DO MONOLINGUAL LANGUAGE MODELS LEARN CROSS-LINGUAL UNIVERSAL CONCEPTUAL REPRESENTATIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Do language models learn universal, language-agnostic conceptual representations, or merely language-specific features that appear aligned under correlational analysis? We investigate this question using *Goldfish*, a family of independently trained small monolingual causal language models spanning 350 languages that share architecture and training budgets but differ entirely in data, vocabulary, and parameters, enabling the study of emergent conceptual structure without multilingual supervision. We first evaluate cross-lingual representational alignment using centered kernel alignment (CKA) at sentence and token levels on semantically matched parallel data, showing that independently trained models exhibit robust alignment beyond architectural baselines, with alignment strength scaling with training data and linguistic proximity. We then introduce **cross-lingual activation patching** as an interventional framework for testing concept validity by injecting hidden representations from a source-language model into a target-language model without learned projection or alignment. Across controlled case studies and large-scale contrastive evaluations, patched activations systematically steer target predictions in semantically consistent directions, with the strongest causal effects emerging in early and intermediate layers. These results provide evidence that monolingual language models learn partially compatible concept-level representations that support cross-model semantic transfer, positioning activation patching as a scalable technique for evaluating learned concepts under causal abstraction and offering new insights into the foundations of emergent universal concept learning and representation in large language models across the world's languages.

1 INTRODUCTION

A central question in multilingual language modelling is whether models develop **language-agnostic conceptual representations**. For example, do lexical items such as DOG (English), HUND (German), and CHIEN (French) in independently trained monolingual language models converge toward a shared internal representation (Zhou & Salhan, 2025)? Prior work has identified observational evidence for shared multilingual semantic spaces using probing or representational similarity analyses (Brinkmann et al., 2025; Dumas et al., 2025). However, these approaches remain correlational and do not establish whether cross-lingual representations are *functionally interchangeable*.

In this paper, we study this question using a family of independently trained monolingual language models spanning 350 languages, GOLDFISH (Chang et al., 2024b). Each monolingual GOLDFISH model share Transformer-based architecture and training budgets (for each language, a GOLDFISH LANGUAGE MODEL is trained on 5MB, 10MB, 100MB, and 1GB of text), sourced from multilingual corpora (Chang et al., 2024a; Imani et al., 2023; Kudugunta et al., 2023) . The GOLDFISH models provide a controlled testbed for emergent cross-lingual conceptual structure without shared supervision and allowing us to isolate the effect of language exposure on learned representations. Our analysis proceeds in three stages. We first evaluate **cross-lingual representational alignment** using centered kernel alignment (CKA) on semantically matched parallel data. We then probe **token-level lexical alignment** using automatically aligned word pairs. Finally, we introduce **cross-lingual activation patching** as a causal intervention to test whether internal activations can transfer concept information across independently trained models. Across sentence- and token-level analyses,

we observe robust alignment that varies with linguistic proximity and training scale. Activation patching experiments demonstrate that internal activations extracted from one monolingual model can causally steer behaviour in another, providing evidence for partially compatible cross-lingual conceptual representations.

2 REPRESENTATIONAL ALIGNMENT

2.1 EXPERIMENTAL SETUP

Models. We study cross-lingual representational similarity among nine monolingual GOLDFISH causal language models (125M parameters) (Chang et al., 2024b) trained independently on English, Chinese, Spanish, Arabic, Hindi, French, Russian, German, and Japanese.[1]

Data. To probe cross-lingual alignment under controlled semantic content, we use parallel sentences from FLORES-200 (Costa-Jussà et al., 2022).[2] For each ordered pair of languages, we randomly sample 200 sentence pairs from the development set. This ensures that representations are compared on semantically matched inputs while avoiding overlap between training data and evaluation text.

Representation extraction. For each model and each sentence, we extract hidden states from all transformer layers, including the embedding layer. Given token-level hidden states of shape $(T \times d)$, we compute a sentence-level representation by mean-pooling across the sequence dimension using the attention mask to exclude padding tokens. This yields one fixed-dimensional vector per sentence, per layer, and per model.

Similarity Metric. We measure cross-model similarity using linear Centered Kernel Alignment (CKA) (Kornblith et al., 2019). Given sentence-level representations $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$ from two models at a given layer, linear CKA is defined as

$$\mathrm{CKA}(X, Y) = \frac{\mathrm{HSIC}(X, Y)}{\sqrt{\mathrm{HSIC}(X, X)\,\mathrm{HSIC}(Y, Y)}},$$

where HSIC is computed using centered Gram matrices $K = XX^\top$ and $L = YY^\top$. CKA is invariant to isotropic scaling and dimensionality differences, making it suitable for cross-model and cross-language comparison.

Matched vs. shuffled controls. To distinguish semantic alignment from spurious similarity induced by architecture or marginal distributional effects, we compute two variants of CKA at each layer. In the *matched* condition, sentence-level representations are compared using aligned parallel sentences. In the *shuffled* control, we apply a fixed random permutation to the sentence order in one language while keeping the other fixed, and recompute CKA using the same permutation across all layers. This preserves the marginal representation distributions while destroying cross-lingual semantic correspondence.

Evaluation protocol. We repeat this procedure for all 36 unordered language pairs. For each pair, we report per-layer CKA scores for both matched and shuffled conditions, enabling a layerwise comparison of cross-lingual alignment under semantic correspondence versus random pairing.

2.2 SENTENCE-LEVEL ALIGNMENT

Figures 1 and 5 show that monolingual Goldfish language models develop aligned internal representations across languages when processing semantically equivalent sentences. Per-layer analyses in Figure 1 reveal that cross-lingual similarity, measured via linear CKA, is consistently higher for semantically matched sentence pairs than for shuffled baselines, with alignment strengthening in higher layers. This pattern indicates that representational similarity is driven by shared semantic

[1] The GOLDFISH models are available here: `https://huggingface.co/goldfish-models`

[2] `https://huggingface.co/datasets/facebook/flores`

structure rather than architectural overlap or marginal feature statistics. Extending this analysis across languages, Figure 5 demonstrates that this alignment generalizes across all examined language pairs at the final layer, while shuffled baselines remain uniformly low. Matched CKA scores increase with depth and substantially exceed shuffled baselines, indicating semantic alignment beyond architectural similarity. The results further show that alignment is stronger for linguistically related languages such as English–French and English–German, suggesting that typological similarity facilitates representational convergence. Additionally, models trained on larger datasets exhibit systematically stronger cross-lingual alignment, indicating that increased training data improves the stability and semantic consistency of learned representations, even under strictly monolingual training conditions.

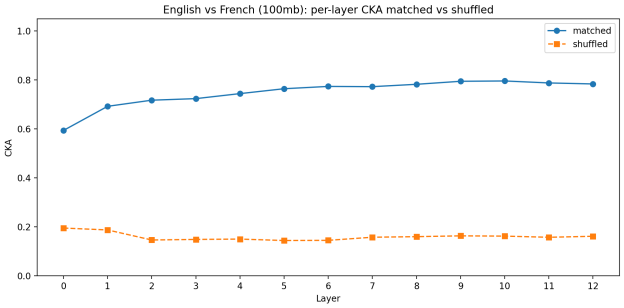

Figure 1: Layerwise English–French CKA on matched and shuffled sentences.

2.3 Token-Level Alignment

Sentence-level similarity conflates lexical and compositional effects. To isolate lexical-level cross-lingual similarity beyond global sentence semantics, we compute token-level alignment using automatically aligned word pairs. We compute token-level representational alignment for GOLDFISH models using parallel FLORES-200 sentences with automatic word alignments. These are obtained via SIMALIGN (Jalili Sabet et al., 2020), an unsupervised word alignment method that identifies corresponding words across parallel sentences by matching contextualized multilingual BERT embeddings using similarity-based matching heuristics. Let $T(w)$ denote the set of subword tokens associated with word w. Given hidden states $\mathbf{h}_{t,\ell} \in \mathbb{R}^d$ for subword token t at layer ℓ, we construct a word-level representation by averaging the hidden states of all subwords assigned to that word, $\mathbf{z}_{w,\ell} = \frac{1}{|T(w)|} \sum_{t \in T(w)} \mathbf{h}_{t,\ell}$. For each aligned word pair across languages, this produces a pair of layer-specific vectors that serve as directly comparable lexical representations, which are subsequently aggregated across aligned pairs to measure cross-lingual representational similarity. This yields aligned word representation pairs across languages, which are aggregated across sentences and used to compute layerwise linear CKA, ensuring invariance to representation dimensionality and scaling. To control for spurious correlations, we include a shuffled baseline that preserves marginal distributions while destroying cross-lingual correspondence. See *Appendix* B for detailed information on the alignment setup.

Figures 11 and 2 show that monolingual Goldfish models exhibit cross-lingual representational alignment at the lexical level, although this alignment is weaker than that observed for sentence-level semantics. Per-layer analyses in Figure 11 demonstrate that token-level similarity, measured via linear CKA, is consistently higher for matched aligned word pairs than for shuffled baselines across nearly all language pairs, confirming the presence of cross-lingual lexical correspondence. However, the overall magnitude of token-level CKA remains substantially lower than sentence-level alignment, indicating that fine-grained lexical representations are less directly shared across languages. Unlike sentence-level alignment, which increases steadily and peaks in higher layers, token-level alignment exhibits a flatter layerwise trajectory, with early and middle layers already capturing substantial cross-lingual correspondence and only modest gains at deeper layers, suggesting that lexical alignment emerges earlier while later layers increasingly encode language-specific compositional or syntactic structure. Extending this analysis across languages, Figure 2 shows that token-level alignment generalizes across all language pairs and remains clearly separated from shuffled baselines, though its strength varies systematically with linguistic relatedness. Closely related languages such

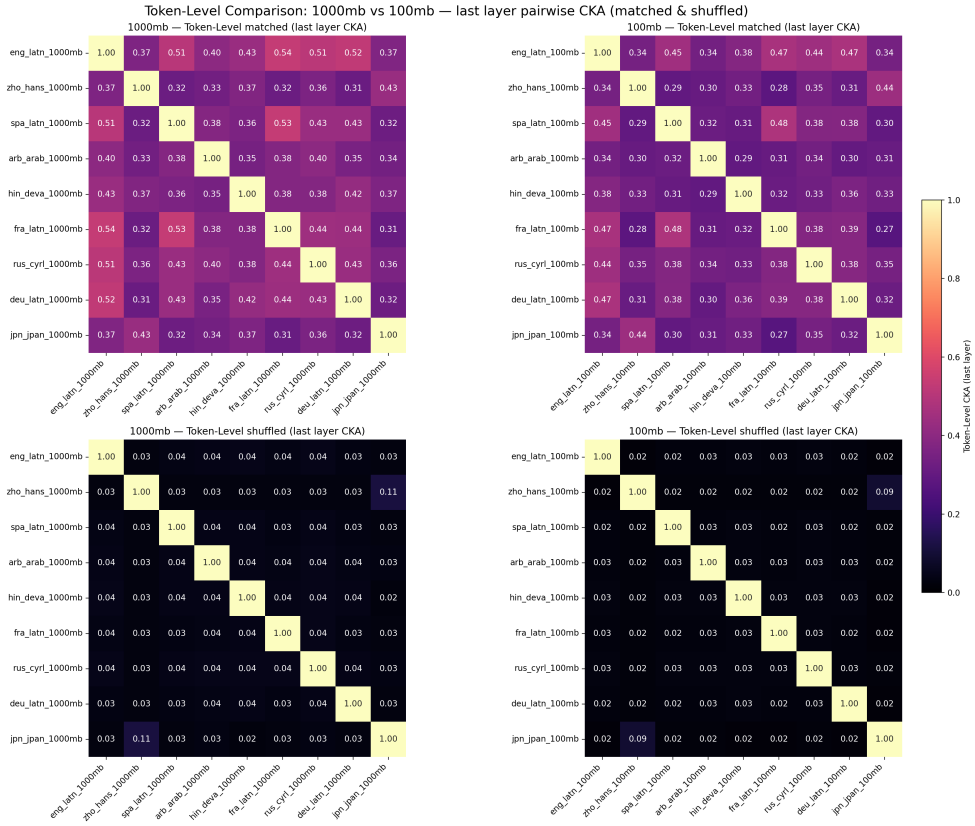

Figure 2: Token-level cross-lingual CKA across languages.

as English–French, English–German, and Spanish–French consistently exhibit stronger alignment than typologically distant pairs, indicating that lexical representations are more sensitive to typological distance than sentence-level semantic representations. Additionally, models trained on larger datasets achieve uniformly stronger token-level alignment across language pairs, demonstrating that increased training data improves the stability and cross-lingual consistency of lexical representations even under strictly monolingual training regimes. Token-level alignment persists across languages but is consistently weaker than sentence-level alignment. Unlike sentence-level similarity, token alignment exhibits weaker layer dependence and is more sensitive to typological distance. Larger training datasets improve token-level alignment across nearly all language pairs. Together, these findings indicate that **lexical semantic structure partially aligns across languages but remains more language-dependent than global sentence representations**.

3 Cross-Lingual Activation Patching

This section investigates whether independently trained monolingual language models develop *functionally compatible* semantic representations across languages. While earlier analyses demonstrated representational similarity, similarity alone does not establish whether these representations are causally interchangeable. To address this limitation, we introduce *cross-lingual activation patching*, a causal intervention that tests whether hidden representations extracted from one model can directly influence the predictions of another.

Activation patching methodology. Activation patching involves extracting hidden states from a *source* model and injecting them into a *target* model during inference. Given a source-language prompt, hidden activations are extracted from each transformer layer at the final token position. These activations are then additively injected into the corresponding layer of a target-language

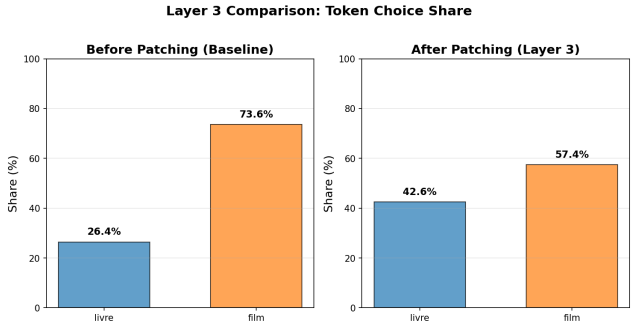

Figure 3: Choice probabilities before and after patching.

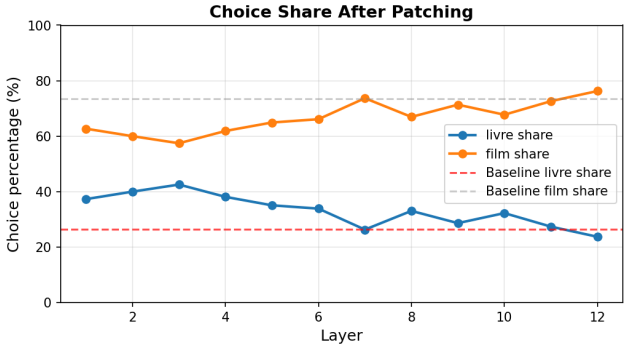

Figure 4: Layerwise patching effect on choice share.

model:

$$\mathbf{h}_\ell^{(t)}[T] \leftarrow \mathbf{h}_\ell^{(t)}[T] + \mathbf{a}_\ell^{(s)}.$$

Additive patching preserves the target model's internal computation while allowing the source activation to provide directional semantic bias. The causal effect is quantified by measuring changes in next-token probability distributions across layers.

French–German case study. We test whether a *reading-related* German representation can override a *watching-related* French cue using the prompts: German *Es ist etwas, das man lesen kann, wie ein"* and French *C'est quelque chose qu'on peut regarder, comme un"*. Before patching, the French model favours *film* (73.7%) over *livre* (26.3%). Injecting the German activation shifts probability toward *livre*, peaking in early and intermediate layers and diminishing in later ones, indicating that transferable conceptual information is encoded mid-layer, while final layers recover language-specific lexical preferences (Figure 7).

Scaled cross-model experiment. To generalise beyond illustrative prompts, we conduct large-scale activation patching using minimal semantic contrast pairs from the XCOMPS dataset (He et al., 2025). Sentences differ only in a single concept token (e.g., *Hund kann bellen* vs. *Katze kann bellen*) and are evaluated across German and French monolingual GOLDFISH models. We directly inject French concept-token activations into the German model without any alignment or projection and measure changes in continuation log-likelihood:

$$\Delta = \log p(\text{continuation} \mid \text{patched}) - \log p(\text{continuation} \mid \text{baseline}).$$

Three controls are used: within-language patching (upper bound), cross-model wrong-concept patching, and shuffled cross-model activations.

Results. Across 5,000 sampled instances, cross-model patching using semantically aligned concepts produces consistent improvements in continuation likelihood in early and mid layers compared

to shuffled baselines, demonstrating emergent cross-lingual compatibility. While within-language patching yields the strongest effect, cross-model interventions exhibit a similar layer-wise profile. In contrast, late-layer patching produces negligible or negative effects across all conditions, indicating that higher layers encode language-specific decision boundaries that limit cross-model transfer.

Conclusion. Cross-model patching produces positive improvements relative to shuffled baselines in early layers, indicating emergent compatibility between independently trained models. Effects diminish in later layers, suggesting language-specific decision computation dominates at higher depths. Together, these experiments provide causal evidence that independently trained monolingual language models develop *partially compatible semantic representations*. Although these representations are not fully aligned, they are sufficiently structured to enable meaningful cross-lingual and cross-model information transfer, particularly within intermediate semantic encoding layers.

Additional methodological details and experimental results are provided in Appendix C.

Condition	Early	Mid	Late
DE→DE	+0.013	+0.001	-0.103
FR→DE (same concept)	+0.015	-0.030	-0.201
FR→DE (wrong concept)	+0.005	-0.068	-0.232
FR→DE (shuffled)	+0.002	-0.032	-0.200

Table 1: Layer-aggregated patching results showing the effect of injecting source activations into a target model across different layers. Positive values indicate an increase in the target probability for the intended continuation, while negative values indicate a decrease. The first row (DE→DE) shows a within-language control. The next three rows illustrate cross-lingual patching from French to German under three conditions: matching concepts, mismatched concepts, and shuffled activations. Effects are strongest at early and mid layers and diminish in late layers, consistent with the hypothesis that transferable conceptual information is primarily encoded in intermediate representations.

4 Discussion and Conclusion

Recent methods aim to make neural representations interpretable by linking them to human-understandable concepts. Concept Bottleneck Models (CBMs) enforce intermediate concept layers before task outputs (Koh et al., 2020; Chauhan et al., 2023; Yuksekgonul et al., 2022; Oikarinen et al., 2023), while concept layer approaches (Bidusa & Markovitch, 2025) extract latent concept activations for intervention and causal analysis. Both focus on *intervenability* but typically require labeled data or architectural changes, limiting their use in large pretrained or cross-lingual models. Our approach uses activation patching on hidden states of monolingual Transformers, enabling causal probing of concept-level information without annotations or model modifications. Analyses show that independently trained monolingual models develop partially compatible semantic representations across languages. Sentence-level CKA alignment emerges in higher layers, while token-level lexical representations align earlier but remain sensitive to typological distance. Layerwise patterns suggest intermediate layers encode broadly transferable conceptual information, whereas later layers capture language-specific features. Cross-lingual activation patching confirms these representations influence predictions: early and intermediate layer patches bias target models toward semantically aligned concepts, while late-layer interventions have minimal effect.

Overall, our results suggest that multilingual conceptual structure can arise from strictly monolingual training, with partially shared semantic features encoded primarily in intermediate Transformer layers. Activation patching establishes that these representations are causally effective across models, providing a scalable framework for evaluating concept-level compatibility in language models. This work underscores the potential for monolingual models to develop emergent, language-agnostic conceptual knowledge and offers a practical method for probing and manipulating these internal representations without requiring aligned multilingual supervision or architectural modifications.

REFERENCES

Or Raphael Bidusa and Shaul Markovitch. Concept layers: Enhancing interpretability and intervenability via llm conceptualization. *arXiv preprint arXiv:2502.13632*, 2025.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. Large language models share representations of latent grammatical concepts across typologically diverse languages. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6131–6150, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.312. URL https://aclanthology.org/2025.naacl-long.312/.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.236. URL https://aclanthology.org/2024.emnlp-main.236/.

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. Goldfish: Monolingual language models for 350 languages. *arXiv preprint arXiv:2408.10441*, 2024b.

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *Proceedings of the aaai conference on artificial intelligence*, volume 37, pp. 5948–5955, 2023.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31822–31841, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1536. URL https://aclanthology.org/2025.acl-long.1536/.

Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Adrian Florea, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, Hinrich Schütze, and Nima Mesgarani. XCOMPS: A multilingual benchmark of conceptual minimal pairs. In Michael Hahn, Priya Rani, Ritesh Kumar, Andreas Shcherbakov, Alexey Sorokin, Oleg Serikov, Ryan Cotterell, and Ekaterina Vylomova (eds.), *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pp. 75–81, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-281-7. doi: 10.18653/v1/2025.sigtyp-1.9. URL https://aclanthology.org/2025.sigtyp-1.9/.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. Glot500: Scaling multilingual corpora and language models to 500 languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1082–1117, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.61. URL https://aclanthology.org/2023.acl-long.61/.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1627–1643, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.147. URL https://aclanthology.org/2020.findings-emnlp.147/.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset, 2023. URL https://arxiv.org/abs/2309.04662.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Ej Zhou and Suchir Salhan. Extended abstract for "linguistic universals": Emergent shared features in independent monolingual language models via sparse autoencoders. In David Ifeoluwa Adelani, Catherine Arnett, Duygu Ataman, Tyler A. Chang, Hila Gonen, Rahul Raja, Fabian Schmidt, David Stap, and Jiayi Wang (eds.), *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pp. 128–130, Suzhuo, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-345-6. doi: 10.18653/v1/2025.mrl-main.9. URL https://aclanthology.org/2025.mrl-main.9/.

A Sentence-Level Alignment

A.1 Cross-Lingual Representational Alignment

Semantic alignment beyond architectural similarity. Figure 1 shows per-layer linear CKA between the English and French Goldfish models when evaluated on semantically matched FLORES sentence pairs, compared against a shuffled baseline. Across nearly all layers, matched CKA scores are substantially higher than their shuffled counterparts. While shuffled CKA remains low and relatively flat across depth, matched CKA increases steadily from early layers and peaks in higher layers. This sharp divergence indicates that the observed representational similarity cannot be explained by shared architecture or marginal feature statistics alone, but instead reflects alignment induced by shared semantic content.

Alignment is consistent across language pairs. This pattern generalizes across all language pairs. Figure 5 reports pairwise CKA at the final layer for all 36 language pairs, under both matched and shuffled conditions. In the matched setting, all language pairs exhibit strong cross-lingual similarity, whereas shuffled baselines remain uniformly low. This confirms that cross-lingual representational alignment is a systematic phenomenon rather than an isolated effect driven by specific languages.

Stronger alignment for linguistically closer languages. Although alignment is present across all pairs, its strength varies with linguistic relatedness. Closely related languages—such as English–French, English–German, and French–German—exhibit consistently higher matched CKA scores than more distant pairs involving typologically dissimilar languages (e.g., English–Japanese or English–Chinese). This gradient suggests that while semantic structure is shared across languages, representational alignment is modulated by linguistic proximity, likely reflecting overlap in syntactic patterns, lexical semantics, or inductive biases induced by similar scripts and morphology.

Effect of training data scale. Figure 5 also compares models trained on different corpus sizes (100MB vs. 1000MB). Across nearly all language pairs, models trained on larger datasets exhibit systematically higher matched CKA scores than their smaller-data counterparts, while shuffled baselines remain comparable. This indicates that increased training data leads to more stable and semantically aligned internal representations, improving cross-lingual generalization even in strictly monolingual training regimes.

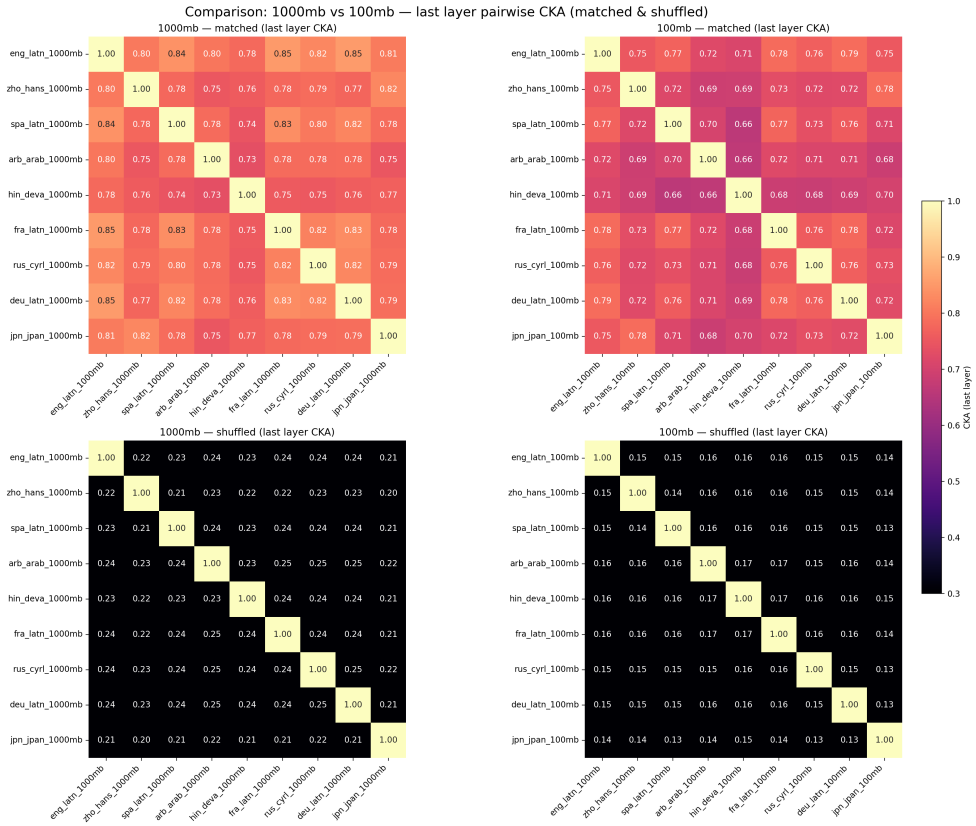

Figure 5: Pairwise cross-lingual representational similarity across languages and training scales. Heatmaps show last-layer linear CKA between all pairs of monolingual Goldfish models evaluated on semantically matched FLORES-200 sentences (top) and shuffled baselines (bottom). Results are shown for models trained on 1000MB (left) and 100MB (right) of monolingual data.

Summary. Taken together, these results demonstrate that (i) monolingual language models learn representations that align across languages when evaluated on semantically matched inputs, (ii) this alignment is robust across all language pairs but stronger for linguistically closer languages, and (iii) larger training corpora substantially enhance cross-lingual representational alignment.

B TOKEN-LEVEL ALIGNMENT

B.1 DETAILED TOKEN-LEVEL CROSS-LINGUAL ALIGNMENT SETUP

Motivation. While sentence-level CKA captures global semantic alignment, it conflates lexical, syntactic, and compositional effects. To more directly probe whether individual lexical units are represented similarly across languages, we extend our analysis to token-level representations aligned at the word level.

Data and alignments. We use the same parallel FLORES-200 sentence pairs as in the sentence-level experiments. For each sentence pair, we compute word-level alignments using an automatic alignment model. By default, we use SIMALIGN with contextualized BERT embeddings and cosine matching. When alignment fails, we fall back to a simple heuristic that aligns identical lowercased surface forms. This produces a set of aligned word index pairs for each sentence pair.

Mapping words to model tokens. Because Goldfish models operate on subword units, we map each aligned word to its corresponding tokenizer tokens using offset mappings provided by fast tokenizers. For each word, we collect all subword tokens whose character spans overlap with the word span. This yields a variable-length set of token indices per word.

Token-level representation extraction To obtain token-level representations that are comparable across languages, we operate at the level of aligned words rather than raw subword tokens. For each parallel sentence pair, we compute word-level alignments using SIMALIGN (Jalili Sabet et al., 2020), which leverages contextualized multilingual BERT embeddings to identify semantically corresponding word pairs across languages. Let $w_i^{(A)}$ and $w_j^{(B)}$ denote an aligned word pair in languages A and B, respectively.

Because the underlying language models operate on subword units, we map each word to its corresponding set of tokenizer subwords using character-level offset mappings. Specifically, for a word span $w = [c_s, c_e)$ and a tokenizer subword with span $t = [c'_s, c'_e)$, the subword is assigned to the word if the spans overlap. This yields a set of subword indices $T(w)$ for each word.

Given the hidden states $\mathbf{h}_{t,\ell} \in \mathbb{R}^d$ of subword token t at transformer layer ℓ, we compute a word-level representation by averaging over its associated subwords:

$$\mathbf{z}_{w,\ell} \;=\; \frac{1}{|T(w)|} \sum_{t \in T(w)} \mathbf{h}_{t,\ell}.$$

For each aligned word pair $(w_i^{(A)}, w_j^{(B)})$, this procedure yields a pair of vectors $(\mathbf{z}_{w_i,\ell}^{(A)}, \mathbf{z}_{w_j,\ell}^{(B)})$ at every layer. Aggregating across all aligned word pairs and all sentence pairs produces a set of token-level representations per layer for each model, which we then use to compute layerwise representational similarity.

Representational similarity metric. Given aligned word-level representations $X \in \mathbb{R}^{n \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$ at a given layer, we compute linear CKA using the same formulation as in the sentence-level analysis. As before, CKA is invariant to representation dimensionality and isotropic scaling.

Matched and shuffled controls. To control for spurious similarity, we again compute a shuffled baseline. For each layer, we randomly permute the aligned word representations in one language while keeping the other fixed, and recompute CKA using the same permutation across all layers. This preserves marginal feature distributions while breaking word-level semantic correspondence.

Evaluation protocol. We report per-layer token-level CKA scores under both matched and shuffled conditions for each language pair. For reference, we also compute sentence-level CKA using mean-pooled final-layer representations on the same sentence subset, enabling a direct comparison between token-level and sentence-level alignment.

B.2 Token-Level Cross-Lingual Alignment

Token-level alignment exists but is weaker than sentence-level alignment. Figure 11 reports per-layer token-level CKA for all language pairs under matched and shuffled conditions. Across nearly all pairs, matched token-level CKA scores are consistently higher than shuffled baselines, demonstrating that cross-lingual alignment persists even when representations are compared at the level of aligned lexical units. However, the absolute magnitude of token-level CKA is substantially lower than in the sentence-level setting, indicating that fine-grained lexical representations are less directly aligned across languages than sentence-level semantic representations.

Layerwise behavior differs from sentence-level alignment. Unlike sentence-level CKA, which increases monotonically and peaks in higher layers, token-level CKA exhibits a flatter or mildly increasing trajectory across depth. Early and middle layers already exhibit non-trivial alignment, with only modest gains in later layers. This suggests that lexical correspondence across languages is established relatively early in the network, while higher layers increasingly encode language-specific compositional or syntactic structure rather than word-level semantics.

Alignment holds across all language pairs, with typological effects. Figure 2 shows last-layer token-level CKA for all language pairs. As in the sentence-level case, all language pairs exhibit clear separation between matched and shuffled conditions, confirming that alignment is not driven by architectural similarity alone. Nevertheless, alignment strength varies systematically across pairs. Linguistically closer languages (e.g., English–French, English–German, Spanish–French) tend to exhibit higher token-level CKA than more distant pairs (e.g., English–Japanese or English–Chinese), indicating that lexical alignment is more sensitive to typological distance than sentence-level semantic alignment.

Effect of training data scale persists at the token level. Comparing models trained on 100MB versus 1000MB of data reveals a consistent data-scale effect. As shown in Figure 2, larger-data models achieve uniformly higher matched token-level CKA scores across almost all language pairs, while shuffled baselines remain comparably low. This mirrors the sentence-level findings and suggests that increased training data improves the stability and cross-lingual consistency of lexical representations, even in strictly monolingual training regimes.

Summary. Taken together, these results show that (i) cross-lingual alignment is present at the lexical level when evaluated on aligned word pairs, (ii) token-level alignment is weaker and less layer-dependent than sentence-level alignment, and (iii) both linguistic proximity and training data scale play a stronger role in shaping fine-grained representational similarity

C Detailed Patching Experiment Results

C.1 Case Study in French and German

In the previous sections, we showed that independently trained monolingual language models exhibit strong representational similarity across languages, suggesting the emergence of language-agnostic semantic structure. However, representational similarity alone is correlational: it does not establish whether these representations are *functionally aligned*. In this section, we introduce a causal intervention—*cross-lingual activation patching*—to test whether internal activations extracted from one monolingual model can directly influence the behavior of another monolingual model in a concept-consistent manner.

C.1.1 Activation Patching

Activation patching is a causal analysis technique in which internal hidden states from a *source* model are injected into a *target* model during a forward pass, allowing us to test whether the patched activations carry behaviorally meaningful information. In our cross-lingual setting, the source and target models are trained on different languages but share the same architecture.

Concretely, given a source-language prompt $x^{(s)}$ and a target-language prompt $x^{(t)}$, we first run the source model with hidden-state extraction enabled and collect the hidden activation at the final token position from each transformer layer. Let $\mathbf{a}_\ell^{(s)} \in \mathbb{R}^d$ denote the hidden state at layer ℓ and the last token position of the source prompt. This vector can be interpreted as the contextual representation induced by the source-language cue.

We then intervene on the target model by injecting this source activation into the target model's computation. Specifically, during the forward pass of the target model, we modify the hidden state at layer ℓ and the final token position as

$$\mathbf{h}_\ell^{(t)}[T] \leftarrow \mathbf{h}_\ell^{(t)}[T] + \mathbf{a}_\ell^{(s)},$$

where T denotes the final token position of the target prompt. We use *additive* patching rather than replacement patching in order to preserve the target model's existing contextual computation while testing whether the source activation provides an additional directional bias. Replacement patching would constitute a much stronger intervention that overwrites the target representation entirely and may introduce artifacts unrelated to natural computation. Additive patching therefore serves as a more conservative and interpretable causal probe.

After patching, we measure the effect of the intervention on the target model's next-token distribution. By comparing token probabilities before and after patching, we can quantify how much the injected source activation causally shifts the target model's preferences. We repeat this procedure layer by layer, yielding a layerwise profile of cross-lingual causal influence.

C.1.2 Experiment and Results

We illustrate this methodology with a deliberately constructed cross-lingual case study between German and French, designed to test whether a *reading-related* representation extracted from German can override a *watching-related* cue in French.

Setup. The source model is a German monolingual language model, and the target model is a French monolingual language model of the same architecture. We use the following prompts:

- **German source cue (reading):**
  "Es ist etwas, das man lesen kann, wie ein"
- **French target prompt (watching):**
  "C'est quelque chose qu'on peut regarder, comme un"

The French prompt is intentionally biased toward a watching-related continuation. We evaluate two competing next-token candidates in French: *livre* (book) and *film* (film).

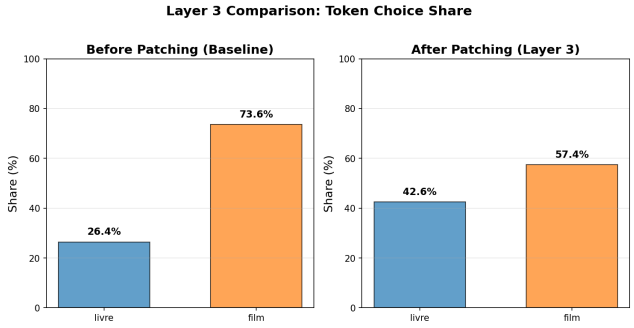

Figure 6: Layerwise comparison of the French model's choice distribution between *livre* and *film* before and after cross-lingual activation patching. Before patching, the model strongly prefers *film* across all layers. After injecting a German reading-related activation, the probability mass shifts toward *livre* in early and intermediate layers, despite the French prompt remaining unchanged.

Baseline behavior. Before any patching, the French target model strongly prefers *film*. Specifically, the normalized choice probabilities are:

$$P(\textit{livre}) = 26.3\%, \qquad P(\textit{film}) = 73.7\%.$$

This confirms that the target prompt successfully induces a watching-oriented bias in the model.

Effect of cross-lingual patching. Figure 6 visualizes the effect of cross-lingual activation patching by directly comparing the French model's choice distribution before and after patching at each layer. Prior to any intervention, the target model consistently prefers *film* (73.7%) over *livre* (26.3%), reflecting the watching-oriented bias of the French prompt. After patching in the German reading-induced activation, this distribution shifts markedly across multiple layers. In early and intermediate layers, the probability mass assigned to *livre* increases substantially—often exceeding 40%—while the share of *film* correspondingly decreases. This shift occurs despite the surface-level French context remaining unchanged, indicating that the injected German activation provides a semantically meaningful signal that alters the target model's internal decision process. The magnitude of the effect varies across layers, with the strongest rebalancing observed in mid-layer representations, suggesting that these layers encode abstract, transferable semantic features.

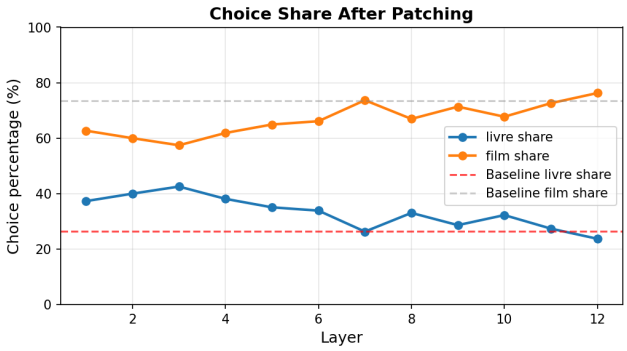

Figure 7: Choice share between *livre* and *film* after cross-lingual activation patching as a function of layer. Dashed lines indicate the baseline (unpatched) preference. Patching systematically increases the share of *livre* and decreases the share of *film*, with the strongest effect at intermediate layers, revealing where transferable semantic representations exert the greatest causal influence.

Interpretation. Figure 7 summarizes the same phenomenon as a continuous layerwise trajectory, plotting the normalized choice share of *livre* and *film* after patching relative to their baseline levels. The figure highlights two key observations. First, patching consistently pushes the model away from its baseline preference for *film* and toward *livre* across a broad range of layers, providing causal

evidence that the source-language activation aligns with the target-language concept. Second, the effect is not uniform: the preference shift peaks at intermediate depths and diminishes in later layers, where the model increasingly reasserts its original lexical bias. Together, these results demonstrate that monolingual language models do not merely learn superficially similar representations across languages, but encode concept-level features that are sufficiently aligned to be *functionally interchangeable*—capable of causally steering predictions in a different language model when injected at the appropriate representational depth.

C.2 Scaled Experiment: Cross-Model Activation Patching without Explicit Alignment

C.2.1 Problem Formulation and Experimental Design

We study whether independently trained monolingual language models develop *partially compatible semantic representations* that support cross-model information transfer, even in the absence of any explicit alignment mechanism. To this end, we design a controlled activation patching experiment that isolates the causal contribution of concept-level representations across languages.

Controlled concept contrast. We consider minimal sentence pairs that differ only in a single concept token while sharing identical syntactic structure. For example, in German:

Hund kann bellen. (acceptable)
Katze kann bellen. (unacceptable)

and their French counterparts:

Chien peut aboyer.
Chat peut aboyer.

Here, the predicate ***bellen / aboyer*** (to bark) is semantically compatible with ***Hund / chien*** (dog) but incompatible with ***Katze / chat*** (cat). This design allows us to attribute differences in model behavior specifically to the representation of the concept token.

Models and representations. Let M^{DE} and M^{FR} denote independently trained monolingual GOLDFISH causal language models with identical architectures. For a given transformer layer ℓ, we extract the hidden state corresponding to the concept token in the French sentence, $h_\ell^{\text{FR}}(\text{chien})$, and intervene on the German model by replacing the hidden state at the corresponding token position:

$$h_\ell^{\text{DE}}(\text{Katze}) \leftarrow h_\ell^{\text{FR}}(\text{chien}).$$

Crucially, no alignment map, projection, or normalization is applied; the French hidden state is directly injected into the German model's residual stream.

Causal intervention objective. This intervention tests whether the French representation of the concept *chien*, when inserted into the German model, causally shifts the model's belief about the continuation of an otherwise identical sentence. If the two models encode semantically related concepts in a compatible manner, such patching should increase the likelihood of a continuation consistent with the predicate *bellen*.

Evaluation metric. We quantify the effect of patching by measuring the change in continuation log-likelihood:

$$\Delta = \log p_{M^{\text{DE}}}(\text{“kann bellen.”} \mid \text{patched input}) - \log p_{M^{\text{DE}}}(\text{“kann bellen.”} \mid \text{original input}),$$

where a positive Δ indicates that the injected representation increases the model's confidence in the semantically appropriate continuation.

Baselines and controls. To contextualize cross-model effects, we compare against three controls: (i) *within-language patching* (DE→DE), which serves as an upper bound; (ii) *cross-model wrong-concept patching*, where an incompatible French concept (e.g., *chat*) is injected; and (iii) a *shuffled cross-model baseline*, where French hidden states are sampled from unrelated sentences. These controls allow us to disentangle semantic transfer from generic or distributional perturbations.

C.2.2 SETUP

Dataset. We evaluate cross-model activation patching on the XCOMPS dataset, which provides contrastive minimal sentence pairs designed to isolate concept-level semantic compatibility. Each instance consists of an *acceptable* and an *unacceptable* sentence that differ only in a single concept token while sharing identical syntactic structure across languages.

Sampling. From the full XCOMPS minimal pairs dataset, we randomly sample 5,000 contrastive instances for evaluation. Each instance includes parallel German–French sentence pairs, enabling controlled cross-lingual interventions such as *Hund* vs. *Katze* in German and *chien* vs. *chat* in French. This sampling strategy balances statistical robustness with computational feasibility while preserving the semantic contrast required for causal analysis.

Evaluation protocol. For each instance, we perform activation patching at every transformer layer of the German model, intervening on the hidden state corresponding to the concept token. We then compute the change in continuation log-likelihood (Δ) relative to the unpatched baseline. Results are aggregated across instances and grouped into early, mid, and late layer bands for reporting.

Reproducibility. All experiments use fixed random seeds for dataset sampling and model execution. Token-level alignment is performed using tokenizer offset mappings to ensure that patching targets the correct concept position in each sentence. Further implementation details and per-layer results are provided in Appendix C Table 3.

C.2.3 RESULTS AND ANALYSIS

Condition	Early (0–3)	Mid (4–7)	Late (8–11)
Within-language (DE→DE, Hund→Katze)	$+0.013$ (58.1%)	$+0.001$ (58.5%)	-0.103 (47.4%)
Cross-model same concept (FR→DE, chien→Katze)	$+0.015$ (53.0%)	-0.030 (47.3%)	-0.201 (27.0%)
Cross-model wrong concept (FR→DE, chat→Katze)	$+0.005$ (51.8%)	-0.068 (43.5%)	-0.232 (24.2%)
Cross-model shuffled (FR→DE, random→Katze)	$+0.002$ (51.4%)	-0.032 (49.8%)	-0.200 (27.5%)

Table 2: **Layer-aggregated effects of activation patching.** We report the mean change in continuation log-likelihood (Δ) aggregated over early (layers 0–3), mid (4–7), and late (8–11) layers of the German model. Percentages denote the fraction of test items with $\Delta > 0$. Cross-model patching from French to German improves continuation likelihood relative to shuffled baselines in early and mid layers, even without explicit alignment, indicating partial emergent compatibility between independently trained models.

Table 2 reports layer-aggregated results across early (layers 0–3), mid (4–7), and late (8–11) transformer blocks.

Emergent cross-model compatibility. Even without explicit alignment, cross-model patching from French to German using the same concept (*chien→Hund*) yields a consistent positive increase in continuation likelihood in early and mid layers. Importantly, this improvement is systematically larger than that obtained from shuffled French representations, indicating that the effect cannot be explained by random or purely mechanical perturbations.

Comparison to within-language patching. Within-language patching produces the strongest effects, as expected, but the layer-wise profile of cross-model patching closely mirrors this upper bound: gains are concentrated in early and intermediate layers and diminish in later layers. This similarity suggests that the German model responds to French hidden states at roughly the same representational stages where it responds to its own internal representations.

Layer dependence. Across all conditions, effects collapse or become negative in late layers. This behavior is consistent with prior observations that higher transformer layers encode model-specific decision boundaries and language-dependent logits, making them less amenable to cross-model interventions. The concentration of positive effects in earlier layers supports the interpretation that emergent compatibility arises primarily at the level of semantic feature encoding rather than final decision computation.

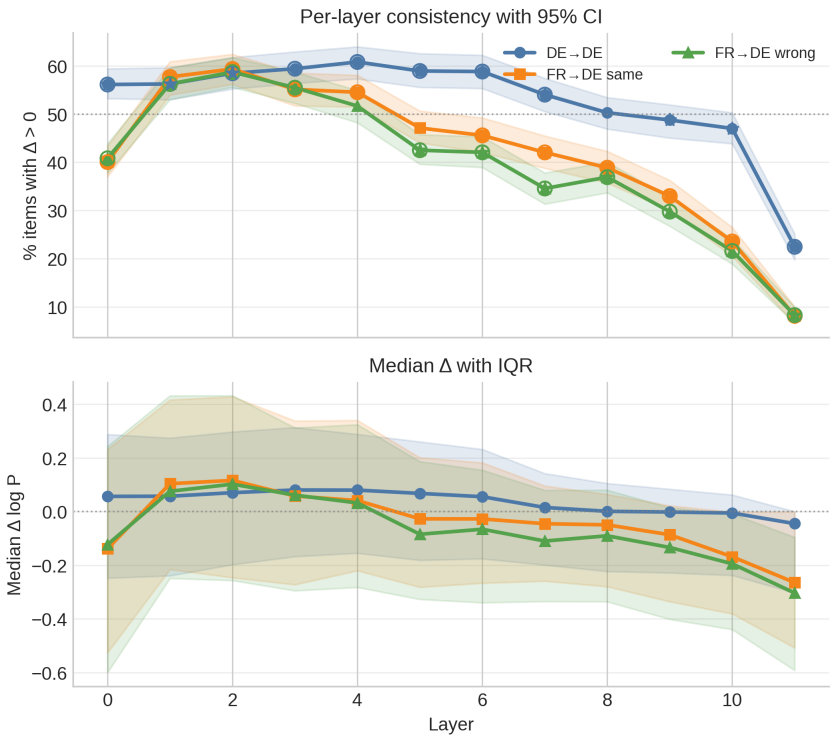

Figure 8: Caption

Takeaway. Together, these results provide causal evidence that independently trained monolingual models develop *partially compatible semantic representations*. While these representations are not fully aligned, they are sufficiently structured to support meaningful cross-model information transfer without any explicit alignment mechanism.

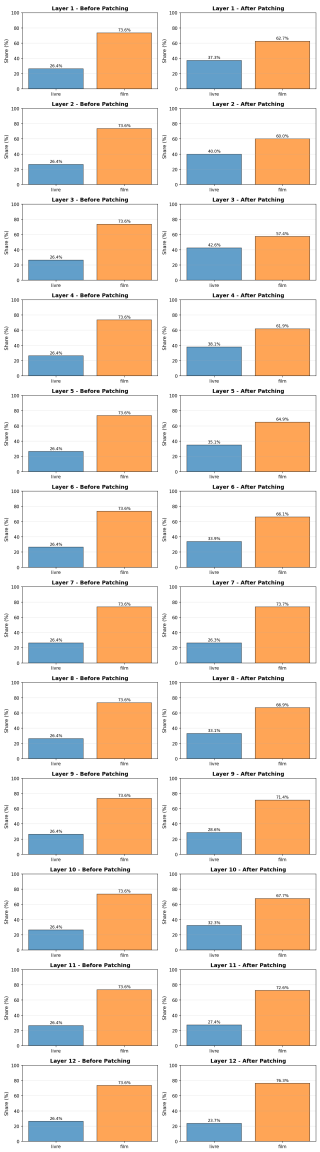

Figure 9: Layerwise comparison of the French model's choice distribution between *livre* and *film* before and after cross-lingual activation patching. Before patching, the model strongly prefers *film* across all layers. After injecting a German reading-related activation, the probability mass shifts toward *livre* in early and intermediate layers, despite the French prompt remaining unchanged.

D Additional Results

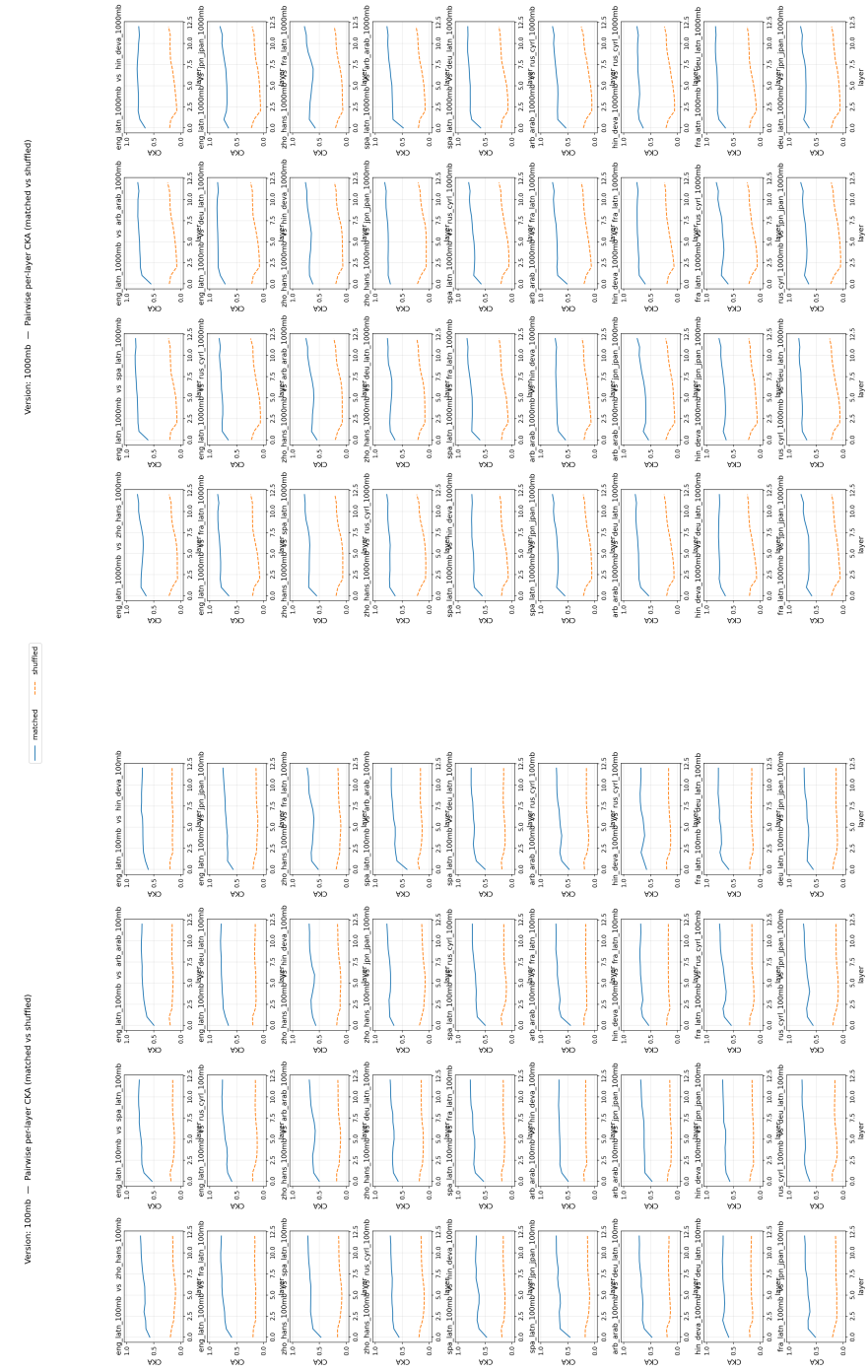

Figure 10: Per-layer cross-lingual representational similarity across all language pairs and training scales.

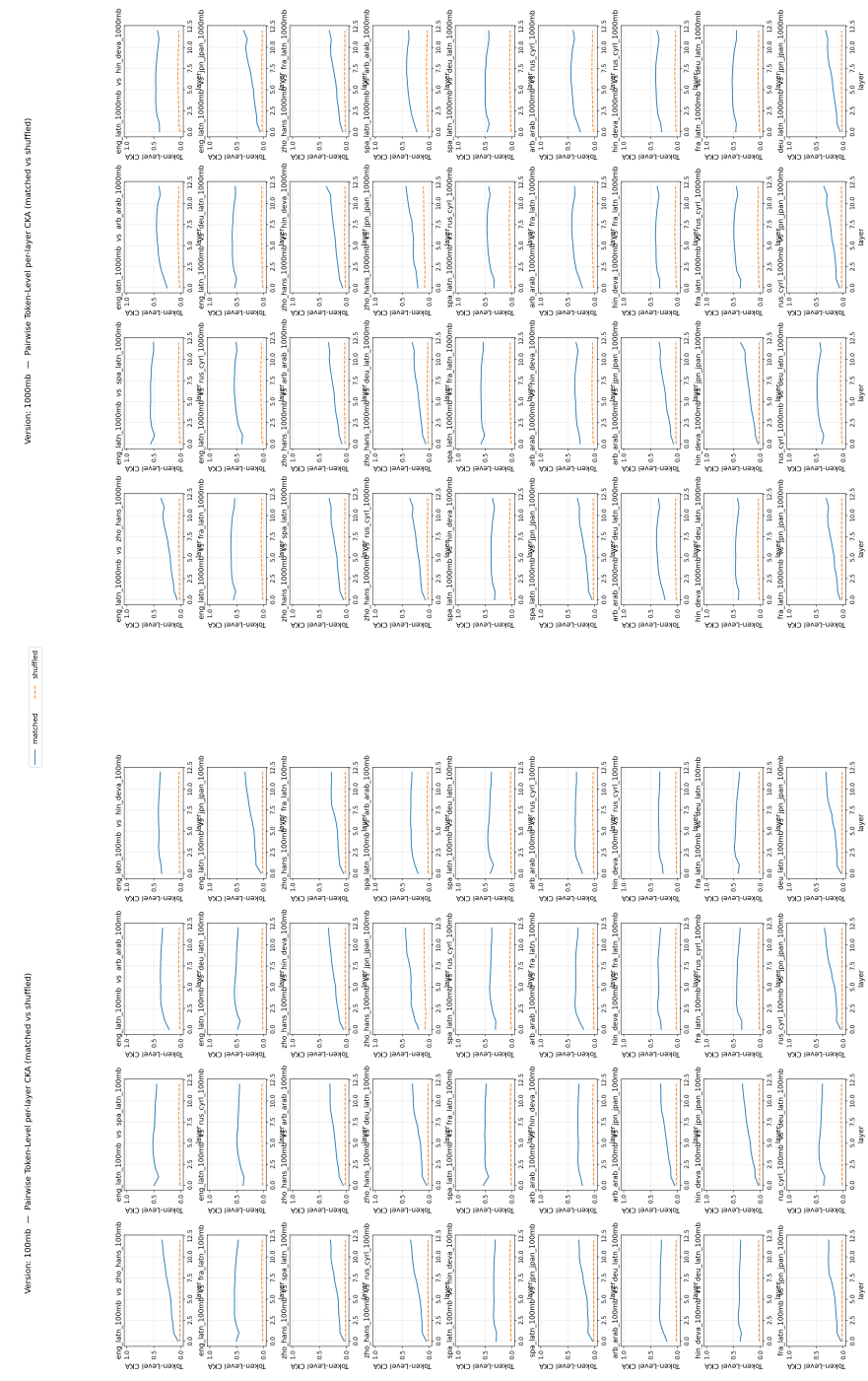

Figure 11: Per-layer cross-lingual representational similarity across all language pairs and training scales. Token-level alignment version.

E Activation Patching Across All Layers

Table 3 summarizes the effects of activation patching across the layers of the German model under four conditions. For the within-language condition (DE→DE, Hund→Katze), early layers show small but significant positive changes in continuation log-likelihood (Δ), with the percentage of positive items increasing up to layer 4, after which deeper layers exhibit negative Δ values and a sharp drop in positive cases. In the cross-model same-concept condition (FR→DE, chien→Katze), the effects are generally weaker and mostly non-significant, indicating limited transfer of concept-specific activations from French to German. The cross-model wrong-concept condition (FR→DE, chat→Katze) displays a similar pattern to the same-concept case in early layers, but deeper layers show larger negative Δ and lower percentages of positive items, reflecting the interference of incorrect concept patching. Finally, the cross-model shuffled baseline (FR→DE, random→Katze) exhibits broadly similar trends across layers, providing a reference for statistical comparison. Overall, the table highlights that activation patching effects are strongest in intermediate layers for correct within-language mappings, while cross-model effects are attenuated and concept-sensitive.

Condition	Layer	Mean Δ	Std	p vs. shuf.	% pos
Within-language (DE→DE, Hund→Katze)	0	-0.005	0.546	$< 10^{-3}$	56.1
	1	-0.002	0.541	$< 10^{-3}$	56.3
	2	0.024	0.520	0.002	58.5
	3	0.035	0.504	0.716	59.4
	4	0.037	0.485	0.871	60.8
	5	0.013	0.471	0.012	58.9
	6	0.003	0.454	0.014	58.8
	7	-0.047	0.406	0.048	54.0
	8	-0.074	0.388	0.211	50.3
	9	-0.085	0.384	$< 10^{-3}$	48.8
	10	-0.095	0.378	$< 10^{-3}$	47.0
	11	-0.158	0.347	$< 10^{-3}$	22.5
Cross-model same concept (FR→DE, chien→Katze)	0	-0.174	0.605	0.252	40.0
	1	0.091	0.528	0.994	57.7
	2	0.107	0.511	0.823	59.4
	3	0.038	0.480	0.595	55.1
	4	0.052	0.445	0.613	54.5
	5	-0.037	0.391	0.869	47.1
	6	-0.049	0.372	0.905	45.6
	7	-0.086	0.315	0.837	42.0
	8	-0.101	0.305	0.706	38.9
	9	-0.161	0.307	0.750	33.0
	10	-0.216	0.306	0.985	23.6
	11	-0.325	0.363	0.649	8.1
Cross-model wrong concept (FR→DE, chat→Katze)	0	-0.188	0.632	0.505	40.8
	1	0.082	0.530	0.717	56.3
	2	0.098	0.527	0.911	58.7
	3	0.029	0.483	0.886	55.5
	4	0.025	0.457	0.489	51.7
	5	-0.079	0.399	0.042	42.5
	6	-0.091	0.384	0.015	42.1
	7	-0.128	0.334	0.004	34.6
	8	-0.131	0.332	0.023	36.9
	9	-0.185	0.328	0.065	29.8
	10	-0.243	0.311	0.074	21.6
	11	-0.371	0.361	0.034	8.3
Cross-model shuffled (FR→DE, random→Katze)	0	-0.208	0.608	–	37.0
	1	0.091	0.529	–	58.9
	2	0.101	0.524	–	59.6
	3	0.026	0.485	–	54.9
	4	0.041	0.452	–	57.0
	5	-0.041	0.396	–	48.0
	6	-0.047	0.376	–	47.4
	7	-0.082	0.331	–	43.9
	8	-0.096	0.316	–	40.5
	9	-0.156	0.323	–	34.4
	10	-0.215	0.327	–	25.4
	11	-0.333	0.383	–	8.9

Table 3: **Activation patching effects across layers.** We report the mean and standard deviation of the change in continuation log-likelihood (Δ) for the German model under four conditions: within-language patching (DE→DE), cross-model patching from French without explicit alignment (FR→DE) using the same concept (*riz*) or a mismatched concept (*maïs*), and a shuffled French-source baseline. p-values compare each condition against the shuffled baseline at the same layer. "% pos" denotes the percentage of items with $\Delta > 0$.