BEYOND LINEAR CONCEPTS: DISCOVERING AND ALIGNING NON-LINEAR CONCEPT MANIFOLDS IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding information processing in large language models (LLMs) requires dissecting the geometric organization of their internal token representations. While existing mechanistic interpretability (MI) methods seek to extract human-understandable concepts, they are constrained by a strong linearity assumption—an assumption increasingly challenged by evidence of non-linear feature manifolds. In this work, we move beyond linear concepts by adapting Non-Linear Multi-Dimensional Concept Discovery (NLMCD) from computer vision to token-level LLM activations, modeling concepts as low-dimensional manifolds. To enable quantitative comparison of these independently discovered concept manifolds across layers and models, we introduce a concept-based alignment (CBA) score, a generalized Rand index that measures geometric proximity without requiring explicit feature matching. Our analysis yields six key findings: (i) a neighboring-layer sanity check shows CBA is more sensitive than PCA- or CKA-based linear baselines; (ii) layer×layer alignment matrices reveal two characteristic block structures concentrated in intermediate and late layers, consistent across models and obscured by linear metrics; (iii) concept composition remains syntax-dominated through most of the network before giving way to increasingly mixed syntactic–semantic concepts in later layers, with increasing output-orientation toward the final layers; (iv) multilingual concept sharing between English and Mandarin is training-dependent rather than universal — strongest in Qwen, weaker in Llama, and absent in GPT-2; (v) inter-model alignment mirrors this structure: strong correspondence between same-family Qwen models of different scale, consistent with the block structures in (ii), but weak alignment across model families; and (vi) tracing representational alignment across the training stages of the Tulu-3 pipeline shows alignment is highest between adjacent stages, validating CBA, with the largest shift occurring between the base model and the SFT stage, while subsequent preference-alignment stages (DPO, RLVR) leave early layers largely unchanged and RLVR mostly preserves DPO's concepts in late layers.

1 INTRODUCTION

Large Language Models (LLMs) are gaining traction in a variety of tasks, most notably as assistant chatbots, as agents that solve complex tasks, and as coding support, making them among the most influential machine-learning paradigms of the present day. As LLMs are trained, scaled, finetuned, and adapted across languages, a recurring question is whether and how their internal representations change: *Does finetuning reorganize concepts or merely refine them? Do larger models within a family preserve the representational structure of smaller ones? Do multilingual models encode shared, language-agnostic concepts, or largely separate ones per language?* Answering such questions requires comparing internal representations across layers, model scales, training stages, and languages — a task that depends entirely on having a reliable notion of representational alignment.

The standard tools for this comparison are global alignment metrics such as Centered Kernel Alignment (CKA) or canonical correlation analysis (CCA). These methods compare entire activation spaces under an implicit linear representation hypothesis: similarity is measured via linear projections or kernel similarities over the full activation vector, rather than over the discrete, interpretable units

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(concepts, features) that individual activations may encode. This is a poor fit for evidence that some computationally relevant structures in LLMs are intrinsically multi-dimensional and non-linear (Engels et al., 2025), and it makes such metrics comparatively insensitive: two representation spaces can appear highly similar under CKA/CCA even when their underlying concept structure diverges substantially at a finer grain.

We investigate an alternative based on *Non-Linear Multi-Dimensional Concept Discovery (NLMCD)* (Vielhaben et al., 2026), which treats concepts as density-supported manifolds rather than linear directions. It builds on evidence, including from sparse autoencoders (SAEs), that concept-level decompositions of LLM activations are a productive unit of comparison (Bricken et al., 2023; Gao et al., 2024), while relaxing the linearity assumption most such methods share. Activations are embedded with UMAP, clustered with HDBSCAN, and represented by fuzzy token-to-concept memberships. These memberships enable Concept-Based Alignment (CBA), a fuzzy generalization of the Rand index that compares the pairwise relations induced by two independently discovered concept spaces — without requiring the same architecture, tokenizer, or training stage on both sides, unlike global alignment metrics that presume a shared coordinate system. While the discovered concepts could also support other uses, such as mechanistic interpretability, we focus entirely on their use for alignment, and present a comprehensive view of intra- and inter-model representational alignment across a variety of LLMs.

Our technical contributions and most important findings are: (R0) We adapt NLMCD to residual stream hidden activations and validate the resulting clustering via coverage, reconstruction, and density-based quality metrics. (R1) We propose a neighboring-layer sanity check showing CBA is more sensitive than PCA- and CKA-based baselines. (R2) We find qualitatively similar block structures in layer×layer alignment across models, with two characteristic blocks in intermediate and late layers. (R3) We characterize discovered concepts quantitatively in terms of syntactic/semantic content, finding a relative decrease of purely syntactic concepts in favor of joint syntactic-semantic concepts toward later layers, alongside increasing output-orientation toward final layers. (R4) We find multilingual concepts shared across an aligned English–Mandarin corpus, strongest in Qwen, weaker in Llama, and absent in GPT-2. (R5) We find strong inter-model alignment between same-family Qwen models of different scale, consistent with the block structures identified in (R2), but weak structural alignment across model families. (R6) We trace representational alignment across Tulu-3 training stages: alignment is highest between adjacent stages, validating CBA; the largest shift occurs between the base model and SFT; and preference-alignment (DPO, RLVR) leaves early layers largely unchanged while RLVR mostly preserves DPO's concepts in late layers. Fig. 1 summarizes the method and the three main uses studied in this paper.

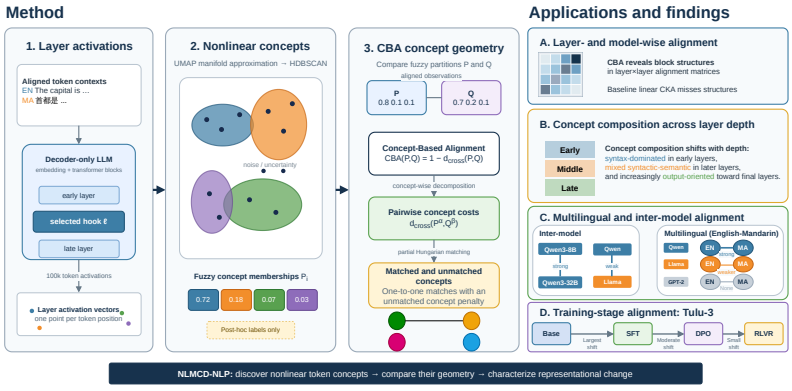


Figure 1: Overview of NLMCD-NLP. Layer-wise token activations are clustered into nonlinear concepts with fuzzy memberships. Using a concept-based alignment (CBA) score, we conduct several experiments on these concept manifolds.

2 Non-linear concept discovery

Activation extraction. For every model and language, we process 5,000 sentences in batches of eight, truncate them to 128 tokens, and retain a shared reservoir of 100,000 token positions using

reservoir sampling (Vitter, 1985). A common sampling stream ensures that row i denotes the same token position at every layer. We record the residual output of each of the transformer blocks for all layers of the respective model. We look at models with a hidden dimension ranging from 768–5,120 across 12–64 layers. Residual activations are L2 normalized before concept discovery.

UMAP, HDBSCAN, and fuzzy memberships. We follow the concept discovery approach of Vielhaben et al. (2026) even including hyperparameter settings. For each layer independently, UMAP maps the activation reservoir to 50 dimensions with 30 neighbors, minimum distance 0.01, and seed 42 (McInnes et al., 2018). HDBSCAN then discovers density-supported clusters with minimum cluster size and minimum samples both set to 50, leaf cluster selection, and explicit noise assignments (Campello et al., 2013). We use one fixed profile across layers and checkpoints so that comparative results are not confounded by per-layer hyperparameter optimization. Let $P \in [0,1]^{N \times C}$ be the resulting proximity matrix. P_{ic} is token i's soft membership in concept c. Rows are not forced to sum to one. Missing mass represents HDBSCAN noise or uncertainty.

Concept-Based Alignment. CBA compares two fuzzy partitions P and Q defined on aligned observations (Hüllermeier et al., 2012; Vielhaben et al., 2026). For two fuzzy clusterings P and Q and feature vectors $(\phi_i)_{i=1\dots N}$, define the membership similarity $d_{ms}(P(\phi_i), P(\phi_j)) = 1 - \frac{1}{2}\|P(\phi_i) - P(\phi_j)\|_1$. The *cross-distance* is defined by

$$d_{\text{cross}}(P,Q) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} |d_{ms}(P(\phi_i), P(\phi_j)) - d_{ms}(Q(\phi_i), Q(\phi_j))|. \tag{1}$$

We define the *Concept-based alignment (CBA)* score as $\text{CBA}(P,Q) = 1 - d_{\text{cross}}(P,Q)$. CBA equals one when both spaces induce identical pairwise membership relations and reduces to the Hüllermeier Rand index for the case of crisp concept assignments. As in (Vielhaben et al., 2026), we also define an *alignment score between two clusters* P^α and Q^β from two clusterings P and Q via

$$d_{\text{cross}}(P^\alpha, Q^\beta) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left| |P^\alpha(\phi_i) - P^\alpha(\phi_j)| - |Q^\beta(\phi_i) - Q^\beta(\phi_j)| \right|. \tag{2}$$

Selected model families We evaluate nine checkpoints from four families: GPT-2 (Radford et al., 2019), our MI baseline for comparison with prior SAE/NMF work and our English-only contrast for the multilingual analysis (Result 4); Llama-3.1-8B-Instruct (Grattafiori et al., 2024), our main cross-architecture reference (Results 2, 3, 5), kept over Llama 4 since the latter is multimodal MoE; and Qwen3-4B/8B/32B (Yang et al., 2025), providing a scale sweep (Result 2) and multilingual coverage (Result 4). Qwen3-8B and Llama-3.1-8B-Instruct form our matched-scale cross-architecture pair (Results 2–4), chosen for their contrasting data provenance and tokenizers. We use dense Qwen3 rather than Qwen3.5/3.6, which are natively multimodal and would confound scale/language effects with training-paradigm differences; Qwen3 remains the most recent text-only checkpoint in this family, a tradeoff we make explicit.

Finetuning case study For the finetuning-drift case study (Result 6), we use the Tulu-3-8B suite (base, SFT, DPO, final) (Lambert et al., 2025), built on Llama-3.1-8B and, unlike most released models, exposing every intermediate training stage. This lets us separate drift due to supervised finetuning from drift due to preference/RL optimization (DPO, then RLVR), and its shared initialization with our Llama-3.1-8B checkpoint additionally enables a secondary comparison against Meta's own RLHF-tuned Instruct release. More details on the desiderata for model choice can be found in App. A.1.

Experimental setup Concept discovery uses 5,000 sentences per language. English is sampled from WikiText-103; the joint English–Mandarin discovery corpus is described in Result 4. Coverage is the non-noise token fraction, and Density-Based Clustering Validation (DBCV) assesses cluster separation (Moulavi et al., 2014). Embedding diagnostics include raw NRMSE over 20,000 sampled pairs. For the R1/R2 within-model comparisons, CBA, PCA-512, linear CKA, and RBF CKA use the same deterministic sample of 5,000 token rows across layers. PCA-512 treats positive principal-component scores as soft concepts.

3 Related work

Alignment measures Representational alignment measures have been categorized in recent reviews (Sucholutsky et al., 2025; Klabunde et al., 2025). The most widely used alignment measure for neural representations is centered kernel alignment (CKA) (Kornblith et al., 2019), which is based on the Hilbert-Schmidt Independence criterion (Gretton et al., 2005). It is invariant to orthogonal transformations and isotropic scaling. Often linear kernels are used in CKA for computational efficiency and similar performance compared to non-linear kernels. In vision, NLMCD (Vielhaben et al., 2026) has been proposed as a way to capture the more finegrained alignment scores that incorporate the finestructure. The score is based on a generalized Rand-index between fuzzy clusterings. This work explores the use of the NLMCD alignment measure in the context of LLMs.

Zero-shot representation alignment A separate line of work addresses cross-space alignment without assuming explicit feature correspondence or a shared coordinate system. Moschella et al. (2023) propose relative representations, enabling zero-shot stitching between independently trained latent spaces. Alvarez-Melis & Jaakkola (2018) instead frame cross-lingual word embedding alignment as an unsupervised Gromov-Wasserstein optimal transport problem. Both approaches share our goal of comparing representation spaces without explicit matching, but operate at a coarser granularity: they align whole embedding spaces or point clouds, whereas CBA compares discovered concept manifolds using alignment scores to be attributed to specific, interpretable concepts rather than reported as a single global statistic.

Concept discovery The NLMCD hinges on representational substructures, which are often referred to as concepts. The most commonly considered structures are linear subspaces, in the case of LLMs typically discovered by means of sparse autoencoders with a linear decoder model. This generalized concept definition aligns with recent findings on non-linear substructures in LLM representation spaces (Levy & Geva, 2024; Geiger et al., 2026; Wurgaft et al., 2026a;b). Instead of combining SAE directions, this work takes a more direct path and leverages NLMCD's concept definition as general manifolds.

Representational alignment in LLMs Cross-architectural alignment has been investigated in (Piepereit, 2026) revealing that trained models show above-random representational alignment, which is weak in absolute terms and strongly architecture-dependent. It has also been used in (Potter et al., 2026) to assess representational similarity in multi-agent interactions. Shen et al. (2026) investigated latent reasoning using CKA. CKA-based alignment scores have also been used for specific applications, such as REEF (Zhang et al., 2025) and AWM (Zeng et al., 2026) to derive representation-based fingerprints for LLMs. Overall, CKA has been used to assess representational alignment within and across LLMs, but CKA as alignment measure has rarely been contested. Second, there is not a dedicated alignment study that investigates representational similarity intra- and inter-model in a comprehensive fashion. This is the gap that this submission aims to fill.

4 Results

Result 0: Concept Quality Across Layers Across models, we observe a median coverage, i.e., fraction of non-noise tokens, of 0.4469, substantially higher than the $\sim$0.2 typical of vision foundation models (Vielhaben et al., 2026), likely reflecting the greater semantic richness of language relative to vision. Coverage declines from $\sim$0.65 in early layers to $\sim$0.4 in later layers across all models, consistent with the increasing difficulty of capturing the richer representational structure of higher-level features; even so, late-layer coverage remains substantial by vision-domain standards. DBCV, a density-based validity index contrasting intra- vs. inter-cluster density, bounded in $[-1, 1]$ with higher indicating better-separated clusters, shows a median of 0.2743, somewhat lower than the $\sim$0.5 typical in vision foundation models, and correlates strongly with coverage, again pointing to superior clustering quality in earlier layers. Reconstruction quality (NRMSE, see App. A.2) follows the same trend, falling from 10-12 in early layers to 3-4 in later layers, converging to values comparable to those reported in the vision setting. Taken together, these results support the technical validity of the concept discovery step itself. We highlight the robustness of the NLMCD approach which transfers from vision to LLMs without the necessity to adapt hyperparameter settings.

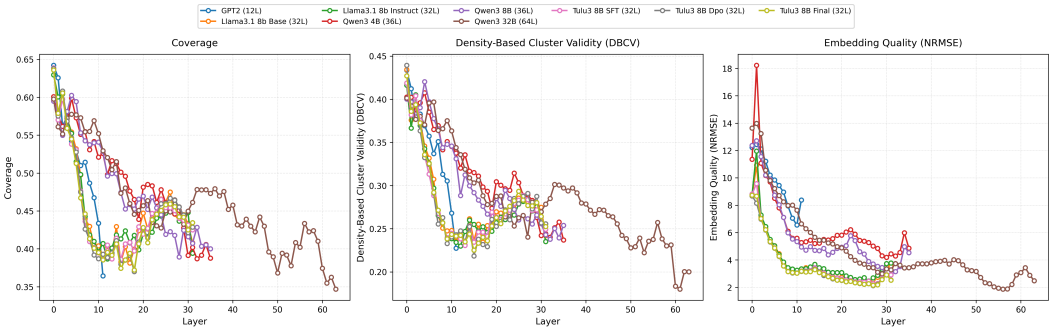


Figure 2: Concept Quality Across Layers: The quality of the concepts extracted from each model is shown across the network depth. For each model, concepts are extracted and evaluated at every layer. (Left) Coverage: the fraction of non-noise tokens. (Middle) Density-based clustering validation (DBCV), which assesses cluster separation. (Right) Embedding quality (NRMSE): the raw normalized root-mean-square error between pairwise distances in the activation space and the reduced embedding.

Result 1: Sanity check As a sanity check for the soundness of our alignment measure, we exploit the assumption that adjacent layers should show the strongest representational correspondence: for each source layer, we ask whether its best-aligned target layer under a given alignment method, excluding self-alignment, is a direct neighbor. The *NN alignment score* reports the fraction of layers for which this holds; *Neighbor z* additionally reports the maximum alignment score (z-standardized across rows) of the direct neighbors to rule out uninformative flat alignment profiles. We compare CBA to CKA with RBF kernel as non-linear baseline and two linear baselines, linear CKA and PCA with 512 dimensions, averaged over 9 model checkpoints (Tab. 1; per-model results in App. Tab. 4). All four methods show nearly perfect NN alignment scores, with slight advantages for the non-linear alignment measures CKA(RBF) and CBA, with 0/1 mispredicted neighbors across all layers and models. The Neighbor z score represents the more informative metric in this case and clearly identifies CBA as most informative measure. While the two non-linear methods perform on par in terms of NN Alignment Score, CBA enjoys the additional benefit that it yields concepts as substructures and allows for a more finegrained characterization of the alignment between these substructures.

Method	NN Alignment Score ↑	Neighbor z ↑
PCA (512-d)	0.983 ± 0.025	1.867 ± 0.244
CKA (linear)	0.978 ± 0.029	1.287 ± 0.486
CKA (RBF)	$\mathbf{1.000 \pm 0.000}$	1.714 ± 0.182
CBA (NLMCD)	0.997 ± 0.009	$\mathbf{2.515 \pm 0.256}$

Table 1: Nearest-neighbor alignment score and best-neighbor z-score averaged equally over 9 model checkpoints (mean $\pm$ population standard deviation). All four methods recover adjacent-layer correspondence almost perfectly, so the NN alignment score alone provides limited discrimination between them. Neighbor z measures the strongest neighbor relative to the row mean and standard deviation. Per-model results are reported in App. Tab. 4.

We also compare the expressiveness of each metric choice in Fig. 3. When using linear CKA as the alignment metric we can identify two clear block structures but all finer details between layers 6-36 are not visible. In contrast both the PCA-512 and NLMCD methods reveal structures that linear CKA misses. RBF CKA also reduces the near-saturation seen with linear CKA. This stresses that the choice of the alignment metric strongly affects the observable substructures within the model.

Result 2: Layer-wise alignment Fig. 4 shows the CBA layer-alignment matrices for four models spanning different model families and parameter counts. All matrices exhibit a pronounced diagonal structure, with model-specific variations in the arrangement of aligned layers. For the larger-scale models, except for GPT, the matrices display a band-diagonal structure in the first region, followed by a block structure, a second band-diagonal region, and a final block structure. These patterns

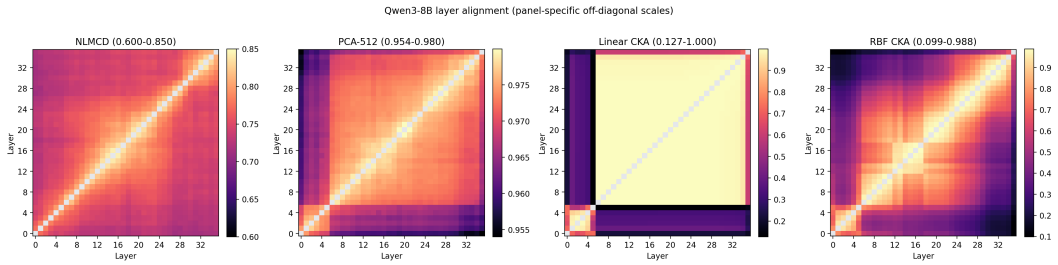
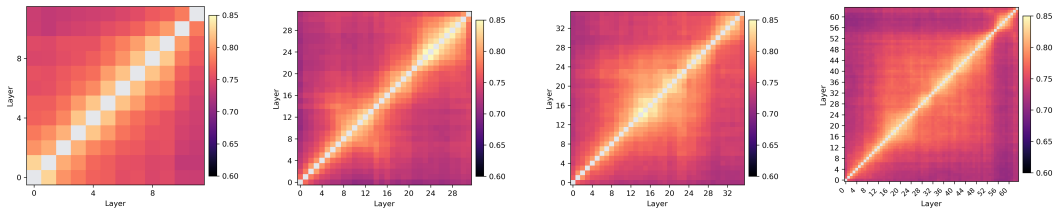


Figure 3: Qualitative comparison of alignment plots for Qwen3 8B

are qualitatively similar across the models. All Qwen models share a larger block structure in their middle layers. Compared with the 8B models, the 32B models exhibit a more finegrained structure in the central block. Overall, several larger models exhibit a similar sequence of band-diagonal and block-structured regions, suggesting recurring alignment patterns across model families, see Result 5 for details.

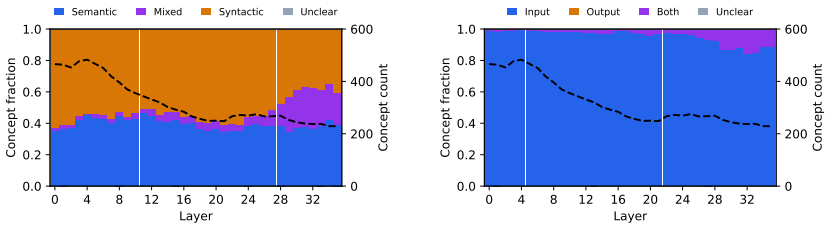


(a) GPT2 (b) Llama3.1 8B Instruct (c) Qwen3 4B (d) Qwen3 32B

Figure 4: CBA layer alignment. Shared color scale $[0.60, 0.85]$; self-alignment diagonals masked. Layer indices are zero-based. All panels use 5,000 samples. Refer to the first subplot in Fig. 3 for the Qwen3 8B result and to Fig. 9 for an extended plot comparing all models.

Result 3: Descriptive/qualitative insights We compare properties of annotated concepts aggregated over all residual layers for Llama-3.1-8B-Instruct and Qwen3-8B. Those models were chosen as they are comparable in parameter count and number of layers. Concepts were annotated by a labeling LLM along two axes: concept type (syntactic, semantic, mixed, or unclear) and evidence orientation (input, output, both, or unclear). Orientation describes which token and next-token evidence best explains the concept. Semantic and syntactic concept percentages are roughly on-par in Llama (45.7% and 45.1%), while syntactic concepts dominate in Qwen (53.2% versus 39.8% semantic). Regarding the orientation annotation results, input-oriented concepts clearly dominate in both models (93.4% and 96.3% respectively), while concepts with both input and output evidence account for 6.6% and 3.7%. Output-oriented concepts are omitted from the table as neither model has any, see Tab. 5 for details. Fig. 5 and App. Fig. 10 summarize the layer-dependence of both quantities. The fraction of syntactic concepts stays relatively high across the first two thirds (Qwen)/the first half (Llama) of the model and then shrinks as the fraction of jointly syntactic and semantic concepts starts to rise. This is partly consistent with probing studies on BERT, which found that syntactic information is more locally encoded in earlier layers while semantic tasks benefit from later-layer representations (Tenney et al., 2019). Our unsupervised concepts show the syntactic-dominance side of this pattern, but show a more nuanced picture, with joint semantic/syntactic concepts becoming more prevalent in later layers. For orientation-annotated concepts the fraction of output-oriented concepts stays close to 0% across all layers. There is, however, a clear increase of concepts annotated as having both input and output evidence towards the later layers in both models.

Result 4: Multilingual activation In order to quantify the amount of multi-lingual activation taking place within the selected models, we measure joint parallel activation on a parallel corpus of English and Mandarin sentences. To do this, we first conduct concept discovery on a separate joint English-Mandarin text corpus based on English WikiText records and Mandarin Wikipedia sentences. 5,000 sentences are taken in streaming order after filtering from each corpus for concept discovery.



(a) Qwen3-8B: semantic/syntactic

(b) Qwen3-8B: input/output

Figure 5: English concept categories and evidence orientation over residual depth. Bars show concept fractions (left axis). The dashed black line shows the total number of annotated concepts at each layer (right axis). Analogous plots for Llama 3.1 8B are shown in Fig. 10.

After this a deterministic random 5,000-pair sample of Tatoeba English-Mandarin translations is used as a parallel sentences probe. Concept assignments are computed on the joint clusters for each token in both sentences. Memberships are summed over tokens within each sentence and normalized across concepts to obtain distributions p_{en} and p_{zh}. Shared mass is $\sum_c \min(p_{\text{en},c}, p_{\text{zh},c})$, averaged over valid translation pairs.

Model	Paired mass	Shuffled mass	Mean excess	Peak layer	Peak excess
GPT-2	0.141	0.140	0.0009	5	0.0039
Llama-3.1-8B-Instruct	0.176	0.152	0.0242	2	0.0980
Qwen3-4B	0.245	0.196	0.0487	7	0.1187
Qwen3-8B	0.244	0.201	0.0430	8	0.1012
Qwen3-32B	0.286	0.247	0.0392	10	0.0817

Table 2: English–Mandarin shared concept mass and shuffled-pair controls over residual layers; peak layer (zero-based) indicates the layer with the largest excess with excess indicated in peak excess.

Tab. 2 shows the results of the multilingual activation experiment. Paired mass is the raw shared mass computed from the parallel sentences probe. However, especially in the larger models we observed layer-specific peaks that did not seem to correlate with multilingual activation, see App. Fig. 11. We therefore computed a baseline using 200 permutations of the Mandarin sentences per layer, recomputing shared mass and averaging over permutations (shuffled mass). Mean excess is paired mass minus shuffled mass, averaged equally over residual layers. Paired mass and shuffled mass in the table are also layer averages.

In line with our expectation the excess shared mass is close to zero for GPT-2, our predominantly English-trained comparison model and increases for the multilingual Llama and Qwen models. The mean excess increases from 0.0242 to 0.0487 from Llama to Qwen3-4B. After that there is a slight drop of mean excess for Qwen3-8B (0.0430) and Qwen3-32B (0.0392). The peak occurs at layer 7 for Qwen3-4B, layer 8 for Qwen3-8B, and layer 10 for Qwen3-32B. The finding that multilingual concept sharing between typologically distant languages is highly training-dependent rather than a universal property provides an empirical counterpoint to the Platonic Representation Hypothesis (Huh et al., 2024), in line with prior work on linear SAE representations (Brinkmann et al., 2025).

In addition to reporting the mean excess for shared mass across layers we also annotated multilingual concepts using a labeling LLM, labeling them as either semantic, syntactic or mixed. At each layer, a concept is selected as translation-shared if active in both sentences of at least 2% of valid translation pairs and at least five pairs. These are 100 pairs with our 5,000 valid pairs. Each sentence independently satisfies the active criterion through top-10 membership or a normalized concept score ≥ 0.03. The labeler uses bilingual training tokens and representative parallel examples. Semantic denotes shared meaning, syntactic includes grammar and surface form, mixed requires both, and unclear denotes insufficient evidence. Fig. 6 shows the results of the multilingual concept annotation. All three models show an early-layer peak in semantic shared-mass excess that gradually decreases towards the middle layers. Qwen3-8B and Qwen3-32B also show a local syntactic peak within their first four layers. Among labeled translation-shared concepts, semantic excess is prominent in

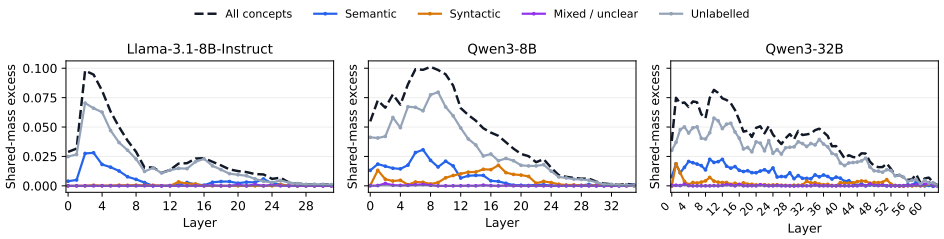


Figure 6: Semantic, syntactic, mixed, and unlabeled shared-concept mass above shuffled-pair controls for Llama-3.1-8B-Instruct, Qwen3-8B, and Qwen3-32B. Negative excess is retained; all panels share the same vertical scale.

early layers. Much of the total excess is unlabeled, so this does not establish semantic dominance across all concepts. Shared punctuation and sentence structure are possible explanations for syntactic contributions.

Result 5: Cross-model alignment Fig. 7 shows CBA matrices for comparisons between Qwen and Llama model variants. Because Qwen and Llama use different tokenizers, we align their tokens using character spans within the same sentence. We retain a target token only when sampled source-token spans lie wholly within and together cover its complete span. The source tokens' concept memberships are then mean-pooled before comparison. All R5 CBA comparisons use 1,024 aligned observations per model. Both within-family Qwen comparisons exhibit pronounced depth-structured alignment. High CBA scores form a broad diagonal band, indicating that layers at comparable relative depths encode similar concepts. The alignment comprises an early band, a broad middle block, and a smaller high-alignment region near the output layers; the main difference between Qwen 8B and 32B is an extended middle block. This pattern persists despite differences in scale and layer count, suggesting that Qwen variants preserve a common progression of internal representations. In contrast, Llama 8B Instruct versus Qwen 8B yields substantially lower CBA scores, although a weak diagonal and middle-layer block structure remain. This indicates limited, locally structured cross-family concept alignment, consistent with prior work reporting weak overall cross-family but strong within-family alignment (Piepereit, 2026). Nevertheless, the result shows that the internal organization does not look similar, see Result 2, but rather supports the hypothesis of corresponding functional blocks even across different model families.

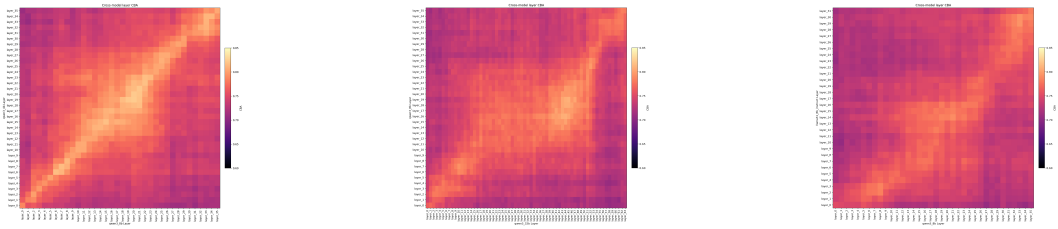


Figure 7: CBA Similarity Scores between models. Left: Qwen 4b vs 8b, Center: Qwen 8b vs. 32b, Right: Llama 8b instruct vs. Qwen 8b

Result 6: Finetuning case study We compare the three finetuned Tulu-3-8B variants with their base model, Llama-3.1-8B. Fig. 8 reports CBA scores for corresponding layers across model pairs. Alignment between adjacent training stages is high, consistent with the claim that objective consistency supports consistent representational similarity patterns (Ciernik et al., 2025). DPO and the final RLVR model are particularly closely aligned, with scores remaining near 0.85–0.86 across most layers, suggesting that RLVR largely preserves the concept geometry established during DPO rather than reorganizing intermediate representations. Comparisons involving the SFT model show lower, intermediate alignment, indicating a larger representational change during supervised finetuning. This is consistent with parameter-level evidence on the same checkpoints: SFT induces dense updates, including from Llama-3.1-8B to Tulu-3-8B-SFT, whereas DPO leaves about 81% of the SFT model's

parameters unchanged (Mukherjee et al., 2025). Across all stages, alignment remains above 0.8, higher than typical cross-model alignment scores, suggesting that most concepts are formed during pretraining and are only moderately modified by subsequent finetuning.

Most model pairs show an alignment drop from layer 14 to layer 15, which is substantially larger in comparisons involving the base model. The effect is not confined to layer 15: alignment remains reduced until approximately layer 22, after which it gradually recovers and, for DPO–RLVR, eventually exceeds its layer-14 value. Alignment also drops sharply at the final layer in every comparison, consistent with increased specialization toward next-token prediction. The same pattern is visible in the alignment matrices in App. Fig. 12. Overall, alignment changes are larger relative to the base model than between successive finetuning stages, with DPO and the final RLVR checkpoint remaining most closely aligned. The fact that across all finetuning stages the drop occurs after layer 14 indicates that this layer is of particular importance for the behavior of the model that is supposedly adapted during post-training. To investigate the alignment drop between layers 14 and 15, we apply partial concept matching based on the Hungarian algorithm. We augment the concept-cost matrix with dummy assignments whose cost equals the match-acceptance threshold, set to 0.0025 in all experiments. This allows concepts to remain unmatched when all real assignments exceed that cost, rather than forcing high-cost one-to-one matches. Fig. 13 overlays the resulting matches on the pairwise cost matrices for SFT and DPO, where layer-15 shows a weaker and less coherent diagonal structure than layer 14. App. Tab. 6 shows representative SFT–DPO concept matches at layers 14 and 15. Layer-14 matches are generally more semantically coherent, with the highest-scoring pairs consisting predominantly of function- or structure-related tokens. In contrast, layer-15 matches are more heterogeneous and often semantically weaker; even highly ranked pairs can mismatch, e.g. "Name beginnings" vs. "Equality symbols". Layer 15 also yields more unmatched concepts in both directions. Together with the lower CBA alignment, these observations suggest that the layer-14 to 15 transition changes the concepts represented after DPO rather than uniformly reducing similarity.

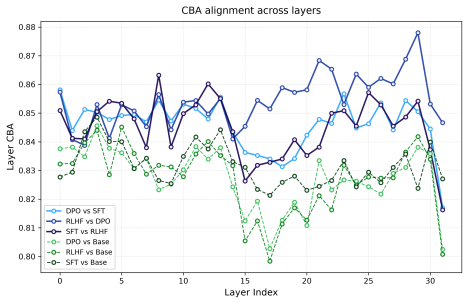


Figure 8: Pairwise same-layer CBA alignment across the Llama-3.1-8B base model and the SFT, DPO, and final RLVR checkpoints of Tulu-3.

5 Limitations and Future Work

First, we do not evaluate models above $\sim$32B parameters or mixture-of-experts architectures. This is partly a compute constraint, and partly because our core claims concern the *existence* and *cross-layer structure* of non-linear concept manifolds rather than their scaling behavior at frontier scale. However, we acknowledge both directions as very interesting directions for future research. Second, the computational complexity of the approach is dominated by the HDBSCAN clustering step. Given our time and budget constraints this limits the number of tokens to $O(100k)$. Devising a computationally less expensive alternative is therefore a worthwhile direction for future research. Third, the discovered concepts rely on the extraction corpus. Even though Wikipedia text provides a broad basis, some of the semantic concepts in intermediate layers are highly specific and further ones at a similar level of detail might not have been sufficiently activated. Finally, the approach can be used for more in-depth studies of the identified concepts along the lines sketched in Result 6.

AI USE STATEMENT

In this work, we used generative AI tools for tasks that require disclosure under current publication guidelines, specifically to assist with method implementation and to suggest improvements to existing implementations. We have not used generative AI tools for conceptual development, framing of results, formulation of hypotheses, design of methodology or experiments, interpretation of results, or the writing of proofs. Tasks such as generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, and providing critical ingredients for proving mathematical claims are not applicable to this work.

Additionally, we used generative AI tools for tasks that fall under recommended disclosure, including: (i) creating or editing software code (e.g., drafting boilerplate, refactoring, and suggesting alternative implementations, refinement of plotting routines); (ii) editing the manuscript to improve readability and language quality; (iii) assisting with model selection among models to be evaluated (e.g., summarizing documented capabilities, constraints, and typical use cases); and (iv) supporting routine writing subtasks such as proposing section structure, phrasing alternatives, and formatting references. No AI tool was used to generate scientific figures or images, to transcribe recordings, or to perform literature discovery or screening.

We have reviewed all AI-assisted work. In particular, LLM-generated or LLM-modified code was inspected line-by-line, executed in our experimental environment, and validated against unit tests and ablation checks by at least two authors; any suggestions affecting experimental logic were cross-checked against the original method descriptions and implementation notes. Language edits and structural suggestions were accepted only when they preserved the technical meaning and did not introduce new claims; all such changes were manually verified. We take responsibility for the final content of this work, including text, claims, code, and any other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All authors have read and agree to comply with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We make the full source code underlying our submission available in an anonymized code repository `https://anonymous.4open.science/r/NLMCD-NLP-C5E7`.

REFERENCES

David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 1881–1890, 2018.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards monosemanticity: Decomposing language models with dictionary learning, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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/.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu (eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 160–172, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

Laure Ciernik, Lorenz Linhardt, Marco Morik, Jonas Dippel, Simon Kornblith, and Lukas Muttenthaler. Objective drives the consistency of representational similarity across datasets. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10920–10948. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/ciernik25a.html.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.14860.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Atticus Geiger, Ekdeep Singh Lubana, Thomas Fel, Jack Merullo, Michael Byun, Owen Lewis, and Thomas McGrath. The world inside neural networks: How neural geometry will unlock understanding and control of ai. *Goodfire Research*, May 2026. URL https://www.goodfire.com/research/the-world-inside-neural-networks. https://www.goodfire.com/research/the-world-inside-neural-networks.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang,

Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov,

Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models. *arXiv preprint 2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint 2405.07987*, 2024. URL https://arxiv.org/abs/2405.07987.

Eyke Hüllermeier, Maria Rifqi, Sascha Henzgen, and Robin Senge. Comparing fuzzy partitions: A generalization of the rand index and related measures. *IEEE Transactions on Fuzzy Systems*, 20(3): 546–556, 2012. doi: 10.1109/TFUZZ.2011.2179303.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys*, 57(9):1–52, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint 2411.15124*, 2025. URL https://arxiv.org/abs/2411.15124.

Amit Arnold Levy and Mor Geva. Language models encode numbers using digit representations in base 10. *arXiv preprint*, (arXiv:2410.11781), 2024. URL https://arxiv.org/abs/2410.11781.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. URL https://arxiv.org/abs/1802.03426.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=SrC-nwieGJ.

Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 839–847. Society for Industrial and Applied Mathematics, 2014. doi: 10.1137/1.9781611973440.96. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611973440.96.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models, 2025. URL https://arxiv.org/abs/2505.11711.

Fernando Cardenas Piepereit. Architecture-dependent causal transfer of activation states across large language models. *arXiv preprint arXiv:2608.16347*, 2026.

Yujin Potter, Seun Eisape, Shiyang Lai, Alexander Huth, James Evans, Been Kim, Jacob Eisenstein, Dawn Song, and Alane Suhr. Representational similarity and model behavior in multi-agent interaction. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=vSKK7ffkfv.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Xuan Shen, Yizhou Wang, Yufa Zhou, Xiangxi Shi, Pu Zhao, Yanzhi Wang, and Jiuxiang Gu. Efficient reasoning with hidden thinking. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Hn81xuLaXY.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine Hermann, Kerem Oktar, Klaus Greff, Martin N Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas O'Connell, Thomas Unterthiner, Andrew Kyle Lampinen, Klaus Robert Muller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Hiq7lUh4Yn. Survey Certification, Expert Certification.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline, 2019. URL https://arxiv.org/abs/1905.05950.

Johanna Vielhaben, Dilyara Bareeva, Jim Berend, Wojciech Samek, and Nils Strodthoff. Beyond scalars: Concept-based alignment analysis in vision transformers. *Advances in Neural Information Processing Systems*, 38:71379–71414, 2026.

Jeffrey S. Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, March 1985. ISSN 0098-3500. doi: 10.1145/3147.3165. URL https://doi.org/10.1145/3147.3165.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior. *arXiv preprint 2605.05115*, 2026a.

Daniel Wurgaft et al. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors. *arXiv preprint 2602.02315*, 2026b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin

Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint 2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Boyi Zeng, Lin Chen, Ziwei He, Xinbing Wang, and Zhouhan Lin. Awm: Accurate weight-matrix fingerprint for large language models. In *International Conference on Learning Representations*, volume 2026, pp. 66850–66873, 2026.

Jie Zhang, Dongrui Liu, Chen Qian, Linfeng Zhang, Yong Liu, Yu Qiao, and Jing Shao. Reef: Representation encoding fingerprints for large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 48092–48117, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/77e830bd802f4b63b682038d090a97b6-Paper-Conference.pdf.

A Appendix

A.1 Model Selection

We evaluate our approach on nine checkpoints drawn from four model families, chosen to jointly satisfy four criteria: (1) *architectural purity* — dense, purely text-trained decoders wherever a claim is language-specific, so that representational geometry is not confounded with joint vision-language training; (2) *comparability to prior mechanistic-interpretability work*, since Result 1 sanity-checks our measure against CKA and PCA baselines reported in the literature; (3) *targeted coverage* of the specific axis each result section requires (sanity-checking, scale, cross-architecture alignment, multilinguality, training stage); and (4) *feasibility under a fixed compute budget*, which rules out broad sweeps across many independent families or model sizes above $\sim$30B parameters. Tab. 3 summarizes the resulting suite and which result section each checkpoint supports.

GPT-2 (Radford et al., 2019). Serves as our mechanistic-interpretability anchor. Despite being the oldest model in our suite, it is the model against which the majority of existing SAE- and NMF-based concept-extraction results have been reported, making it essential for situating our manifold-based proximity measure relative to prior linear approaches (Result 1). Its small size and English-only pretraining also make it our contrast case in the multilingual analysis (Result 4).

Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Provides our primary cross-architecture reference point at the $\sim$8B scale (Results 2, 3, 5), and supplies the base weights underlying our finetuning-stage case study (see below). We retain this mid-2024 checkpoint rather than substituting a newer Llama release because Llama 4 (April 2025) moved to a mixture-of-experts, natively multimodal early-fusion architecture, which no longer satisfies our architectural-purity criterion; Llama-3.1-8B is thus the most recent Llama checkpoint compatible with our design.

Qwen3-4B, Qwen3-8B and Qwen3-32B (Yang et al., 2025). Supply our within-family scale sweep (Result 2) and our multilingual analysis (Result 4), holding architecture and training recipe fixed while varying parameter count roughly eightfold. Qwen3-8B additionally forms our matched-scale cross-architecture pair with Llama-3.1-8B-Instruct (Results 2–4), chosen as the comparison point because Qwen3 and Llama-3.1 differ substantially in data provenance, tokenizer, and multilingual coverage. We use the dense Qwen3 line rather than the more recent Qwen3.5/3.6 releases because, from Qwen3.5 onward, Alibaba's flagship checkpoints are trained natively multimodally via early fusion of vision and text tokens; using these would confound our scale and multilingual claims with differences in training paradigm rather than isolating parameter count and language as intended. Qwen3 dense models remain purely text-trained, cover 119 languages including Mandarin, and are, to our knowledge, the most recent checkpoints from any major open-weight family that satisfy our text-only constraint at time of writing — a recency/architectural-purity tradeoff we make explicit rather than an unexamined limitation.

Tulu-3-8B suite: base, SFT, DPO, final (Lambert et al., 2025). Supports our finetuning-drift case study (Result 6). Unlike most publicly released instruction-tuned models, which expose only base and final checkpoints, the Tulu 3 pipeline — built on Llama-3.1-8B base weights — publishes each intermediate training stage explicitly. This lets us attribute representational drift separately to supervised finetuning and to preference/RL-based optimization (DPO, then RLVR), rather than collapsing both into a single base-to-instruct delta, which a single additional checkpoint could not support. We chose this suite over comparable alternatives (e.g., OLMo 2) because it shares its base initialization with the Llama-3.1-8B checkpoint used elsewhere in our suite, allowing a direct comparison between Meta's RLHF-tuned Instruct release and AI2's independently-tuned Tulu 3 endpoint as a secondary check on how sensitive our drift measure is to the specific finetuning pipeline, not only its stages. We summarize the selected models in Tab. 3.

A.2 Embedding distance diagnostic

For $M = 20{,}000$ sampled pairs of distinct token rows, let d_m be the Euclidean distance between their L2-normalized residual activations and $\hat{d}_m$ the Euclidean distance between their saved UMAP embeddings. Pairs are drawn uniformly with seed 42; a row may occur in multiple pairs. We

Model	Release	HF identifier	Result section(s)
GPT-2 (124M)	02/2019	`openai-community/gpt2`	0, 1, 4
Llama-3.1-8B-Instruct	07/2024	`meta-llama/Llama-3.1-8B-Instruct`	0, 1, 2, 3, 4, 5
Qwen3-4B	05/2025	`Qwen/Qwen3-4B`	0, 1, 2, 4, 5
Qwen3-8B	05/2025	`Qwen/Qwen3-8B`	0, 1, 2, 3, 4, 5
Qwen3-32B	05/2025	`Qwen/Qwen3-32B`	0, 1, 2, 4
Llama-3.1-8B (base)	06/2024	`meta-llama/Llama-3.1-8B`	1, 6
Tulu-3-8B (SFT)	11/2024	`allenai/Llama-3.1-Tulu-3-8B-SFT`	1, 6
Tulu-3-8B (DPO)	11/2024	`allenai/Llama-3.1-Tulu-3-8B-DPO`	1, 6
Tulu-3-8B (RLVR)	11/2024	`allenai/Llama-3.1-Tulu-3-8B`	1, 6

Table 3: Model suite used in this work.

normalize the error by the mean original distance $\bar{d} = M^{-1} \sum_m d_m$:

$$\mathrm{NRMSE}_{\mathrm{raw}} = \frac{\sqrt{M^{-1} \sum_m (d_m - \hat{d}_m)^2}}{\bar{d}} \tag{3}$$

Fig. 2 reports this score.

A.3 Concept annotation

We use Mistral-Small-24B-Instruct-2501 to label each cluster from its up to 20 most frequent decoded tokens, four causal sentence contexts, and ten most frequent next-token predictions derived from the extraction corpus. Greedy decoding produces a 2–6-word English label. A separate prompt classifies the English concepts in Result 3 by type (semantic, syntactic, mixed, or unclear) and evidence orientation (input, output, both, or unclear). A third prompt classifies translation-shared concepts in Result 4. All three tasks use a single user message formatted with the labeler's chat template and greedy decoding (`do_sample=False`). The generation limits are 30 tokens for concept names and 100 tokens for each classification task. Prompts, evidence, raw generations, and cleaned labels are retained. Labels are hypotheses for inspection that do not enter UMAP, HDBSCAN, CBA, or any other discovery metric. The R3 and R4 category summaries do depend on these annotations (Bills et al., 2023).

A.3.1 Concept names

This prompt supplies the descriptive labels used for concept inspection and matching examples. These names are not supplied to the R3 or R4 classifiers.

> You are analysing the internal representations of a language model (<model identifier>, <number of layers> transformer layers total). You are looking at <layer description>. The tokens below come from <corpus description> text, split into subwords by its tokenizer (<tokenizer revision>).
>
> The following subword tokens cluster together in the model's activation space at this layer — meaning they trigger a similar internal representation pattern.
>
> Input tokens: <top input tokens>
>
> Predicted next tokens (what the full model predicts will follow these input tokens): <top predicted next tokens>
>
> <depth-dependent note>
>
> Example sentences:
>
> - '<token>' in: <causal sentence prefix>
>
> <up to four examples>
>
> Task: Give a short label (2–6 words) naming what these tokens have in common.
>
> Guidelines:
>
> • For morphological fragments (prefixes, suffixes, stems), describe their grammatical function rather than calling them 'fragments' or 'subwords' (e.g. 'verb suffix -ing', 'plural noun endings -s/-es', 'past tense -ed').

> • Avoid vague catch-alls such as 'abbreviations', 'subwords', 'legal terms', 'fragments', or 'German text' unless all tokens genuinely share that specific property.
> • Prefer a specific, distinctive label over a generic one.
> • Output ONLY the label — no explanation, no punctuation at the end, no quotes.

The layer description identifies the residual block and its relative depth. The next-token note varies with rounded relative depth, as follows; this is part of the labeling prompt, not a measured property of the layer. *Early (below 40%).*

> Note: at this early depth the predicted next tokens reflect the full model but may not align closely with the current representations — use them as a secondary signal.

Intermediate (40–69%).

> Note: at this intermediate depth the predicted next tokens are moderately informative about the layer's function.

Late (at least 70%).

> Note: at this late depth the predicted next tokens are closely connected to the layer's representations — weight them heavily when labelling.

For boundary-only or empty decoded tokens, the prompt additionally states:

> The visible token is only a standalone boundary/whitespace marker or otherwise decodes to an empty string. Use the causal sentence prefixes and predicted next tokens below to characterise the contextual representation captured by this cluster. Do not label it merely as an unknown concept.

When begin-of-sequence markers dominate the displayed tokens, an additional guideline states:

> Most tokens here are '<s>' (begin-of-sequence marker). Use the example sentences and predicted next tokens to characterise what type of content or article this cluster captures (e.g. 'Biographical article opening', 'Technical Wikipedia intro', 'Sentence-initial Die/The').

A.3.2 English concept type and evidence orientation (R3)

The input contains up to 20 input tokens, ten predicted next tokens, and four contexts. Each context string is truncated to its first 200 characters by the classifier. Concept names, model identity, corpus description, and layer depth are withheld. Type and orientation are requested jointly with the following prompt.

> You are classifying a concept cluster from a language model's internal representations using only the token and context evidence below.
> Top tokens in cluster: <top input tokens>
> Predicted next tokens: <top predicted next tokens>
> Example sentences (cut at the token position):
> - '<token>' in: <causal sentence prefix>
> <up to four examples>
> Classify two independent properties.
> concept_type must be one of semantic, syntactic, mixed, unclear. Semantic means content-bearing meaning, topic, entities, or world knowledge. Syntactic means grammatical function, word form, punctuation, formatting, subword structure, or sentence position. Use mixed when both are material.
> orientation must be one of input, output, both, unclear. Input means the current token and causal context best explain the cluster. Output means the predicted next-token evidence best explains it. Use both only when both evidence sources independently support the interpretation. The predicted tokens come from the final language-model head and are auxiliary evidence, not a direct readout from this intermediate layer.
> Return only JSON in this form:
> {"concept_type":"<semantic—syntactic—mixed—unclear>","orientation":"<input—output—both—unclear>","reason":"brief evidence"}

A.3.3 BILINGUAL SHARED-CONCEPT TYPE (R4)

This prompt is applied to concepts selected as translation-shared by the activation-frequency criterion in Result 4, independently of statistical significance. It uses up to 12 training tokens per language with counts and four representative parallel examples. Each sentence is truncated to 320 characters; these bilingual examples are sentence-level evidence, rather than the causal prefixes used above. Layer identity and shared-pair counts and rates are withheld. This task returns a type, a short English description, and a reason.

> You are classifying a concept discovered jointly from two languages in a language model's residual-stream activations.
>
> Languages: English and Mandarin Chinese
>
> English training tokens: <token-count list>
>
> Mandarin Chinese training tokens: <token-count list>
>
> Representative parallel examples:
>
> - English: <English sentence>
>
> Mandarin Chinese: <Mandarin sentence>
>
> activated tokens: English=[<tokens>], Mandarin Chinese=[<tokens>]
>
> <up to four parallel examples>
>
> Classify the organizing principle of this concept:
>
> - semantic: shared meaning, topic, entity type, lexical meaning, or a translation-equivalent semantic category;
> - syntactic: grammar, morphology, function words, punctuation, orthography, token fragments, or sentence position;
> - mixed: both semantic and syntactic evidence are clearly central;
> - unclear: the evidence is insufficient or incoherent.
>
> Judge the activation evidence, not merely whether the two sentences are translations. Return exactly one JSON object with no markdown:
>
> {"classification":"semantic—syntactic—mixed—unclear","label":"2-6 word English description","reason":"one short English sentence"}

A.4 SUPPLEMENTARY RESULTS FOR RESULT 1: SANITY CHECK

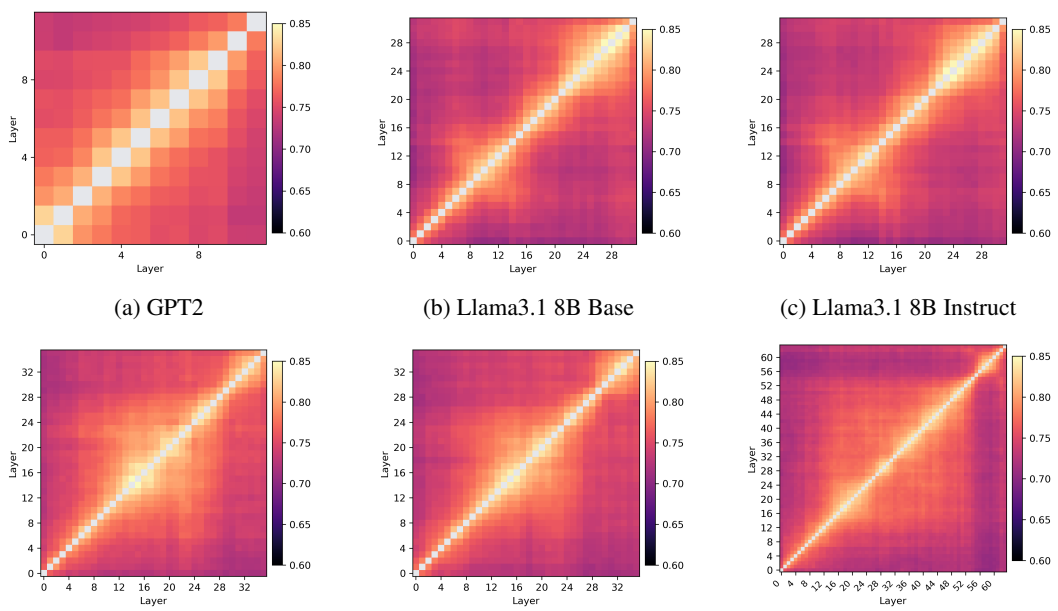
Tab. 4 shows the result of the sanity check stratified by model.

	PCA (512-d)		CKA (linear)		CKA (RBF)		CBA (NLMCD)	
Model	NN ↑	z ↑	NN ↑	z ↑	NN ↑	z ↑	NN ↑	z ↑
GPT-2	1.000	1.407	0.917	0.461	1.000	1.288	1.000	1.893
Llama3.1 8B Instruct	0.938	1.827	1.000	1.702	1.000	1.694	1.000	2.546
Llama3.1 8B Base	1.000	1.896	1.000	1.683	1.000	1.680	1.000	2.569
Qwen3 4B	0.944	1.749	0.944	0.911	1.000	1.787	0.972	2.481
Qwen3 8B	1.000	1.744	0.972	0.931	1.000	1.796	1.000	2.621
Qwen3 32B	1.000	2.395	0.969	0.744	1.000	2.038	1.000	2.945
Tulu SFT	1.000	1.898	1.000	1.715	1.000	1.714	1.000	2.551
Tulu DPO	1.000	1.936	1.000	1.717	1.000	1.718	1.000	2.524
Tulu final	0.969	1.953	1.000	1.716	1.000	1.716	1.000	2.504

Table 4: Per-model nearest-neighbor alignment score and mean best-neighbor z-score (z) for all nine checkpoints underlying Tab. 1. NN measures whether the highest-scoring other layer is adjacent. z measures how strongly the best immediate neighbor stands above the row mean relative to its standard deviation.

A.5 SUPPLEMENTARY RESULTS FOR RESULT 2: LAYER-WISE ALIGNMENT

In Fig. 9, we show CBA layer-alignment matrices for six models under consideration. This supplements Fig. 4 in the main text.



(a) GPT2 (b) Llama3.1 8B Base (c) Llama3.1 8B Instruct

(d) Qwen3 4B (e) Qwen3 8B (f) Qwen3 32B

Figure 9: CBA layer alignment. Shared color scale $[0.60, 0.85]$; self-alignment diagonals masked. Layer indices are zero-based. All panels use 5,000 samples.

A.6 Supplementary Results for Result 3: Descriptive/Qualitative Insights

In Tab. 5, we show the summary statistics (summarized across all layers) of the descriptive concept analysis for Llama-3.1-8B-Instruct and Qwen3-8B. In Fig. 10, we visualize the layer dependence of the concept properties for Llama-3.1-8B-Instruct, complementing the same analysis for Qwen3-8B in the main text.

A.7 Supplementary Results for Result 4: Multilingual Activation

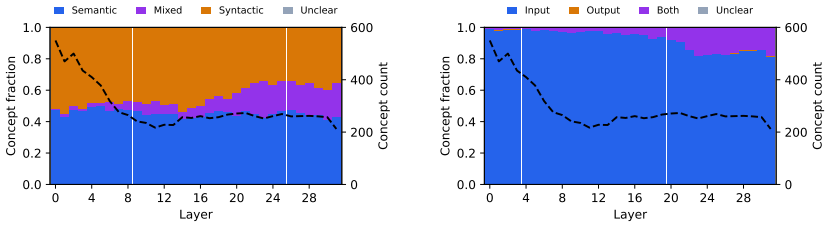
Fig. 11 summarizes shared concept mass as a function of the relative position of a layer in the architecture. In particular, it shows that in certain cases (such as Qwen3-32B close to the top layer), the token mass and the shuffled token mass largely coincide, indicating concepts that are always activated. This motivates the use of the difference between the original and the shuffled token mass as a measure for the multilingual activation.

A.8 Supplementary Results for Result 6: Finetuning Case Study

In Fig. 12, we show CBA alignment matrices between different finetuned models. All of them show strong alignment of corresponding layers (diagonal). In Fig. 13, we show a pairwise-distance concept cost matrix, which represents the basis for the application of the Hungarian algorithm for concept matching, as well as identified concept matches. In Tab.6, we show representative concept matches between SFT and DPO stages for two layers (14 and 15). Tab.7 shows analogous matching results between the final Tulu model and the respective base model.

Model	Concepts	Semantic	Syntactic	Mixed	Input	Both
Llama-3.1-8B-Instruct	9,398	45.7%	45.1%	9.1%	93.4%	6.6%
Qwen3-8B	11,496	39.8%	53.2%	7.0%	96.3%	3.7%

Table 5: English concept annotations aggregated over residual layers. Percentages count concepts. Both denotes joint input/output evidence.



(a) Llama-3.1-8B-Instruct: semantic/syntactic

(b) Llama-3.1-8B-Instruct: input/output

Figure 10: English concept categories and evidence orientation over residual depth. Bars show concept fractions (left axis). The dashed black line shows the total number of annotated concepts at each layer (right axis).

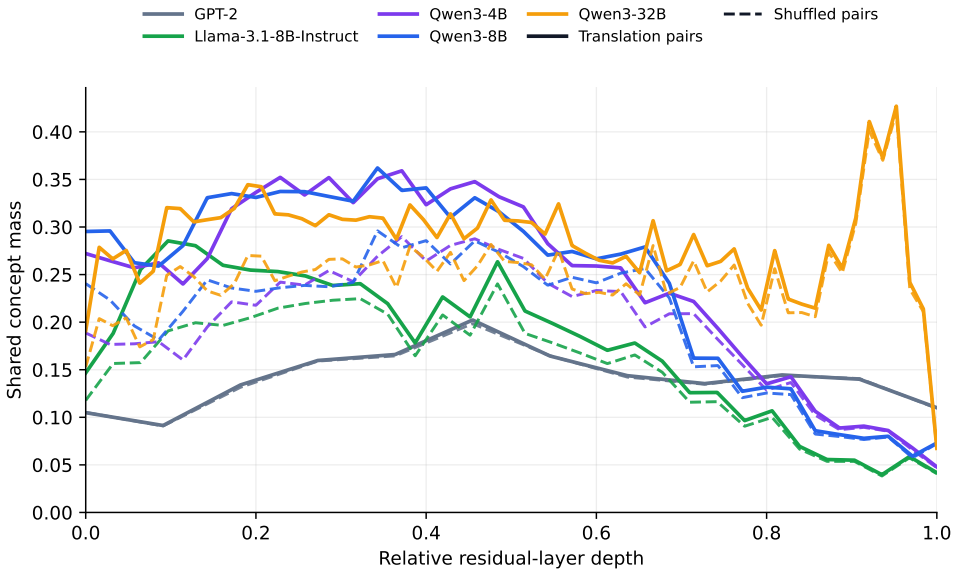


Figure 11: Shared concept mass over normalized residual depth for English–Mandarin translation pairs and shuffled-pair controls.

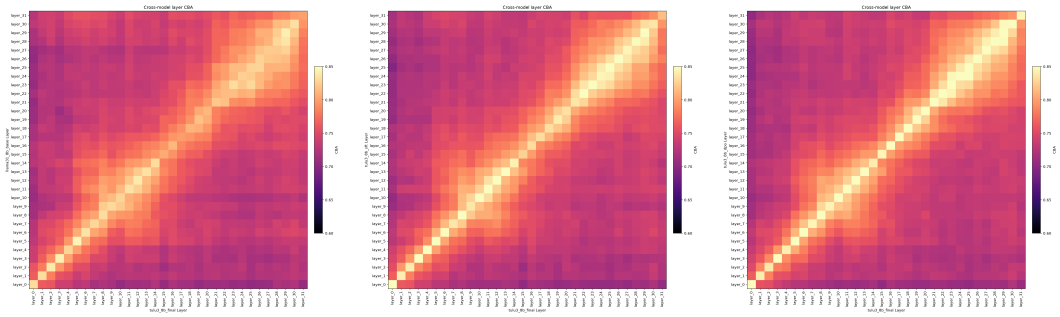


Figure 12: CBA Similarity Scores between finetune models. Left: Base vs. RLVR, Center: SFT vs. RLVR, Right: DPO vs. RLVR

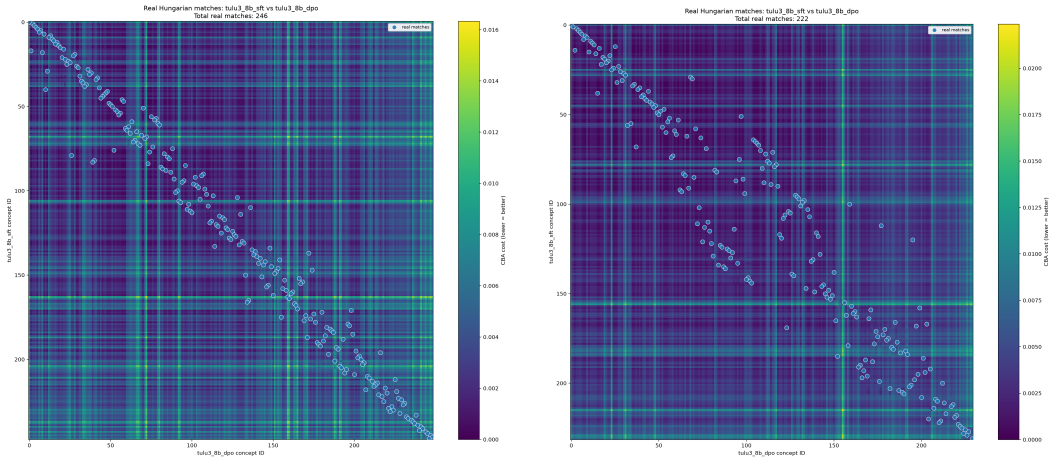


Figure 13: Pairwise-distance concept cost matrix and concept matches for SFT vs DPO. Left: Layer 14 Right: Layer 15

Table 6: Representative Hungarian matches between SFT and DPO at Layers 14 and 15. Lower cost indicates greater similarity. "Unmatched" concepts are assigned to private dummy nodes. For all experiments the threshold for assigning matches to dummy nodes was set to 0.0025.

Layer	Type	SFT concept	DPO concept	Cost
14	Top	section delimiters	section delimiters	1.0×10^{-4}
	Top	Prepositions and conjunctions	Prepositions and conjunctions	1.1×10^{-4}
	Top	Month names	Month names	1.3×10^{-4}
	Top	sentence connectors	sentence starters	1.4×10^{-4}
	Top	sentence starter	demonstrative pronouns	1.4×10^{-4}
	Imperfect	definite article	definite article	0.0048
	Imperfect	prepositions and verbs	Prepositional phrase start	0.0047
	Imperfect	physical structures	places and structures	0.0045
	Imperfect	Last names	Surnames and Names	0.0044
	Imperfect	sentence initial the	sentence initial capitalized the	0.0043
	Unmatched	definite article, degree modifiers	Last names, military terminology, First names	—
15	Top	Prepositions and conjunctions	Prepositions and conjunctions	0.6×10^{-4}
	Top	section delimiters	section delimiters	0.7×10^{-4}
	Top	Name beginnings	Equality symbols	0.7×10^{-4}
	Top	email symbols	email symbols	0.8×10^{-4}
	Top	punctuation markers	number separators	0.8×10^{-4}
	Imperfect	past tense verbs	past tense verbs	0.0049
	Imperfect	Government Bodies	Building materials	0.0045
	Imperfect	sentence start capitalized	definite article	0.0042
	Imperfect	common English prefixes	consonant clusters	0.0042
	Imperfect	punctuation markers	sentence boundary markers	0.0041
	Unmatched	Indefinite articles, Important Descriptors, conjunctions, auxiliary verbs, Prepositional start	Common English words, subject pronouns, auxiliary verbs, auxiliary verbs and prepositions, Building and structure terms	—

Table 7: Representative Hungarian matches between Llama-3.1-8B-Base and Tulu-3-8B-Final at early, middle, and deep layers. Lower cost indicates greater similarity. "Unmatched" concepts are assigned to dummy nodes. For all experiments the threshold for assigning matches to dummy nodes was set to 0.0025.

Layer	Type	Llama concept	Tulu concept	Cost
0	Top	Preposition *in*	Preposition *in*	6.0×10^{-5}
	Top	Equality sign	Punctuation and digits	7.0×10^{-5}
	Top	Temporal conjunctions	Temporal conjunctions	7.0×10^{-5}
	Imperfect	Punctuation comma	Sentence boundary	0.0044
	Imperfect	Definite article	Sentence start	0.0041
	Imperfect	Past-tense verb	Past-tense verb	0.0040
	Unmatched	Article introducer; conjunction after; sentence boundary	Verb *to*; punctuation comma	—
16	Top	Equality symbols	Equality symbols	7.0×10^{-5}
	Top	Punctuation markers	Month names	1.3×10^{-4}
	Top	ISBN and numeric sequences	Email addresses	1.4×10^{-4}
	Imperfect	Sentence-ending punctuation	Sentence-ending punctuation	0.0047
	Imperfect	Past and present verb forms	Auxiliary verbs	0.0045
	Imperfect	Past-tense action verbs	Past-tense action verbs	0.0045
	Unmatched	Wikipedia article opening; sentence-boundary markers	Punctuation markers; common English prefixes; surname beginnings	—
31	Top	Lowercase prepositions	Capitalization markers	9.0×10^{-5}
	Top	Capitalized names	Capitalized words	1.8×10^{-4}
	Top	Cricket terminology	Capitalized words	1.9×10^{-4}
	Imperfect	Ordinal and temporal indicators	Common English words	0.0048
	Imperfect	Common English words	Common English words	0.0048
	Imperfect	Common English words	Function words and metals	0.0046
	Unmatched	Year numbers; verb suffixes; verb forms and related words	Year numbers; verb suffixes; initial consonants	—