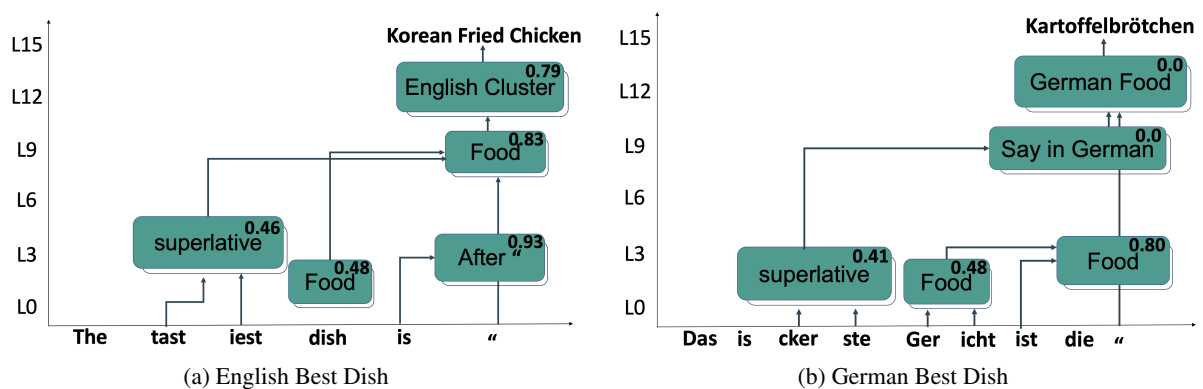


Figure 40: Circuits identified for culturally conditioned prompts across languages.

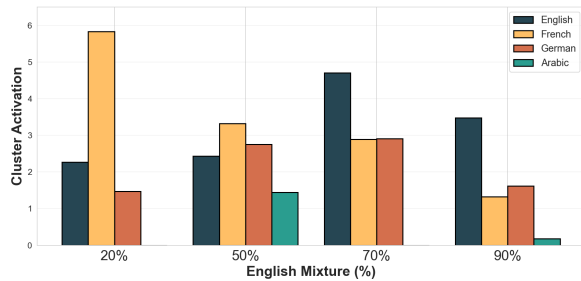
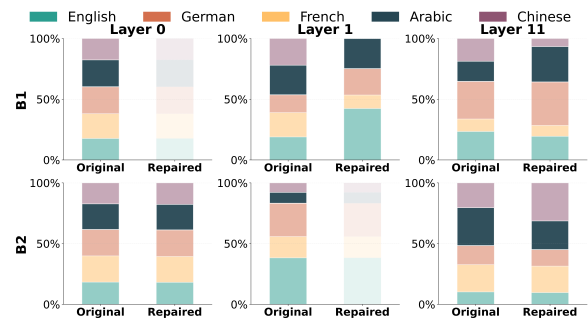
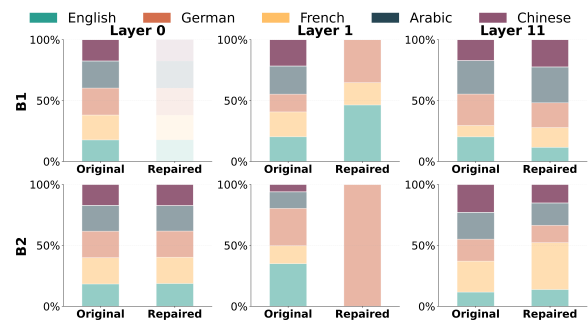


Figure 41: Cluster activation strength across languages and mixture ratios. English shows consistently higher activations for task-relevant clusters (Man&Woman for Antonym, Say Sports for Category Completion), with disparities diminishing at the 20% mixture.



(a) Arabic



(b) Chinese

Figure 43: Language distribution across layers and feature groups for Arabic (top) and Chinese (bottom). Blurred bars indicate insufficient activations for reliable estimation.

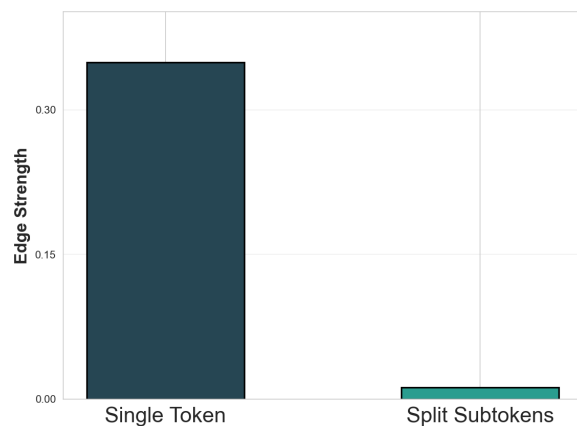


Figure 42: Edge strength from token embeddings to target clusters versus sub-tokenization rate. Higher sub-tokenization correlates with weaker edge strength ($r=-0.82$), explaining reduced circuit activation in fragmented languages like Arabic.

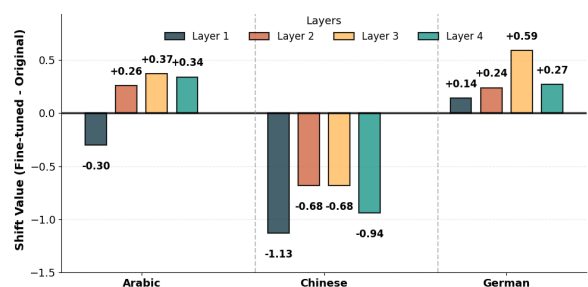


Figure 44: Logit Lens token assembly results. We find that models tend to repair early layers by making them better at assembling tokens, thus generalizing to abstractions earlier. This especially applies to Arabic and German. Surprisingly, Chinese finetuning results in less assembly across layers. This is in accordance with the higher-rank update of Chinese that shows that the model is possibly pattern-matching.