

# From Weights to Representations: How Pre-Training Data Order Shapes Language Models

Luca Dini<sup>1,\*</sup>, Lucia Domenichelli<sup>1,2</sup>, Dominique Brunato<sup>1</sup> and Felice Dell’Orletta<sup>1</sup>

<sup>1</sup>*Institute of Computational Linguistics “Antonio Zampolli” (CNR-ILC), ItaliaNLP Lab, Via Giuseppe Moruzzi, 1, Pisa, 56124, Italy*

<sup>2</sup>*University of Pisa, Largo Bruno Pontecorvo 3, Pisa, 56127, Italy*

## Abstract

Neural Language Models can exhibit similar behavior while converging to different internal solutions, in terms of both learned weights and induced sentence representations. In this work, we ask whether the order of pre-training examples contributes to this divergence. Prior studies have shown that pre-training data ordering affects both learning dynamics and the properties encoded in the resulting representations, but its role in shaping the similarity between models trained on different orders remains less understood. We study this question in a controlled setting by comparing GPT-style models pre-trained with identical architecture, initialization, and corpus, varying only the sequence in which examples are presented. We measure how training order shapes model similarity across weight and representation spaces, comparing layer-wise weight matrices and the corresponding component-level activations. We also investigate whether distances in parameter space are reflected in representational distances, and examine the linguistic profile of the sentences that induce the largest representational differences. Our findings characterize training order as a source of internal model divergence, linking data sequence to parameter configurations, representational structure, and input-level linguistic properties. The code for the experimental setup, analysis pipeline, and data processing is available at: [https://github.com/lucadinidue/all\\_the\\_models\\_distances](https://github.com/lucadinidue/all_the_models_distances).

## Keywords

Neural Language Models, Model similarity, Representation similarity, Centered Kernel Alignment, Data Ordering, Curriculum Learning

## 1. Introduction

The training dynamics of neural language models are shaped not only by architectural choices and data composition, but also by the *order* in which training examples are presented. Prior work has shown that pre-training data ordering can significantly influence convergence and generalization, affecting downstream performance, learning efficiency, and the geometry of learned representations [1, 2, 3].

However, these analyses focus on the observable properties of representations, such as linguistic quality, predictive behavior, and geometric organization, rather than on the internal structures from which these properties emerge. It remains unclear whether different data orderings lead models to organize their internal knowledge in similar ways: models may acquire representations with comparable properties while relying on fundamentally different parameter configurations, or encode similar information through only weakly aligned activation spaces.

This issue is particularly relevant in light of recent work showing that neural networks with comparable performance can nonetheless develop different internal solutions depending on their optimization trajectory [4, 5, 6]. Pre-training data order can thus be viewed not merely as a factor affecting optimization efficiency, but as a potential source of *path dependence* in the emergence and organization of linguistic representations.

In this work we address the following research question: “**To what extent are model weights and learned representations sensitive to the order of pre-training data?**” To this end, we analyze GPT-style language models trained under controlled conditions (identical architecture, initialization,

---

CLiC-it 2026: Twelfth Italian Conference on Computational Linguistics, September 14–16, 2026, Palermo, Italy

\*Corresponding author.

✉ [luca.dini@ilc.cnr.it](mailto:luca.dini@ilc.cnr.it) (L. Dini); [lucia.domenichelli@ilc.cnr.it](mailto:lucia.domenichelli@ilc.cnr.it) (L. Domenichelli); [dominique.brunato@ilc.cnr.it](mailto:dominique.brunato@ilc.cnr.it) (D. Brunato); [felice.dellorletta@ilc.cnr.it](mailto:felice.dellorletta@ilc.cnr.it) (F. Dell’Orletta)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and corpus) where data order is the only source of variation, comparing both curriculum-based and random orderings.

Our contributions are as follows:

- We investigate the impact of pre-training data order on model weights, comparing models that differ only in the order in which training examples are presented.
- We extend the same analysis to sentence representations, examining whether patterns of weight similarity are reflected in patterns of representational similarity.
- We conduct both analyses at a fine-grained level, considering each layer separately and each architectural module independently, in order to assess which components are most sensitive to data order.
- We provide a linguistically grounded interpretation of representational divergence by identifying the properties of sentences that induce the highest and lowest cross-model similarity.

Together, these analyses offer a unified view of how training order shapes the internal organization of language models, linking parameter configurations, representation spaces, and input-level linguistic properties.

## 2. Related Work

Understanding how pre-training data order shapes the internal structure of language models requires drawing on two distinct lines of research: work on curriculum learning and data ordering, which examines how training sequences affect optimization and model behavior, and work on model similarity, which provides tools for comparing models in weight and representation space.

### 2.1. Data Ordering and Curriculum Learning

The order in which training examples are presented can substantially affect the optimization dynamics and final behavior of neural networks. Even when trained on the same data, models may converge to different solutions depending on the sequence in which examples are processed, with consequences for convergence speed, generalization, and the internal features that emerge during training [7, 8]. This sensitivity suggests that data ordering shapes the learning trajectory, influencing not only how fast models learn, but also *what* they learn.

In natural language processing, curriculum learning [9] provides a prominent framework for explicitly controlling data order, typically by organizing training examples according to a notion of difficulty. A large body of work has explored curriculum strategies for neural language models, often relying on human-inspired, model-agnostic complexity measures such as sentence length, readability, or lexical features [10, 11, 12, 3]. In parallel, other approaches define difficulty through model-dependent signals, such as loss or confidence, leading to adaptive curricula that evolve with the model’s competence [13, 14, 15, 16].

These studies show that data ordering can influence training efficiency, downstream behavior, and the properties of learned representations, but typically assess its effects through performance metrics, probing tasks, or geometric analyses. Whether different orderings lead models to occupy similar or distinct regions of weight and representation space remains an open question.

### 2.2. Model Similarity in Parameter and Representation Space

A complementary line of research investigates how to compare neural networks beyond aggregate performance metrics, by analyzing their internal structure directly. One approach operates in parameter space, comparing the weights learned by different models. Weight-level comparison is less straightforward than it might appear: neural networks exhibit well-known symmetries, such as permutation-equivalent neuron configurations, that cause models implementing the same function to occupy arbitrarily different

positions in weight space [17]. Rather than treating raw weight distances as a proxy for functional divergence, structured methods such as biCKA [18] bypass this problem by comparing Transformer weight matrices through their induced similarity structure, enabling fine-grained, layer-wise and module-wise analyses across models.

A second perspective focuses on the representations induced by different models over a shared set of inputs. A range of methods has been proposed to measure alignment between activation spaces, including Canonical Correlation Analysis (CCA) [19], Singular Vector CCA (SVCCA) [20], Projection-Weighted CCA (PWCCA) [21], and Centered Kernel Alignment (CKA) [22], each making different assumptions about the geometric structure of representation spaces. These methods have been applied to study layer-wise correspondence, learning dynamics, and representational stability across models and training stages [20, 22, 23]. More recent work has moved beyond pairwise layer comparisons, revealing implicit groupings of functionally related layers [24].

Despite these advances, the role of data ordering as a source of both parametric and representational variation remains comparatively underexplored.

### 3. Approach

Our goal is to assess whether the order of pre-training data affects the internal structure of neural language models. We adopt a controlled comparison framework in which models share the same architecture, training corpus, and optimization setup, and differ only in the sequence in which training examples are presented. This allows us to treat data ordering as the only source of variation among the trained models.

We analyze model differences along two complementary dimensions. First, we compare the corresponding weight matrices of the models, asking whether different training sequences lead to different parameter configurations. Second, we compare the representations induced by different models on a shared evaluation set of sentences, asking whether models trained with different orderings encode the same inputs in similar ways. Both analyses are performed in a fine-grained manner, across layers and Transformer components, rather than only at the level of whole-model parameters or final hidden states. In particular, we distinguish between components involved in the self-attention mechanism and in the feed-forward sublayer. This allows us to assess whether the effect of data ordering is distributed uniformly across the model, or whether it is concentrated in specific functional parts of the Transformer, such as the projections used to compute attention patterns, the transformations applied to token information, or the representations passed between layers.

Beyond measuring weight and representation similarity separately, we also study the relationship between the two, asking whether models that are closer in parameter space also tend to produce more similar representations, or whether comparable representations can emerge from substantially different weight configurations.

Finally, we connect model-level differences back to the input data. Aggregate similarity scores may not show the fact that ordering effects are unevenly distributed across inputs. We therefore identify the sentences for which representations vary most across training orders and analyze their linguistic properties, asking which input characteristics are associated with greater sensitivity to data ordering.

## 4. Experimental Setting

### 4.1. Models and Pre-Training Data Orderings

We analyze a set of GPT-2-style language models trained by Dini et al. [3]. All models share the same architecture: a decoder-only Transformer with 8 layers, 8 attention heads, and a hidden size of 512, for a total of approximately 40M parameters.

All models were trained on the same corpus of 10M sentences sampled from Italian Wikipedia, using identical training configurations. Each data ordering is used to train three models initialized with

the same set of random seeds. Consequently, comparisons across orderings can isolate the effect of training example order for a fixed initialization, while the use of multiple seeds enables us to evaluate the consistency of the observed effects across different initializations.

The training orderings fall into three main groups. The first group consists of *complexity-based curricula*, where sentences are presented in increasing order according to different complexity proxies: sentence length, the Gulpease index [25], and Read-IT readability score [26]. Sentence length represents a relatively shallow indicator of complexity, whereas Gulpease and Read-IT are more directly related to readability assessment. In particular, Gulpease relies mainly on raw textual features, while Read-IT is computed from a broader set of lexical, morphosyntactic, and syntactic features. The second group consists of the corresponding *anti-curricula*, obtained by reversing each complexity-based ordering and therefore presenting sentences in decreasing order of complexity.

The third group includes *non-curriculum orderings*. These comprise models trained on randomly shuffled data using four different shuffling seeds, together with the corresponding reversed orderings, as well as models trained on the original Wikipedia dump order and its reverse. This original ordering may retain residual document-level or temporal structure, such as local coherence introduced by the original dump organization or by the sampling process, thus providing a weakly structured ordering, distinct from both explicit curricula and fully random orderings.

Overall, we analyze 48 trained models, corresponding to 16 data orderings and 3 initialization seeds. All reported results are computed by first comparing models sharing the same initialization seed and then averaging the resulting measurements across the three seeds.

## 4.2. Evaluation Data

To compute representation similarity, we use a held-out set of 10,000 sentences sampled from Italian Wikipedia and excluded from the models’ pre-training corpus. This provides an in-domain but unseen evaluation set. Following the preprocessing criteria used during pre-training, we retain only sentences containing at least six tokens.

The sample size follows empirical practice in neural representation similarity analysis. Prior work has shown that CKA-based comparisons tend to stabilize with approximately 10,000 input examples, making this scale a practical reference for estimating representation-level similarity [27, 28].

### 4.2.1. Linguistic Annotation

To investigate whether training order sensitivity is associated with specific linguistic phenomena, we annotate the evaluation set with Profiling-UD [29], a linguistic profiling tool based on the Universal Dependencies annotation framework [30] that extracts a broad range of surface, morpho-syntactic, and syntactic features from parsed sentences.

We use these annotations to characterize the linguistic properties of the sentences for which representations diverge most across training orderings, thereby identifying which linguistic phenomena are more sensitive to differences in training trajectory.

## 4.3. Weight and Representation Similarity

We analyze model similarity at two complementary levels: the weights learned by each model and the representations produced during inference. We describe the representation-level analysis first, since it relies on CKA, the metric from which our weight-level measure (biCKA) is derived.

At the representation level, for each input sentence, we run a forward pass and extract the activations produced by a selected set of the models components: the query ( $Q$ ), key ( $K$ ), value ( $V$ ), and attention output ( $OA$ ) projections, together with the feed-forward intermediate ( $DI$ ) and output ( $DO$ ) projections. We also include the standard hidden states, which correspond to the layer output. In GPT-style models, queries, keys, and values are produced by a single joint attention projection; we split this output into separate  $Q$ ,  $K$ , and  $V$  representations before comparison. Attention heads are kept concatenated, so that each representation captures the full output of a given module at a given layer.

Since activations are computed at the token level, we aggregate them into sentence-level vectors using last-token pooling. This yields, for each combination of model, layer, and component, a representation matrix whose rows correspond to sentences and whose columns correspond to activation dimensions. We compare these matrices across models using linear CKA, one of the most widely adopted metrics for representation comparison [22]. Linear CKA quantifies how similarly two representation matrices organize the same set of sentences: rather than comparing activation values directly, it compares the pattern of pairwise similarities that each representation induces over the sentences, where the similarity between two sentences is given by the inner product of their activation vectors. Two representations are therefore judged similar when they arrange the sentences in a comparable configuration relative to one another, irrespective of the coordinate system in which the activations are expressed. As a result, the comparison is invariant to orthogonal transformations and isotropic rescaling while remaining sensitive to differences that alter the relative geometry of the sentences. These invariances are desirable when comparing distinct models, whose activation spaces need not share a common basis or scale.

We then perform an analogous analysis at the weight level. Here we compare the parameter matrices of the same components considered above,  $Q$ ,  $K$ ,  $V$ ,  $OA$ ,  $DI$ , and  $DO$ , excluding the hidden states, which have no corresponding parameter matrix. We measure their similarity using biCKA, a variant of linear CKA designed to compare weight matrices [18]. biCKA represents each matrix through a block-diagonal structure that preserves information from both its row and column spaces, allowing similarity to be assessed along both dimensions of the weight matrix. This makes it more appropriate for comparing parameter configurations than applying CKA directly to the raw matrices. Because biCKA is derived from CKA, the two levels of analysis share a common, similarity-based framework.

Both analyses are performed separately for each Transformer layer and, within each layer, for each architectural component considered. Since both linear CKA and biCKA scores range from 0 to 1, higher values indicate greater similarity.

#### 4.3.1. Sentence-level Similarity Analysis

Although CKA provides a global measure of similarity between two representation spaces, it does not assign scores to individual sentences. To obtain a sentence-level estimate, we decompose the CKA score into per-sentence contributions, where each sentence is assigned a value reflecting how consistently its relational profile (i.e. its pattern of similarity to all other sentences) is preserved across the two compared models. Sentences with lower contribution scores are those whose representations are least stable across models, and therefore most influenced by data ordering. We use these scores to identify the sentences most affected by pre-training data order. For a technical description of the decomposition, see Appendix A.

## 5. Results

### 5.1. Weight Similarity

Figure 1 reports the average biCKA score across layers for each weight module, computed over all non-identical model pairs and averaged across initialization seeds. A more detailed breakdown by type of data ordering (random, forward-curriculum, and backward-curriculum) is provided in Appendix D.1. The layer-wise trends are generally smooth, suggesting that the effect of data ordering on the learned parameters is not tied to isolated layers, but rather follows module-specific patterns that persist across depth. Among the considered modules,  $DI$  shows the highest and most stable similarity, while  $Q$  and  $K$  are consistently among the least aligned components.

The only modules that don’t show a stable pattern are the  $Q$  and  $K$  projections. These modules start from relatively high similarity in the first layer, but drop sharply at layer 2 and remain lower in the following layers. This suggests that the parameters involved in query-key matching are especially sensitive to differences in the order in which training examples are presented. By contrast, the feed-forward intermediate projection appears more robust, indicating that some parts of the network converge

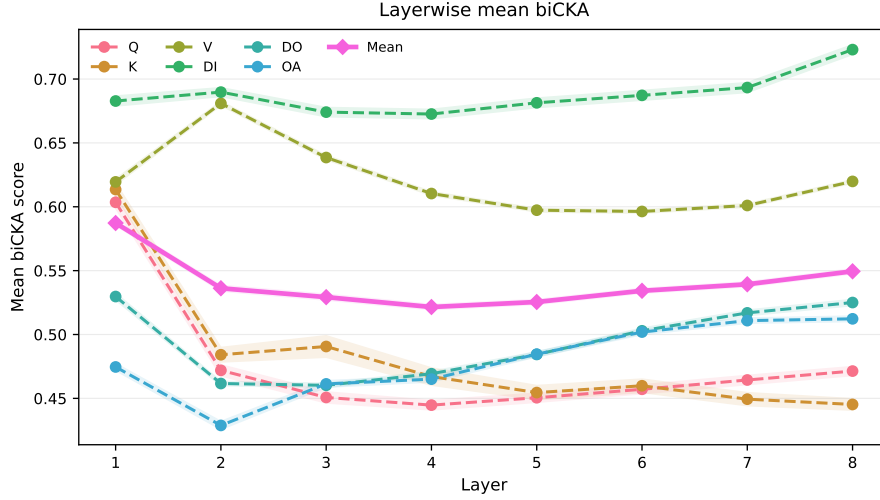


**Figure 1:** Layer-wise average biCKA score. Each line in the plot corresponds to a different model’s module. In addition the "Mean" line reports the average score computed across all the modules. Error bars correspond to the standard deviation across the initialization seeds.

to more similar parameter configurations despite changes in data ordering.

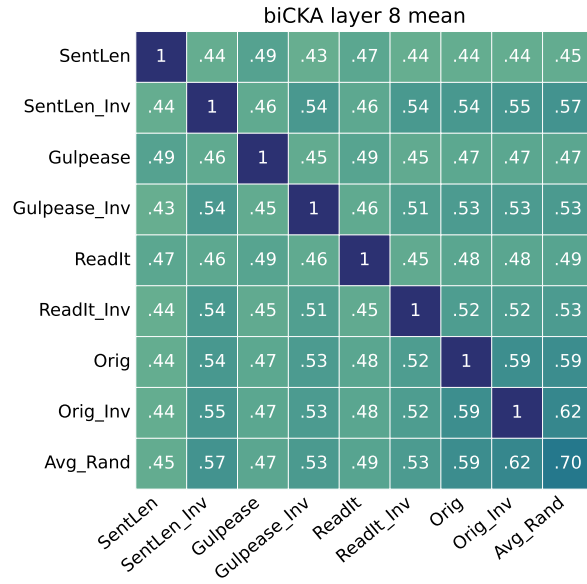


**Figure 2:** Average biCKA scores of model pairs across all weight modules in the last layer. *Avg\_Rand* denotes the average over models trained on randomly shuffled data.

Figure 2 reports final-layer biCKA similarities between model pairs, averaged across all weight modules. Since the observed patterns are largely consistent across modules, this average provides a compact summary of the global parameter-space structure at the last layer while avoiding redundant module-wise visualizations. For completeness, module-specific plots are provided in Appendix B.

For models trained under non-curriculum orderings, those based on different randomly shuffled versions of the training data are aggregated as *Avg\_Rand*, while those based on the original Wikipedia dump order *Orig* and its reverse *Orig\_Inv* are kept separate because they show lower similarity scores with respect to the fully randomized baselines. Even within this group of non-curriculum models, similarities are not close to identity, confirming that independently trained models do not converge to the same parameter configuration, even when trained under comparable ordering conditions. Nevertheless, random-order models remain the most internally aligned group, providing a reference point against

which the curriculum models can be compared.

The heatmap reveals a clear effect of curriculum direction. Models trained with inverse curricula (i.e. from more complex to easier sentences) remain relatively close to the random baseline and to each other, whereas forward curricula (i.e. from easier to more complex sentences) lead to a distinct region of parameter space and show lower similarity to both random and inverse models. This indicates that curriculum direction leaves a measurable trace in the learned weights, with forward orderings producing more pronounced parameter-level deviations.

The picture that emerges is therefore not one of a uniform shift in parameter space. Differences are concentrated mainly in the attention projections, and in particular in  $Q$  and  $K$ , suggesting that the mechanisms used to score token interactions are among the most sensitive to training order. This effect is strongest for forward curricula. A plausible explanation is that starting from simpler or more regular examples constrains the early optimization trajectory, leading models toward parameter configurations that are only partly reshaped by later training. Inverse curricula, instead, expose the model to more heterogeneous examples from the beginning, which may make their updates less systematically different from those produced by random orderings.  $DI$ , which corresponds to the intermediate representation passed from the attention block to the feed-forward network, appears substantially more stable. This suggests that, unlike the attention projections that directly parameterize token-to-token interactions, this component encodes a more aggregated representation that is less sensitive to training order.

## 5.2. Representation Similarity

Figure 3 reports the corresponding average CKA similarity across layers for each representation module computed over all non-identical model pairs and averaged across initialization seeds. Again, a more detailed breakdown by type of data ordering is provided in Appendix D.2. Compared to the weight-level analysis, representation similarities are generally higher, but they show clearer changes across depth. For  $V$ ,  $DI$ ,  $DO$ , and hidden states, similarity is highest in the first layer, often above 0.9, and then decreases smoothly. This suggests that lower layers encode more shared and stable representations, while deeper layers are more affected by the specific training trajectory induced by data ordering.

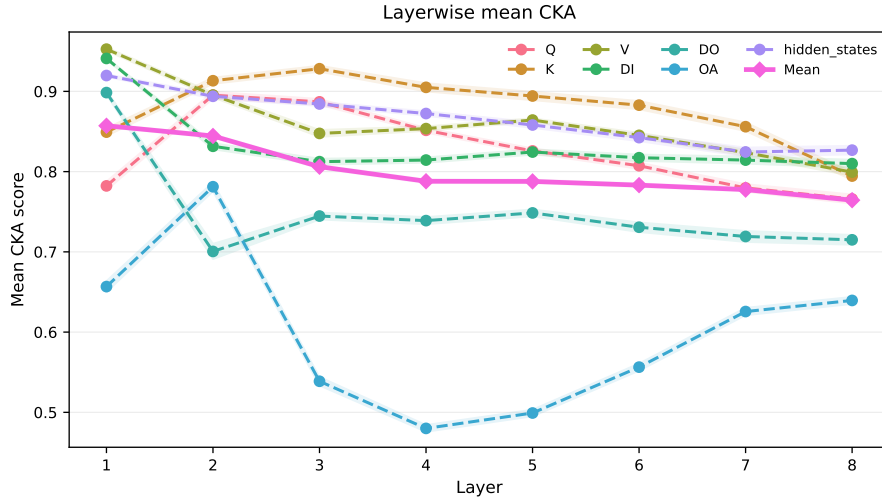


**Figure 3:** Layer-wise average CKA score. Each line in the plot corresponds to representations extracted from a different model’s module. In addition the "Mean" line reports the average score computed across all the modules. Error bars correspond to the standard deviation across the initialization seeds.

$Q$  and  $K$  follow a slightly different trend: their highest similarity is reached around layers 2–3, before gradually decreasing in the following layers. The attention output projection ( $OA$ ) is the most distinct component, with lower similarity overall, a marked drop in the middle layers, and a partial recovery toward the end of the network. Since  $OA$  projects the concatenated attention-head outputs back into the model space, its lower similarity may reflect differences not only in the attended information, but

CKA layer 8 mean

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .52 | .83 | .48 | .82 | .76 | .80 | .81 | .81 |
| SentLen_Inv  | .52 | 1   | .52 | .57 | .54 | .60 | .57 | .57 | .58 |
| Gulpease     | .83 | .52 | 1   | .46 | .82 | .76 | .82 | .82 | .82 |
| Gulpease_Inv | .48 | .57 | .46 | 1   | .48 | .55 | .49 | .50 | .50 |
| ReadIt       | .82 | .54 | .82 | .48 | 1   | .76 | .82 | .82 | .83 |
| ReadIt_Inv   | .76 | .60 | .76 | .55 | .76 | 1   | .79 | .80 | .80 |
| Orig         | .80 | .57 | .82 | .49 | .82 | .79 | 1   | .86 | .85 |
| Orig_Inv     | .81 | .57 | .82 | .50 | .82 | .80 | .86 | 1   | .86 |
| Avg_Rand     | .81 | .58 | .82 | .50 | .83 | .80 | .85 | .86 | .89 |

SentLen SentLen\_Inv Gulpease Gulpease\_Inv ReadIt ReadIt\_Inv Orig Orig\_Inv Avg\_Rand

**Figure 4:** Average CKA scores of model pairs across all weight modules in the last layer. *Avg\_Rand* denotes the average over models trained on randomly shuffled data.

also in how attention heads specialize and how their outputs are recombined into the residual stream. The particularly low scores in the middle layers therefore suggest that models trained with different data orderings may differ most in how they transform and reintegrate attention information during intermediate processing.

Figure 4 reports final-layer CKA similarities between model pairs, averaged across all representation modules. As in the weight-level analysis, we report the module average because model-pair patterns are consistent across modules, allowing us to summarize final-layer representational similarity while avoiding redundant module-wise heatmaps, which are provided in Appendix C.

Again, random-order models are aggregated into *Avg\_Rand*, while *Orig* and *Orig\_Inv* are kept separate. Overall, random-order models provide a highly aligned representation-level baseline, confirming that independently trained models can produce similar activation spaces even when their parameter configurations are not identical. This contrasts with the weight-level results, where similarities were lower and curriculum effects were more directly visible in parameter space.

The heatmap shows that curriculum direction also affects representations, but with a different pattern from the one observed for weights. Forward curricula remain broadly close to the random baseline, suggesting that their stronger parameter-level deviations do not translate into equally large representational differences at the final layer. Inverse curricula, instead, produce more marked representational shifts: *ReadIt\_Inv* remains relatively aligned with the other models, whereas *SentLen\_Inv* and especially *Gulpease\_Inv* show lower similarities. This indicates that reversed complexity-based orderings can lead to more distinct final-layer representations.

Taken together, the CKA results suggest that representations are more robust to data ordering than weights, but not insensitive to it. The main source of variation appears to be the direction of the curriculum, with inverse orderings producing the clearest representational divergence. This effect may be driven by a recency effect in training: inverse curricula expose the model, in the final stages of training, to shorter and syntactically simpler sentences, which have a stronger influence on the final state of the representations. This is consistent with the pattern observed in the layer-wise CKA analysis (Figure D.6 in Appendix D.2), where the divergence of inverse curriculum models is more pronounced in deeper layers, which are typically more sensitive to higher-level syntactic and semantic structure. This phenomenon is also related to catastrophic forgetting, where later training stages overwrite previously learned representations; here, the final exposure to low-complexity data may partially degrade the representational structure built during earlier, more heterogeneous training. This

interpretation also reconciles the apparent contrast with the weight-level results: forward curricula shift models toward different parameter configurations early in training, an effect that persists in weight space but does not strongly alter the final representations; inverse curricula, by contrast, leave the weight space relatively unaffected but influence the organization of the activation space through their late-stage data distribution.

### 5.3. Relation Between Weights and Representations

To investigate the relationship between parameter-space and representation-space similarities, we examine whether models that are closer in weight space also tend to induce more similar internal representations. Figure 5 reports the correlation between biCKA and CKA scores across model comparisons. Higher values indicate that models that are closer in parameter space also tend to produce more similar representations.

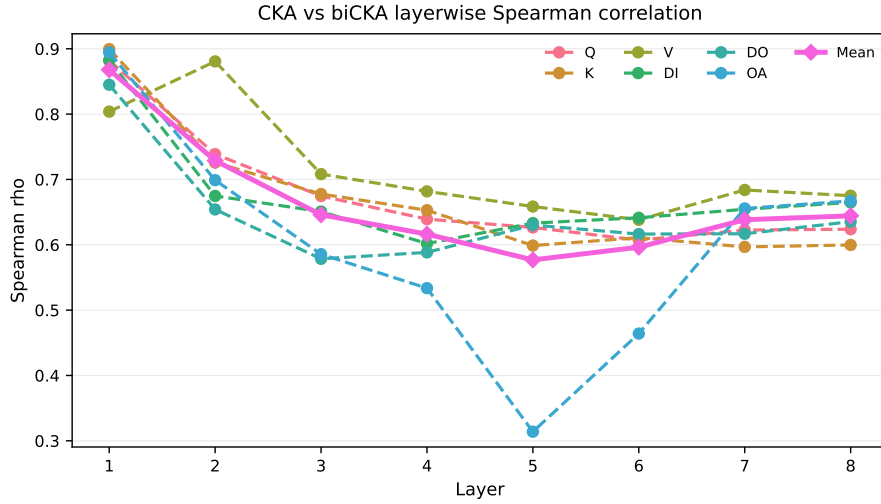


**Figure 5:** Spearman correlation between biCKA and CKA scores computed over pairs of models, measuring the relationship between parameter-space and representation-space similarity.

The strongest association between weights and representations is observed in the first layer, where Spearman’s  $\rho$  values are close to 0.9 for most modules. The only exception is *V*, whose maximum correlation is reached at layer 2. After this initial peak, correlations decrease toward the middle layers and then stabilize around 0.6 in the final layers.

This trend is consistent with the previous analyses. Lower layers show both high weight similarity and highly aligned representations, suggesting a relatively direct correspondence between parameter and activation spaces. In deeper layers, however, the relationship becomes weaker: models can differ more in their weights while still producing partially aligned representations, or conversely reach similar representational behavior through less similar parameter configurations.

The *OA* module again departs from the general pattern. Its correlation drops to a local minimum of around 0.3 at layer 5, before increasing again in the following layers. This reinforces the interpretation that the attention output projection is a particularly unstable point of correspondence between weights and representations. Since *OA* recombines the information produced by attention heads into the residual stream, differences in its parameters may not translate into representation similarity in a simple or linear way, especially in the middle layers.

## 6. Linguistic Analysis of Distant Sentences

The final analysis examines the linguistic profile of the sentences whose representations are most sensitive to pre-training data ordering. For each curriculum, we identify the sentences whose representations

| Top 20 contribution features - layer_08 |          |                 |               |                        |                 |                      |                              |                    |               |                   |               |                  |                           |                 |                      |                           |               |                     |                    |                |
|-----------------------------------------|----------|-----------------|---------------|------------------------|-----------------|----------------------|------------------------------|--------------------|---------------|-------------------|---------------|------------------|---------------------------|-----------------|----------------------|---------------------------|---------------|---------------------|--------------------|----------------|
| SentLen                                 | -0.14    | -0.11           | -0.13         | -0.13                  | 0.33            | -0.07                | -0.11                        | 0.18               | 0.23          | -0.10             | -0.05         | -0.13            | -0.10                     | -0.13           | 0.20                 | 0.11                      | 0.03          | -0.21               | -0.17              | 0.03           |
| SentLen_Inv                             | -0.44    | -0.43           | -0.37         | -0.35                  | -0.00           | -0.33                | -0.30                        | 0.04               | 0.25          | -0.29             | -0.24         | -0.28            | -0.27                     | -0.08           | -0.02                | 0.04                      | 0.21          | -0.10               | -0.22              | 0.18           |
| Gulpease                                | -0.07    | -0.12           | -0.07         | -0.12                  | 0.28            | -0.03                | -0.08                        | 0.18               | 0.16          | -0.09             | -0.00         | -0.09            | -0.05                     | -0.07           | 0.26                 | 0.22                      | 0.03          | -0.19               | -0.18              | 0.07           |
| Gulpease_Inv                            | -0.23    | -0.41           | -0.06         | -0.19                  | 0.01            | -0.05                | -0.00                        | 0.14               | 0.24          | -0.24             | -0.18         | -0.03            | -0.01                     | -0.27           | 0.13                 | 0.14                      | 0.25          | -0.15               | 0.01               | 0.23           |
| ReadIt                                  | -0.30    | -0.02           | -0.24         | -0.16                  | 0.27            | -0.25                | -0.25                        | 0.17               | 0.14          | -0.03             | -0.16         | -0.26            | -0.22                     | -0.17           | 0.22                 | 0.13                      | -0.04         | -0.25               | -0.24              | 0.06           |
| ReadIt_Inv                              | 0.26     | -0.34           | 0.11          | 0.09                   | 0.27            | 0.20                 | 0.12                         | 0.30               | 0.30          | -0.22             | 0.29          | 0.02             | 0.10                      | -0.10           | 0.22                 | 0.26                      | 0.09          | -0.21               | 0.02               | 0.22           |
| Avg_Rand                                | -0.21    | -0.16           | -0.19         | -0.20                  | 0.26            | -0.11                | -0.18                        | 0.20               | 0.22          | -0.13             | -0.07         | -0.24            | -0.13                     | -0.13           | 0.17                 | 0.19                      | 0.04          | -0.22               | -0.24              | 0.14           |
|                                         | n_tokens | upos_dist_PROPN | avg_max_depth | n_prepositional_chains | upos_dist_CCONJ | verbal_head_per_sent | subordinate_proposition_dist | aux_tense_dist_imp | upos_dist_ADV | dep_dist_flatname | max_links_len | subordinate_post | avg_subordinate_chain_len | upos_dist_PUNCT | verbs_tense_dist_imp | verbs_num_pers_dist_Phr+3 | upos_dist_ADJ | aux_tense_dist_Pres | subordinate_dist_1 | upos_dist_NOUN |
| UD feature                              |          |                 |               |                        |                 |                      |                              |                    |               |                   |               |                  |                           |                 |                      |                           |               |                     |                    |                |

**Figure 6:** Linguistic profile of the 5% of sentences exhibiting the highest representational divergence for each training data order. The feature values are reported in terms of z-scores and sorted with decreasing  $\max(|z\text{-score}|)$ . Only the top-20 features with the largest absolute deviations are reported.

are most divergent with respect to random-ordering models. This is done using sentence-level CKA contribution scores (Section 4.3.1), which quantify representational similarity across models. We select the 5% of evaluation sentences with the lowest average scores, since lower CKA values indicate weaker representational alignment and therefore greater divergence across models. Finally, we compute the average z-scores of their linguistic features with respect to the full evaluation set.

Figure 6 reports the 20 linguistic features with the largest absolute average z-score in at least one curriculum condition, together with the profile of the most distant sentences among random-ordering models (*Avg\_Rand*) as a baseline. In this case, *Avg\_Rand* additionally includes also *Orig* and *Orig\_Inv* orderings, which were excluded from this group in the previous analyses. This comparison helps distinguish features associated with sentences that are generally sensitive to random variation from those that become especially divergent under linguistically motivated pre-training curricula.

For additional context, Appendix E reports the five most distant sentences for each curriculum model and for the random-ordering baseline.

Across curricula, the most divergent sentences are consistently short and structurally simple: they exhibit negative z-scores for sentence length (*n\_tokens*), parse-tree depth (*avg\_max\_depth*), prepositional chains (*n\_prepositional\_chains*), and dependency length (*max\_links\_len*). This pattern is also visible in *Avg\_Rand*, indicating that short, simple sentences are generally more sensitive to variation in training trajectory regardless of its specific form. A plausible account is that shorter sentences amplify the representational impact of individual tokens: with fewer items contributing to the sentence-level representation, small differences in how models encode local dependencies or lexical items produce larger overall divergences. The one exception is *ReadIt\_Inv*, whose most distant sentences are longer than average. This result may reflect the composite nature of Read-IT, which combines lexical, morpho-syntactic, and syntactic information, and suggests that different operationalisations of linguistic complexity can induce qualitatively distinct patterns of representational sensitivity.

Across all ordering conditions, proper nouns (*upos\_dist\_PROPN*) are systematically underrepresented among the most divergent sentences, a tendency reinforced by the lower incidence of the *flat* dependency relation (*dep\_dist\_flatname*), which frequently marks multiword proper names in Universal Dependencies. This consistency suggests that named entities function as representational anchors: their presence stabilises cross-model agreement, while their absence leaves representations more exposed to the influence of training-order differences.

Among the inverse curricula, this pattern is accompanied by a complementary signal: common nouns (*upos\_dist\_NOUN*), adjectives (*upos\_dist\_ADJ*), and adverbs (*upos\_dist\_ADV*) are moderately overrepresented among the most distant sentences. The most divergent representations in these

conditions thus tend to arise in sentences that are not only short and lacking in named entities, but also predominantly composed of open-class content words, a sentence type that requires a greater degree of lexical and compositional integration and may therefore be more sensitive to differences in training trajectory.

A secondary signal involves coordinating conjunctions (*upos\_dist\_CCONJ*), which are overrepresented across several orderings including *Avg\_Rand*, but largely absent for *SentLen\_Inv* and *Gulpease\_Inv*. The inconsistency of this pattern suggests that coordination structures are at most a secondary source of representational variability, subordinate to the structural and lexical effects described above.

A final, recurring but less interpretable signal concerns verbal morphology: several curricula and *Avg\_Rand* show a higher frequency of imperfect-tense verbal forms (*aux\_tense\_dist\_Imp*, *verbs\_tense\_dist\_Imp*) among the most distant sentences. The mechanism is less transparent than for the structural or lexical patterns, but the recurrence of this signal across conditions suggests that tense-related encoding may represent an independent, if secondary, source of representational instability.

## 7. Discussion and Conclusion

In this work, we investigated whether pre-training data order affects the internal structure of neural language models. By comparing models trained with the same architecture, data, and optimization setup but different data orderings, we analyzed both learned weights and induced representations. Overall, our results show that models remain broadly similar, but that data order leaves measurable traces in specific layers, modules, and types of input.

At the weight level, training data ordering leaves a measurable trace in parameter space. Models trained under random orderings remain the most internally aligned group, whereas curriculum-trained models occupy distinct regions of parameter space, with forward curricula showing the strongest deviations. Parameter differences are concentrated in the attention projections, especially  $Q$  and  $K$ , suggesting that the mechanisms governing token-to-token interactions are among the components most sensitive to training trajectory, while feed-forward intermediate weights ( $DI$ ) exhibit greater stability.

The representation-level analysis reveals a partially different picture. Similarities are generally higher than their weight-level counterparts, indicating that different parameter configurations can still give rise to relatively similar activation spaces. Similarity tends to decrease with depth, and the attention output projection ( $OA$ ) emerges as the most sensitive component. Crucially, while forward curricula produce the strongest parameter-level deviations, inverse curricula lead to the most distinct final-layer representations, highlighting that parameter and representational divergence do not necessarily follow the same pattern.

The sentence-level analysis further shows that representational divergence is concentrated on short and structurally simple sentences, typically characterized by a lower incidence of proper nouns and named entities. For some inverse curricula, these sentences also exhibit a greater proportion of content-bearing lexical categories, indicating that representational sensitivity is associated not only with sentence length but also with lexical composition.

Taken together, these findings indicate that the effect of pre-training order is selective rather than global. Different orderings do not lead to radically different models, but they leave systematic traces in specific components and representations. In particular, the results highlight a dissociation between parameter-space and representation-space effects: training order can produce substantial differences in weights without an equally strong impact on final representations, and vice versa. Although this study is limited to a single architecture and a controlled set of curricula, it shows that pre-training order is not merely an optimization detail, but a factor that can shape both the internal organization of model parameters and the representations they produce.

## Acknowledgments

This work has been supported by:

- LLMs4EU project, co-funded by the Digital Europe Programme under Grant Agreement No. 101198470.
- Projects “Understanding and Enhancing Preference Alignment in Large Language Models Through Controlled Text Generation” (IsCc8\_ALIGNLLM) and “Ordering Data for Efficient Representation Learning” (IsCd8\_ORDER), funded by CINECA under the ISCRA initiative, for the availability of HPC resources.

## Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

## References

- [1] Y. Tsvetkov, M. Faruqui, W. Ling, B. MacWhinney, C. Dyer, Learning the curriculum with Bayesian optimization for task-specific word representation learning, in: K. Erk, N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 130–139. URL: <https://aclanthology.org/P16-1013/>. doi:10.18653/v1/P16-1013.
- [2] Y. Zhou, B. Yang, D. F. Wong, Y. Wan, L. S. Chao, Uncertainty-aware curriculum learning for neural machine translation, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6934–6944. URL: <https://aclanthology.org/2020.acl-main.620/>. doi:10.18653/v1/2020.acl-main.620.
- [3] L. Dini, L. Domenichelli, D. Brunato, F. Dell’Orletta, On the impact of pretraining data ordering in transformer encoder- and decoder-only language models, Knowledge-Based Systems 342 (2026) 115850. URL: <https://www.sciencedirect.com/science/article/pii/S0950705126005769>. doi:https://doi.org/10.1016/j.knosys.2026.115850.
- [4] N. Saphra, A. Lopez, Understanding learning dynamics of language models with SVCCA, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3257–3267. URL: <https://aclanthology.org/N19-1329/>. doi:10.18653/v1/N19-1329.
- [5] O. van der Wal, P. Lesci, M. Müller-Eberstein, N. Saphra, H. Schoelkopf, W. Zuidema, S. Biderman, Polypythias: Stability and outliers across fifty language model pre-training runs, in: The Thirteenth International Conference on Learning Representations, 2025. URL: <https://openreview.net/forum?id=bmrYu2Ekdz>.
- [6] F. Fehlaue, K. Mahowald, T. Pimentel, Convergence and divergence of language models under different random seeds, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 32982–32991. URL: <https://aclanthology.org/2025.emnlp-main.1675/>. doi:10.18653/v1/2025.emnlp-main.1675.
- [7] J. Mange, Effect of training data order for machine learning, in: 2019 International Conference on Computational Science and Computational Intelligence (CSCI), 2019, pp. 406–407. doi:10.1109/CSCI49370.2019.00078.
- [8] M. Y. Hu, A. Chen, N. Saphra, K. Cho, Latent state models of training dynamics, Transactions on Machine Learning Research (2023). URL: <https://openreview.net/forum?id=NE2xXWo0LF>.

- [9] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 41–48. URL: <https://doi.org/10.1145/1553374.1553380>. doi:10.1145/1553374.1553380.
- [10] L. Ranaldi, G. Pucci, F. M. Zanzotto, Modeling easiness for training transformers with curriculum learning, in: R. Mitkov, G. Angelova (Eds.), Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 937–948. URL: <https://aclanthology.org/2023.ranlp-1.101/>.
- [11] A. Agrawal, S. Singh, L. Schneider, M. Samuels, On the role of corpus ordering in language modeling, in: N. S. Moosavi, I. Gurevych, A. Fan, T. Wolf, Y. Hou, A. Marasović, S. Ravi (Eds.), Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing, Association for Computational Linguistics, Virtual, 2021, pp. 142–154. URL: <https://aclanthology.org/2021.sustainlp-1.15/>. doi:10.18653/v1/2021.sustainlp-1.15.
- [12] R. Diehl Martinez, Z. Goriely, H. McGovern, C. Davis, A. Caines, P. Buttery, L. Beinborn, CLIMB – curriculum learning for infant-inspired model building, in: A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, R. Cotterell (Eds.), Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Singapore, 2023, pp. 112–127. URL: <https://aclanthology.org/2023.conll-babylm.10/>. doi:10.18653/v1/2023.conll-babylm.10.
- [13] A. Graves, M. G. Bellemare, J. Menick, R. Munos, K. Kavukcuoglu, Automated curriculum learning for neural networks, in: international conference on machine learning, Pmlr, 2017, pp. 1311–1320.
- [14] Y. Wan, B. Yang, D. F. Wong, Y. Zhou, L. S. Chao, H. Zhang, B. Chen, Self-paced learning for neural machine translation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 1074–1080. URL: <https://aclanthology.org/2020.emnlp-main.80/>. doi:10.18653/v1/2020.emnlp-main.80.
- [15] Y. Huang, D. Xiong, IT2ACL learning easy-to-hard instructions via 2-phase automated curriculum learning for large language models, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9405–9421. URL: <https://aclanthology.org/2024.lrec-main.822/>.
- [16] X. Tang, J. Wang, Q. Su, C.-R. Huang, J. Gu, An effective incorporating heterogeneous knowledge curriculum learning for sequence labeling, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2025, pp. 495–503.
- [17] R. Entezari, H. Sedghi, O. Saukh, B. Neyshabur, The role of permutation invariance in linear mode connectivity of neural networks, in: International Conference on Learning Representations, 2022. URL: <https://openreview.net/forum?id=dNigytemkL>.
- [18] E. S. Ruzzetti, F. Ranaldi, F. Logozzo, M. Mastromattei, L. Ranaldi, F. M. Zanzotto, Exploring linguistic properties of monolingual BERTs with typological classification among languages, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 14447–14461. URL: <https://aclanthology.org/2023.findings-emnlp.963/>. doi:10.18653/v1/2023.findings-emnlp.963.
- [19] D. R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Computation 16 (2004) 2639–2664. doi:10.1162/0899766042321814.
- [20] M. Raghu, J. Gilmer, J. Yosinski, J. Sohl-Dickstein, Svcca: singular vector canonical correlation analysis for deep learning dynamics and interpretability, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6078–6087.
- [21] A. S. Morcos, M. Raghu, S. Bengio, Insights on representational similarity in neural networks with canonical correlation, in: Proceedings of the 32nd International Conference on Neural Information

- Processing Systems, NIPS'18, Curran Associates Inc., Red Hook, NY, USA, 2018, p. 5732–5741.
- [22] S. Kornblith, M. Norouzi, H. Lee, G. Hinton, Similarity of neural network representations revisited, in: K. Chaudhuri, R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 3519–3529. URL: <https://proceedings.mlr.press/v97/kornblith19a.html>.
  - [23] B. Neyshabur, H. Sedghi, C. Zhang, What is being transferred in transfer learning?, in: *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Curran Associates Inc., Red Hook, NY, USA, 2020.
  - [24] T. Gao, A. Dhurandhar, K. N. Ramamurthy, D. Wei, Representation similarity reveals implicit layer grouping in neural networks, in: *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL: <https://openreview.net/forum?id=Ha5UIyfjTO>.
  - [25] P. Lucisano, M. E. Piemontese, Gulpease: Una formula per la predizione della difficoltà dei testi in lingua italiana, *Scuola e Città* 3 (1988) 57–68.
  - [26] F. Dell'Orletta, S. Montemagni, G. Venturi, READ-IT: Assessing readability of Italian texts with a view to text simplification, in: N. Alm (Ed.), *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 73–83. URL: <https://aclanthology.org/W11-2308/>.
  - [27] E. Boix-Adsera, H. Lawrence, G. Stepaniants, P. Rigollet, GULP: a prediction-based metric between representations, in: *Advances in Neural Information Processing Systems*, volume 35, 2022.
  - [28] L. Ciernik, L. Linhardt, M. Morik, J. Dippel, S. Kornblith, L. Muttenthaler, Objective drives the consistency of representational similarity across datasets, in: *Forty-second International Conference on Machine Learning*, 2025. URL: <https://openreview.net/forum?id=va3zmBXPat>.
  - [29] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, S. Montemagni, Profiling-UD: a tool for linguistic profiling of texts, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2020, pp. 7145–7151. URL: <https://aclanthology.org/2020.lrec-1.883/>.
  - [30] M.-C. De Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, *Computational linguistics* 47 (2021) 255–308.

## A. Sentence-level contribution to CKA

Although CKA quantifies the similarity between two representation spaces, it is a global metric defined over representation matrices rather than individual vectors. Consequently, it cannot be directly used to assign a CKA score to a single sentence. To obtain a sentence-level measure, we therefore decompose the linear CKA computation across examples.

Let  $X, Y \in \mathbb{R}^{n \times d}$  be the centered representations of the same  $n$  sentences extracted from two models, and let  $K = XX^T$  and  $L = YY^T$  be the corresponding Gram matrices. The standard linear CKA score can be written as:

$$CKA(X, Y) = \frac{\sum_{i=1}^n \sum_{j=1}^n K_{ij} L_{ij}}{\sqrt{\sum_{i=1}^n \sum_{j=1}^n K_{ij}^2 \sum_{i=1}^n \sum_{j=1}^n L_{ij}^2}}$$

We define the contribution of sentence  $i$  as the fraction of the normalized CKA numerator associated with the  $i$ -th row of the Gram matrices:

$$c_i(X, Y) = \frac{\sum_{j=1}^n K_{ij} L_{ij}}{\sqrt{\sum_{p=1}^n \sum_{q=1}^n K_{pq}^2 \sum_{p=1}^n \sum_{q=1}^n L_{pq}^2}}$$

This score captures the extent to which the similarity relations of sentence  $i$  with all other sentences are preserved across the two representation spaces. Sentences with larger positive values are those

whose pairwise similarity patterns are more consistently preserved across the two models, and therefore contribute more strongly to the overall CKA similarity. This quantity is not the CKA of an individual sentence, since CKA is not defined for isolated vectors, but allows us to rank sentences according to how strongly they support the observed representational similarity between two models.

## B. biCKA Modules

This section complements the weight-level analysis presented in Section 5.1. While Figure 2 summarizes the average biCKA scores for the last layer, here we report the corresponding module-wise similarity matrices. In particular, Figure B.1 shows the biCKA scores between model pairs separately for each parameter matrix considered in the last Transformer layer.

These plots provide a more fine-grained view of the patterns discussed in the main results. Overall, the module-wise matrices confirm that the effect of pre-training order on the learned parameters is not confined to a single component. Although some modules show stronger sensitivity than others, the same broad organization of model similarities is visible across the different parts of the last layer.

biCKA layer 8 Q

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .34 | .39 | .33 | .36 | .33 | .33 | .33 | .34 |
| SentLen_Inv  | .34 | 1   | .36 | .46 | .37 | .46 | .46 | .47 | .50 |
| Gulpease     | .39 | .36 | 1   | .35 | .39 | .36 | .37 | .37 | .38 |
| Gulpease_Inv | .33 | .46 | .35 | 1   | .37 | .43 | .45 | .44 | .46 |
| ReadIt       | .36 | .37 | .39 | .37 | 1   | .35 | .38 | .38 | .40 |
| ReadIt_Inv   | .33 | .46 | .36 | .43 | .35 | 1   | .44 | .43 | .45 |
| Orig         | .33 | .46 | .37 | .45 | .38 | .44 | 1   | .51 | .52 |
| Orig_Inv     | .33 | .47 | .37 | .44 | .38 | .43 | .51 | 1   | .56 |
| Avg_Rand     | .34 | .50 | .38 | .46 | .40 | .45 | .52 | .56 | .65 |

biCKA layer 8 K

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .28 | .33 | .28 | .31 | .27 | .27 | .27 | .28 |
| SentLen_Inv  | .28 | 1   | .29 | .44 | .32 | .44 | .44 | .45 | .48 |
| Gulpease     | .33 | .29 | 1   | .30 | .31 | .29 | .30 | .30 | .31 |
| Gulpease_Inv | .28 | .44 | .30 | 1   | .32 | .40 | .42 | .42 | .44 |
| ReadIt       | .31 | .32 | .31 | .32 | 1   | .32 | .34 | .33 | .35 |
| ReadIt_Inv   | .27 | .44 | .29 | .40 | .32 | 1   | .42 | .41 | .42 |
| Orig         | .27 | .44 | .30 | .42 | .34 | .42 | 1   | .49 | .51 |
| Orig_Inv     | .27 | .45 | .30 | .42 | .33 | .41 | .49 | 1   | .56 |
| Avg_Rand     | .28 | .48 | .31 | .44 | .35 | .42 | .51 | .56 | .65 |

biCKA layer 8 V

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .57 | .58 | .56 | .59 | .57 | .57 | .58 | .58 |
| SentLen_Inv  | .57 | 1   | .58 | .60 | .59 | .62 | .62 | .62 | .63 |
| Gulpease     | .58 | .58 | 1   | .57 | .59 | .58 | .58 | .59 | .58 |
| Gulpease_Inv | .56 | .60 | .57 | 1   | .58 | .59 | .60 | .60 | .60 |
| ReadIt       | .59 | .59 | .59 | .58 | 1   | .59 | .60 | .60 | .60 |
| ReadIt_Inv   | .57 | .62 | .58 | .59 | .59 | 1   | .61 | .61 | .61 |
| Orig         | .57 | .62 | .58 | .60 | .60 | .61 | 1   | .64 | .64 |
| Orig_Inv     | .58 | .62 | .59 | .60 | .60 | .61 | .64 | 1   | .65 |
| Avg_Rand     | .58 | .63 | .58 | .60 | .60 | .61 | .64 | .65 | .71 |

biCKA layer 8 OA

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .44 | .46 | .43 | .45 | .43 | .43 | .43 | .44 |
| SentLen_Inv  | .44 | 1   | .45 | .50 | .46 | .51 | .51 | .51 | .53 |
| Gulpease     | .46 | .45 | 1   | .43 | .46 | .44 | .45 | .45 | .46 |
| Gulpease_Inv | .43 | .50 | .43 | 1   | .45 | .48 | .49 | .48 | .50 |
| ReadIt       | .45 | .46 | .46 | .45 | 1   | .44 | .46 | .46 | .47 |
| ReadIt_Inv   | .43 | .51 | .44 | .48 | .44 | 1   | .49 | .49 | .50 |
| Orig         | .43 | .51 | .45 | .49 | .46 | .49 | 1   | .53 | .54 |
| Orig_Inv     | .43 | .51 | .45 | .48 | .46 | .49 | .53 | 1   | .55 |
| Avg_Rand     | .44 | .53 | .46 | .50 | .47 | .50 | .54 | .55 | .64 |

biCKA layer 8 DI

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .57 | .72 | .57 | .67 | .57 | .60 | .60 | .59 |
| SentLen_Inv  | .57 | 1   | .61 | .72 | .58 | .67 | .72 | .73 | .73 |
| Gulpease     | .72 | .61 | 1   | .61 | .70 | .61 | .66 | .65 | .64 |
| Gulpease_Inv | .57 | .72 | .61 | 1   | .57 | .66 | .71 | .72 | .71 |
| ReadIt       | .67 | .58 | .70 | .57 | 1   | .57 | .61 | .61 | .62 |
| ReadIt_Inv   | .57 | .67 | .61 | .66 | .57 | 1   | .67 | .67 | .66 |
| Orig         | .60 | .72 | .66 | .71 | .61 | .67 | 1   | .82 | .80 |
| Orig_Inv     | .60 | .73 | .65 | .72 | .61 | .67 | .82 | 1   | .84 |
| Avg_Rand     | .59 | .73 | .64 | .71 | .62 | .66 | .80 | .84 | .87 |

biCKA layer 8 DO

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .44 | .46 | .43 | .45 | .44 | .43 | .44 | .44 |
| SentLen_Inv  | .44 | 1   | .45 | .51 | .46 | .52 | .52 | .53 | .54 |
| Gulpease     | .46 | .45 | 1   | .44 | .46 | .45 | .45 | .46 | .46 |
| Gulpease_Inv | .43 | .51 | .44 | 1   | .45 | .49 | .50 | .50 | .51 |
| ReadIt       | .45 | .46 | .46 | .45 | 1   | .45 | .46 | .47 | .47 |
| ReadIt_Inv   | .44 | .52 | .45 | .49 | .45 | 1   | .50 | .50 | .51 |
| Orig         | .43 | .52 | .45 | .50 | .46 | .50 | 1   | .56 | .55 |
| Orig_Inv     | .44 | .53 | .46 | .50 | .47 | .50 | .56 | 1   | .58 |
| Avg_Rand     | .44 | .54 | .46 | .51 | .47 | .51 | .55 | .58 | .66 |

**Figure B.1:** biCKA of model pairs, for each weight module of the last models' layer.

## C. CKA Modules

This section extends the representation-level analysis reported in Section 5.2. The average last-layer CKA scores are shown in Figure 4; here, we break them down by module to inspect whether the same similarity structure holds for the different representations extracted from the final Transformer layer. Figures C.2 and C.1 report the resulting similarity matrices.

The separate matrices show that the trends observed in the aggregate analysis are largely preserved at the module level. Thus, the impact of pre-training data order on model representations does not appear to be driven by a single activation type, but rather emerges consistently across the considered components, including the hidden states.

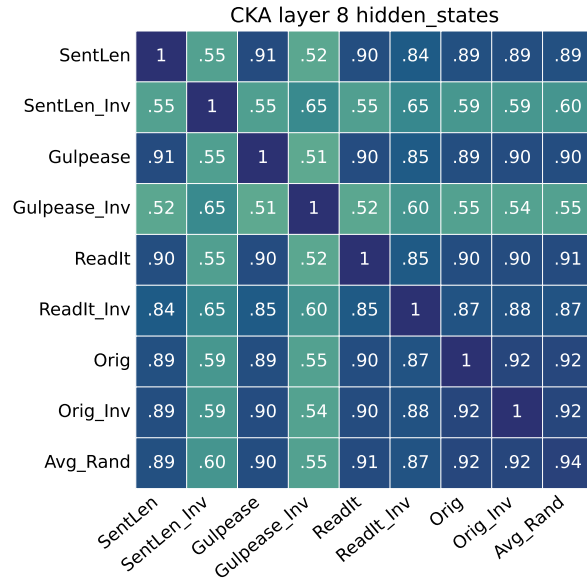


**Figure C.1:** CKA of hidden states (model output)

CKA layer 8 Q

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .52 | .85 | .44 | .85 | .75 | .84 | .83 | .82 |
| SentLen_Inv  | .52 | 1   | .53 | .54 | .52 | .59 | .56 | .56 | .56 |
| Gulpease     | .85 | .53 | 1   | .43 | .84 | .76 | .84 | .84 | .83 |
| Gulpease_Inv | .44 | .54 | .43 | 1   | .43 | .52 | .45 | .46 | .46 |
| ReadIt       | .85 | .52 | .84 | .43 | 1   | .77 | .85 | .85 | .84 |
| ReadIt_Inv   | .75 | .59 | .76 | .52 | .77 | 1   | .79 | .80 | .80 |
| Orig         | .84 | .56 | .84 | .45 | .85 | .79 | 1   | .88 | .86 |
| Orig_Inv     | .83 | .56 | .84 | .46 | .85 | .80 | .88 | 1   | .87 |
| Avg_Rand     | .82 | .56 | .83 | .46 | .84 | .80 | .86 | .87 | .90 |

CKA layer 8 K

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .48 | .84 | .63 | .83 | .82 | .80 | .82 | .84 |
| SentLen_Inv  | .48 | 1   | .48 | .54 | .50 | .46 | .55 | .56 | .54 |
| Gulpease     | .84 | .48 | 1   | .60 | .85 | .79 | .83 | .85 | .85 |
| Gulpease_Inv | .63 | .54 | .60 | 1   | .57 | .68 | .61 | .63 | .63 |
| ReadIt       | .83 | .50 | .85 | .57 | 1   | .77 | .84 | .84 | .85 |
| ReadIt_Inv   | .82 | .46 | .79 | .68 | .77 | 1   | .78 | .79 | .82 |
| Orig         | .80 | .55 | .83 | .61 | .84 | .78 | 1   | .89 | .88 |
| Orig_Inv     | .82 | .56 | .85 | .63 | .84 | .79 | .89 | 1   | .89 |
| Avg_Rand     | .84 | .54 | .85 | .63 | .85 | .82 | .88 | .89 | .92 |

CKA layer 8 V

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .64 | .85 | .53 | .85 | .82 | .84 | .85 | .85 |
| SentLen_Inv  | .64 | 1   | .60 | .61 | .63 | .67 | .64 | .63 | .65 |
| Gulpease     | .85 | .60 | 1   | .50 | .84 | .81 | .84 | .85 | .85 |
| Gulpease_Inv | .53 | .61 | .50 | 1   | .52 | .57 | .53 | .52 | .52 |
| ReadIt       | .85 | .63 | .84 | .52 | 1   | .82 | .85 | .86 | .86 |
| ReadIt_Inv   | .82 | .67 | .81 | .57 | .82 | 1   | .83 | .84 | .84 |
| Orig         | .84 | .64 | .84 | .53 | .85 | .83 | 1   | .89 | .88 |
| Orig_Inv     | .85 | .63 | .85 | .52 | .86 | .84 | .89 | 1   | .88 |
| Avg_Rand     | .85 | .65 | .85 | .52 | .86 | .84 | .88 | .88 | .91 |

CKA layer 8 OA

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .49 | .65 | .38 | .62 | .60 | .57 | .59 | .62 |
| SentLen_Inv  | .49 | 1   | .54 | .53 | .53 | .58 | .54 | .55 | .55 |
| Gulpease     | .65 | .54 | 1   | .37 | .65 | .65 | .66 | .66 | .68 |
| Gulpease_Inv | .38 | .53 | .37 | 1   | .41 | .45 | .36 | .39 | .36 |
| ReadIt       | .62 | .53 | .65 | .41 | 1   | .62 | .63 | .63 | .66 |
| ReadIt_Inv   | .60 | .58 | .65 | .45 | .62 | 1   | .67 | .69 | .69 |
| Orig         | .57 | .54 | .66 | .36 | .63 | .67 | 1   | .70 | .71 |
| Orig_Inv     | .59 | .55 | .66 | .39 | .63 | .69 | .70 | 1   | .72 |
| Avg_Rand     | .62 | .55 | .68 | .36 | .66 | .69 | .71 | .72 | .78 |

CKA layer 8 DI

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .58 | .88 | .51 | .87 | .79 | .86 | .87 | .85 |
| SentLen_Inv  | .58 | 1   | .56 | .61 | .58 | .69 | .63 | .63 | .65 |
| Gulpease     | .88 | .56 | 1   | .49 | .88 | .79 | .87 | .87 | .86 |
| Gulpease_Inv | .51 | .61 | .49 | 1   | .51 | .57 | .54 | .54 | .55 |
| ReadIt       | .87 | .58 | .88 | .51 | 1   | .80 | .88 | .88 | .87 |
| ReadIt_Inv   | .79 | .69 | .79 | .57 | .80 | 1   | .85 | .85 | .86 |
| Orig         | .86 | .63 | .87 | .54 | .88 | .85 | 1   | .91 | .89 |
| Orig_Inv     | .87 | .63 | .87 | .54 | .88 | .85 | .91 | 1   | .90 |
| Avg_Rand     | .85 | .65 | .86 | .55 | .87 | .86 | .89 | .90 | .92 |

CKA layer 8 DO

|              |     |     |     |     |     |     |     |     |     |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| SentLen      | 1   | .39 | .81 | .34 | .80 | .66 | .79 | .79 | .78 |
| SentLen_Inv  | .39 | 1   | .40 | .54 | .44 | .55 | .47 | .47 | .48 |
| Gulpease     | .81 | .40 | 1   | .34 | .80 | .66 | .79 | .78 | .78 |
| Gulpease_Inv | .34 | .54 | .34 | 1   | .37 | .47 | .40 | .39 | .40 |
| ReadIt       | .80 | .44 | .80 | .37 | 1   | .69 | .81 | .79 | .80 |
| ReadIt_Inv   | .66 | .55 | .66 | .47 | .69 | 1   | .72 | .73 | .73 |
| Orig         | .79 | .47 | .79 | .40 | .81 | .72 | 1   | .83 | .83 |
| Orig_Inv     | .79 | .47 | .78 | .39 | .79 | .73 | .83 | 1   | .83 |
| Avg_Rand     | .78 | .48 | .78 | .40 | .80 | .73 | .83 | .83 | .87 |

Figure C.2: CKA of model pairs, for each weight module of the last models' layer.

## D. Trends Across Curriculum Types

This section provides a more detailed analysis of the layer-wise impact of pre-training data order on model similarity. In particular, we aim to disentangle the effect of data ordering due to random shuffling from the effect of an explicit curriculum, where pre-training samples are ordered according to a predefined function, in our case based on linguistic complexity.

To this end, we repeat the similarity analyses presented in the main text while separating models according to the type of training order: random orderings, forward curricula, and backward curricula. For each group, similarity scores are computed by comparing the considered models only against models trained with random orderings. This allows us to isolate the variability induced by random shuffling when analyzing random models, while avoiding possible interactions among different curriculum strategies when analyzing curriculum-trained models.

### D.1. biCKA

Figure D.1 reports the average biCKA scores obtained when comparing random models against other random models. Compared with the corresponding overall plot in Figure 1, the layer-wise patterns are largely preserved, although the similarity scores are generally higher by approximately 0.05–0.1 points. A notable exception is the  $K$  module, which exhibits substantially higher similarity in this setting. This suggests that  $K$  is the weight module most affected by the inclusion of curriculum-trained models in the general analysis.

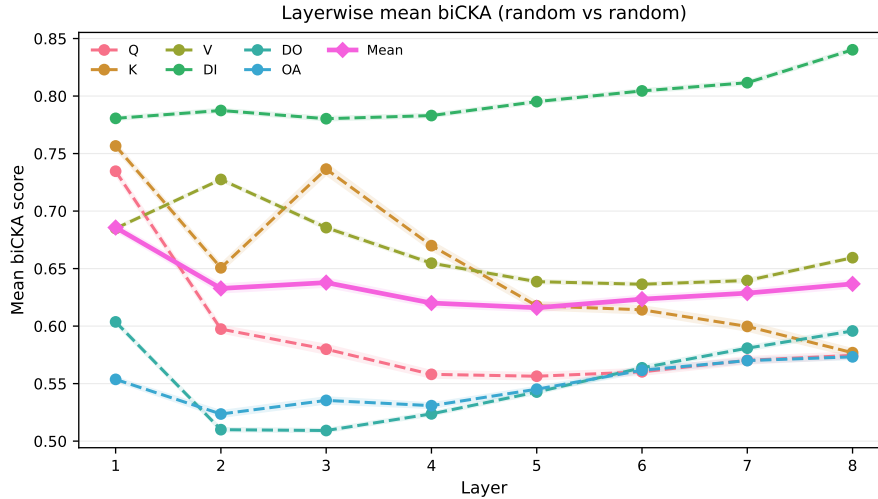


**Figure D.1:** Layer-wise biCKA scores for random models compared against random models.

Figure D.2 reports the results for models trained with forward curricula, compared against random models. In this case, the similarity scores are clearly lower than those observed in the general analysis. Among the attention modules,  $V$  shows the smallest decrease in similarity, whereas  $K$  is the most affected module.

Figure D.3 reports the results for models trained with backward curricula, compared against random models. The resulting pattern is very close to the one observed in the overall analysis, suggesting that backward curricula behave similarly to the aggregate trend when measured with biCKA.

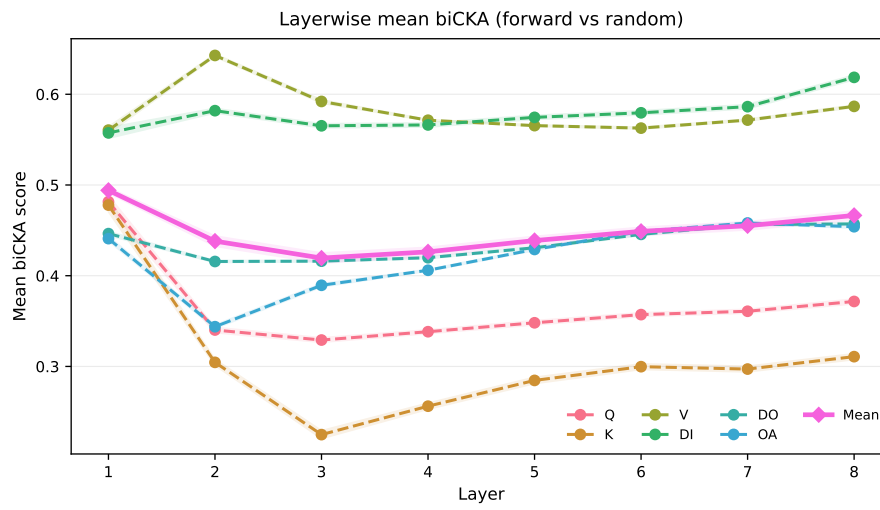


**Figure D.2:** Layer-wise biCKA scores for forward curriculum models compared against random models.

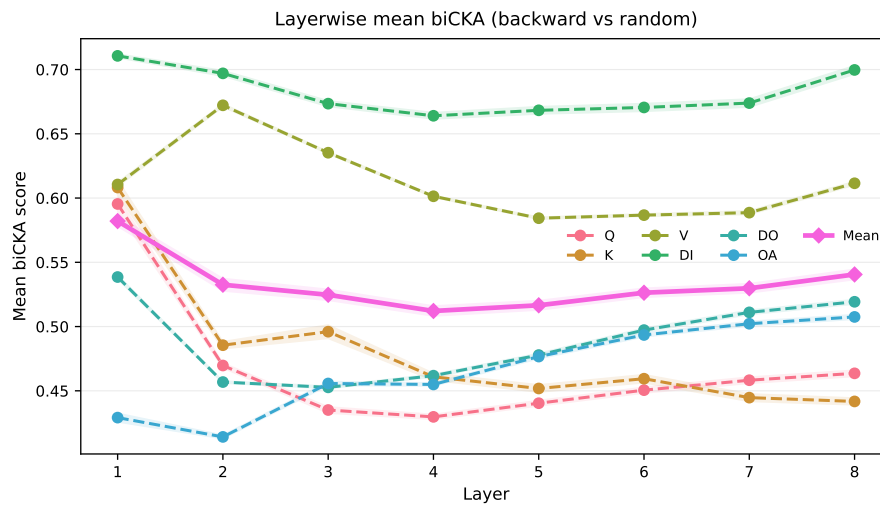


**Figure D.3:** Layer-wise biCKA scores for backward curriculum models compared against random models.

## D.2. CKA Trends Across Curriculum Types

We now report the analysis of the CKA scores, using the global plot in Figure 3 as a reference.

Figure D.4 shows the results computed only among random models. Overall, the scores are higher than in the global analysis, with a positive effect across all modules and values approaching near-identical similarity for most modules in the first layer. Despite this increase, the overall layer-wise behavior remains unchanged, confirming the trends observed in the global analysis.

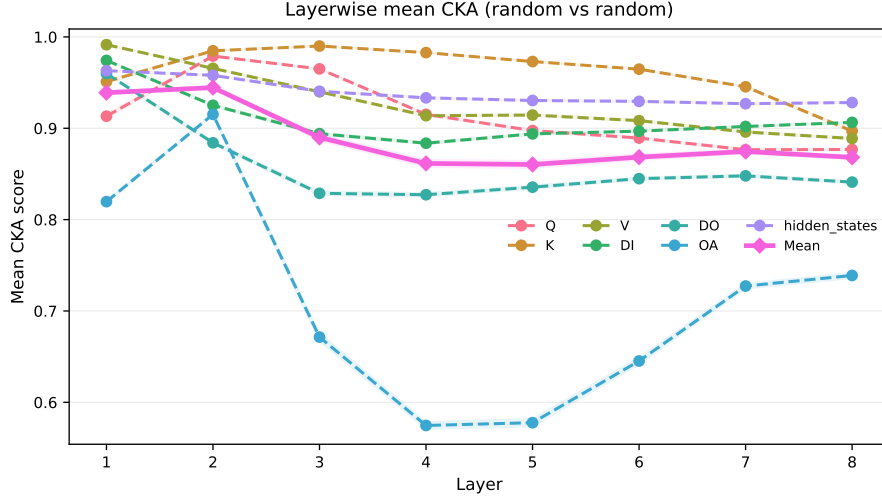


**Figure D.4:** CKA random vs random

Figure D.5 reports the results for models trained with forward curricula. The scores are lower in the first layer for some modules; however, the plot largely mirrors the original analysis, both in terms of layer-wise trends and absolute scores.

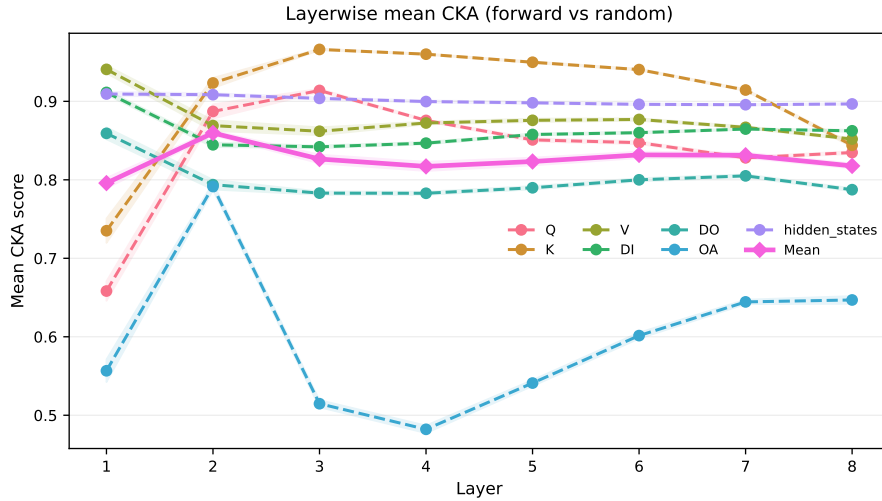


**Figure D.5:** CKA forward curricula vs random

Finally, Figure D.6 reports the analysis for models trained with backward curricula. In contrast to the previous plots, this is the first case in which the error bars become clearly visible, suggesting that backward-curriculum models display greater variability in their similarity scores. Here, the scores are substantially lower than in the original plot, with a steeper decrease in the final layers. This indicates that backward pre-training curricula have the strongest impact on the learned representations.

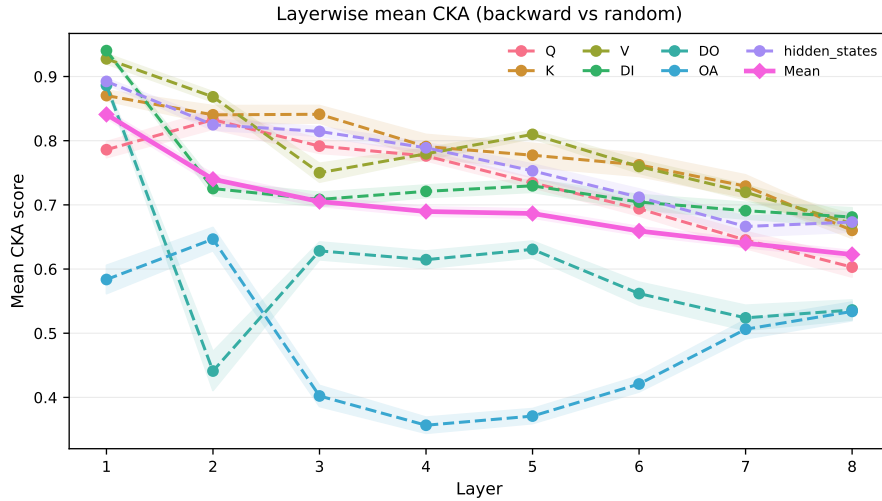


Figure D.6: CKA backward curricula vs random

## E. Most Distant Sentences

Table 1 reports the five held-out sentences with the lowest CKA contribution scores for each training data ordering. The selected examples are highly heterogeneous and do not share a single defining surface characteristic. While some are short, truncated, or structurally simple, others are longer and contain coordination, embedded clauses, enumerations, numerical expressions, or marked punctuation patterns.

Nevertheless, some curriculum-specific tendencies emerge. For *SentLen\_Inv*, all five most distant sentences are truncated examples, whereas for *Gulpease\_Inv* three of the five most distant sentences have the minimum length included in the evaluation set (six tokens). Moreover, several sentences recur across multiple training conditions, suggesting the presence of stable representational outliers that remain sensitive to variation in pre-training trajectory. Overall, the examples indicate that representational divergence cannot be reduced to a single linguistic property and instead arises from a combination of factors that varies across curricula.

Table 1: Text of the five most distant sentences for each training data order.

| Ordering    | Sentence                                                                                                                                                                                                                                                                       |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SentLen     | Nell'occasione venne investito da una moto della polizia, che gli procurò una ferita alla testa, ma riuscì comunque a portare a termine la frazione; nella tappa seguente prese il via da Roma, ma dopo circa 50 km, nei pressi di Segni Scalo, preferì abbandonare.           |
|             | Nel territorio comunale si trovano le cascate Reichenbach formate dal fiume Aar.                                                                                                                                                                                               |
|             | Storia.                                                                                                                                                                                                                                                                        |
|             | Nel corso del XVI secolo la tecnica del "cantus firmus" cominciò ad essere abbandonata, sostituita con la parodia (o imitazione) tecnica, nella quale più voci di una fonte preesistenti vennero inserite in una composizione sacra come una messa.                            |
|             | Il generale Koenig è oramai al corrente che il nemico può passare agevolmente a nord dei campi di mine, la missione affidata ai francesi di disturbare i rifornimenti nelle retrovie del nemico e di presidiare gli sbarramenti di mine al nord, volge così oramai al termine. |
| SentLen_Inv | Relegati oggi in una tra le più povere terre del mondo, il Kalahari, i Boscimani rischiano ancora di dover trasferirsi in quanto il governo del Botswana vorrebbe integrarli ai benefici della civiltà .                                                                       |
|             | Alla Rotonda si può arrivare direttamente prendendo la via del Papa; prima di giungervi si incontreranno a mano destra, oltre il cimitero (arch.                                                                                                                               |

Continued on the next page

| Ordering     | Sentence                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|              | <p>Finite le lezioni, si occupava degli incarichi ricevuti da Hunter, quali dissezioni, preparazione di medicine, ecc.</p> <p>Nonostante il nome, non è una forza, ma una accelerazione; sebbene, nei casi di cui sopra sia una accelerazione relativa al riferimento considerato, determina una forza fittizia (es.</p> <p>quella della tattica nei confronti delle lotte immediate e degli organismi di lotta proletari indipendenti dal collaborazionismo tricolore, e quella nota storicamente come ‘nazionale e coloniale’ (“A proposito di Medio Oriente e Maghreb.</p> <p>Naturalmente opere del genere si basano più che altro sugli scritti di autori greci classici, e includono tutti i significati attribuiti alla parola dai filosofi greci pagani, fra cui “anima” dei morti, “anima, come sussistente senza il corpo, o contrapposta ad esso”, ecc.</p>                                                                                                |
| Gulpease     | <p>La riconversione industriale venne applicata all’industria navale, alla siderurgia ed all’industria tessile, settori tecnologicamente danneggiati, geograficamente mal ubicati e non competitivi rispetto ai prodotti europei ed asiatici.</p> <p>uno per il mercato italiano, due per quello francese, due per quello spagnolo, due per quello olandese e, infine, uno per quello tedesco.</p> <p>A determinate scene dell’Antico Testamento inoltre si contrappongono altre del Nuovo.</p> <p>Il modello E0 è stato il primo modello bipede (a due zampe) prodotto come parte della Honda serie E, che era una linea iniziale sperimentale di robot umanoidi creati tra il 1986 e il 1993.</p> <p>Di recente, alcune caratteristiche insolite, riguardanti principalmente le vertebre stranamente biforcute, hanno mostrato agli scienziati che questo gruppo apparteneva con tutta probabilità a una radiazione di sauropodi endemica dell’Estremo Oriente.</p> |
| Gulpease_Inv | <p>Due delle stazioni sono fuori terra.</p> <p>Il tartufo è grosso e nero.</p> <p>Tutta la zona è densamente abitata.</p> <p>Finite le lezioni, si occupava degli incarichi ricevuti da Hunter, quali dissezioni, preparazione di medicine, ecc.</p> <p>Sophia produsse il mondo tramite una successiva evoluzione di sizigie, in cui la femminile precedeva sempre la maschile, per poi esserne, infine, superata.</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| ReadIt       | <p>Ogni famiglia originaria del luogo fa parte del comune patriziale che ha la responsabilità della manutenzione di ogni bene comune.</p> <p>Ogni famiglia originaria del luogo fa parte del cosiddetto comune patriziale e ha la responsabilità della manutenzione di ogni bene ricadente all’interno dei confini del comune.</p> <p>nel 1931, infatti, esperti calcistici di tutta Europa riconoscono le cavallette tra le quattro migliori squadre del continente.</p> <p>Relegati oggi in una tra le più povere terre del mondo, il Kalahari, i Boscimani rischiano ancora di dover trasferirsi in quanto il governo del Botswana vorrebbe integrarli ai benefici della civiltà .</p> <p>Negli anni successivi inoltre furono introdotte altre cerimonie minori riservate a particolari categorie di programmi.</p>                                                                                                                                               |
| ReadIt_Inv   | <p>Ogni famiglia originaria del luogo fa parte del comune patriziale che ha la responsabilità della manutenzione di ogni bene comune.</p> <p>uno per il mercato italiano, due per quello francese, due per quello spagnolo, due per quello olandese e, infine, uno per quello tedesco.</p> <p>nel 1830 venne introdotto un sistema di taglie da assegnare per ogni animale ucciso, e nel solo 1923 vennero abbattuti più di 900.000 diavoli orsini, sia con armi da fuoco sia attraverso metodi indiretti, come trappole ed esche avvelenate.</p> <p>Relegati oggi in una tra le più povere terre del mondo, il Kalahari, i Boscimani rischiano ancora di dover trasferirsi in quanto il governo del Botswana vorrebbe integrarli ai benefici della civiltà .</p>                                                                                                                                                                                                     |

*Continued on the next page*

| Ordering   | Sentence                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|            | Moai è anche il nome dato a piccole figure scolpite, alte in media quaranta centimetri, della cultura dell'Isola di Pasqua, realizzate principalmente in legno di Toromiro .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| Avg_Random | <p>il maschile, il femminile ed il neutro.</p> <p>Se solo due nazioni fossero state votate, queste avrebbero ottenuto rispettivamente sei e tre punti.</p> <p>Relegati oggi in una tra le più povere terre del mondo, il Kalahari, i Boscimani rischiano ancora di dover trasferirsi in quanto il governo del Botswana vorrebbe integrarli ai benefici della civiltà .</p> <p>nel 1931, infatti, esperti calcistici di tutta Europa riconoscono le cavallette tra le quattro migliori squadre del continente.</p> <p>La bandiera è composta da tre bande verticali di uguali dimensioni, blu, simbolo del cielo e colore nazionale, quella centrale e rosse quelle ai lati.</p> |