Everything, everywhere, all at once: how learning order shapes the embedding space.

Lucia Domenichelli
National PhD in Artificial Intelligence

Abstract

This project investigates how the sequencing of pretraining data, commonly known as curriculum learning, shapes the internal representations of neural language models. We pretrain a BERT-Medium model on an Italian text corpus under multiple curricula organized by sentence complexity, then evaluate each model's performance on downstream NLP benchmarks, its acquisition of linguistic knowledge, and the geometric properties of its embedding space. Although structured curricula yield only modest gains on traditional task metrics, they produce markedly less isotropic representations with reduced intrinsic dimensionality, indicating a tighter, more focused organization of learned features. Our results demonstrate that the order in which models encounter training examples not only affects their external behavior but also fundamentally alters the way they encode and structure linguistic information.

1 Introduction

Natural Language Processing (NLP) develops algorithms that allow computers to understand and generate human language. Early work used task, specific classifiers, such as Support Vector Machines or Recurrent Neural Networks, built on handcrafted syntactic and semantic features. Modern *neural language models* (NLMs) abandon manual engineering: large Transformer networks are first *pre-trained* on massive corpora to learn rich linguistic representations and then *fine-tuned* on each downstream task. This strategy achieves state-of-the-art accuracy and real-time inference, but it sacrifices interpretability because the inner workings of deep networks remain largely opaque.

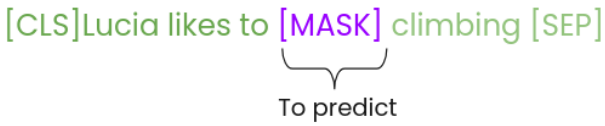

Figure 1: The Masked Language Modeling task.

The Transformer The 2017 paper *"Attention Is All You Need"* [10] introduced the *Transformer*, an encoder–decoder network that relies exclusively on self-attention to capture global dependencies. Given an input sequence of symbol embeddings $x = (x_1, \ldots, x_n)$, the encoder produces contextual vectors $z = (z_1, \ldots, z_n)$. The decoder

then autoregressively generates the output sequence $y = (y_1, \ldots, y_m)$, conditioning on z and previously generated tokens. Both encoder and decoder are stacks of identical layers composed of multi-head self-attention and position-wise feed-forward networks. A single attention head maps a query matrix $\mathbf{Q}$, key matrix $\mathbf{K}$ and value matrix $\mathbf{V}$ to:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) \;=\; \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V},$$

where d_k is the dimensionality of the keys and queries. Self-attention lets every token attend to all others, enabling the model to encode long-range relationships efficiently. Although conceived for *machine translation*, Transformers soon powered state-of-the-art language models such as BERT and GPT, thanks to their superior handling of long-distance dependencies and parallelizable training dynamics.

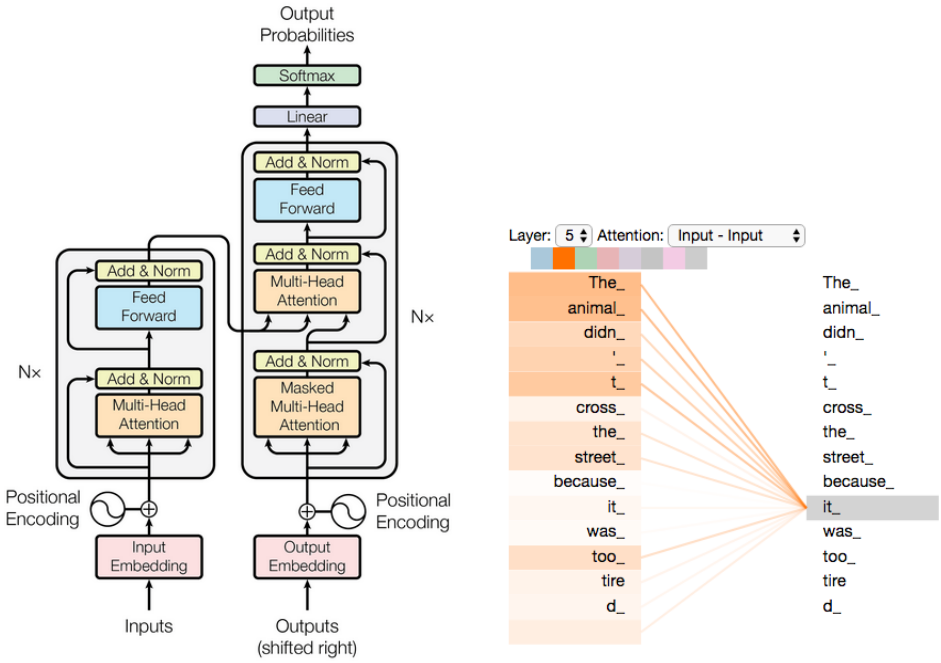

Figure 2: The Transformer architecture and the self-attention mechanism.

BERT Bidirectional Encoder Representations from Transformers (BERT) [5] replaces the encoder–decoder setup with a deep stack of Transformer *encoders* only, allowing each token to attend to all others on both the left and right context. The model is first *pre-trained* on the MLM task (predicting $\approx 15\%$ randomly masked tokens). After pre-training, the same parameters are *fine-tuned* by adding a lightweight task head (classification, span extraction, etc.) and optimizing the corresponding downstream loss. This simple "pre-train once, fine-tune many" recipe yields state-of-the-art results across a wide spectrum of NLP benchmarks with minimal task-specific engineering.

2 Methods

This study asks whether the *order* in which pre-training data is seen changes a neural language model's behaviour. Inspired by *curriculum learning*, we train models on corpora

arranged from "easy" to "hard" according to linguistic and structural heuristics, and compare them with models exposed to the same data in random order.
Outside of this report, all experiments were replicated on 2 others random seeds of data for a more robust analysis.

Research Questions

1. **Downstream performance**: Does curriculum-based pre-training yield higher performance on standard benchmarks?

2. **Linguistic knowledge**: Does the resulting representation encode syntax and semantics more cleanly (e.g., via probing tasks)?

3. **Embeddings space**: How does the "geometry" of the embedding space change under a curriculum?

By answering these questions, we gauge whether a progressive, human-like learning schedule provides tangible gains in both task performance *and* the organisation of internal representations.

Datasets Our experiments rely on two complementary Italian corpora, each fulfilling a distinct role.

- **Italian Wikipedia** *(pre-training & representation analysis)*
  - **Size**: 1,002,001 sentences extracted from a 2024 dump.
  - **Length filter**: 6–60 tokens per sentence.
  - **Splits**: 1,000,001 sentences for pre-training; 20,000 unseen sentences reserved for testing.
  - **Usage**: the training split feeds the Transformer during curriculum and baseline runs, while sentence embeddings (obtained by mean-pooling last-layer token vectors) are computed on the held-out test split for geometric analysis.
- **Universal Dependencies (UD) Italian Treebank** *(probing)*
  - **Size**: 8,000 sentences for training, 2,000 for evaluation.
  - **Annotations**: 136 fine-grained morphological, syntactic, and dependency labels.
  - **Linguistic phenomena probed**:
    1. Raw text properties
    2. Lexical variety
    3. Morphosyntactic information (POS)
    4. Verbal predicate structure
    5. Global and local parse-tree structures
    6. Syntactic relations
    7. Use of subordination

In summary, Italian Wikipedia provides the foundational data for model pretraining and serves as a basis for analyzing the resulting sentence-embedding space. In contrast, UD-IT offers richly annotated data to probe how well these embeddings capture a range of linguistic phenomena, from surface-level features to deep syntactic structures. Although both datasets fall within the same general domain, the probing data introduces distinct challenges due to substantial differences in topical composition and token-frequency distribution (Jensen–Shannon divergence).

Pretraining We pre-trained a BERT-Medium model (8 layers, 512 hidden units) on the Italian Wikipedia dataset composed of 1,002,001 sentences.. Training spanned 112,000 optimization steps across three epochs, and 37 checkpoints were saved, each capturing the model's parameters at a distinct stage of learning. Our analysis adopts two complementary perspectives: treating the network as (i) a *single model that evolves over time*, and (ii) a collection of *37 discrete models*, each reflecting a progressively higher level of linguistic competence.

`learning_rate`	`weight_decay`	`warmup_ratio`	`train_batch_size`	`eval_batch_size`
1×10^{-4}	0.01	0.01	64×2 GPUs	$128 \times$ 2 GPUs

Table 1: Pretraining hyperparameters.

Curriculum learning To isolate ordering effects during pre-training, we derive four linguistically motivated rankings of the Wikipedia corpus and pair each with its reverse (*anti-curriculum*).

- **Length**: sentences are sorted from 6 to 60 tokens, gradually introducing longer inputs.
- **Readability (Gulpease)**: the Italian-specific Gulpease score:

$$\text{Gulpease} = 89 + \frac{300\, N_{\text{sent}} - 10\, N_{\text{char}}}{N_{\text{word}}}$$

  orders the data from easy to difficult prose.
- **ReadIT**: we adopt the `ReadIT` framework [4], which combines lexical, morpho-syntactic, and dependency cues into a single difficulty score.
- **Random**: three independent shuffles act as non-curricular controls.

Each ranking is applied once in ascending order (curriculum) and once in descending order (anti-curriculum), allowing us to assess whether progressive exposure or its opposite influences convergence and representation quality.

2.1 RQ1: Downstream performance

Downstream evaluation. We fine-tune each checkpoint for 10 epochs on three complementary tasks:

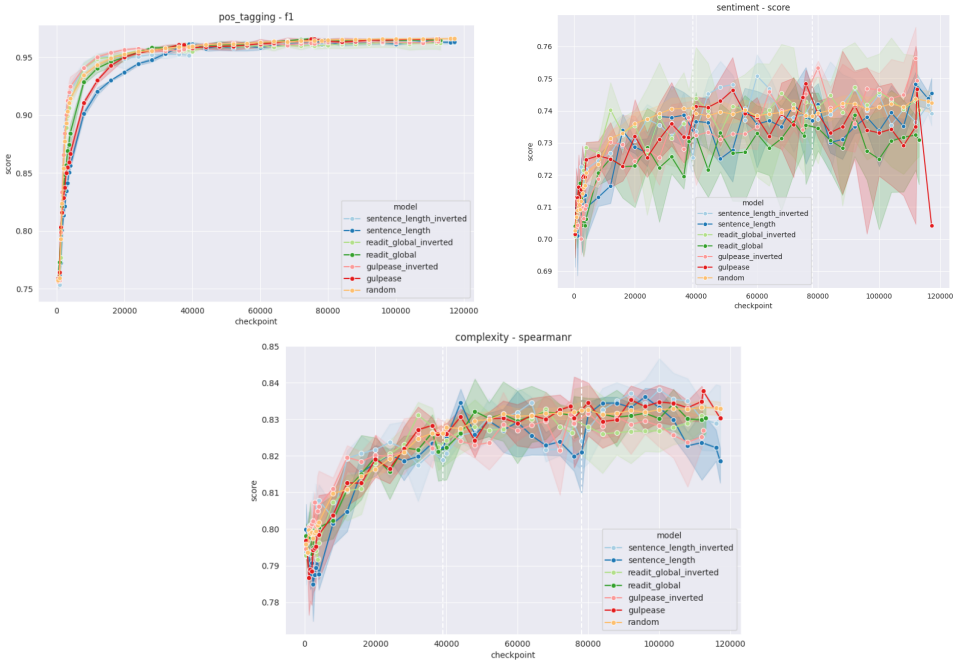

Figure 3: Downstream task scores on our chosen tasks.

- **Sentiment Analysis** (SST-2, 2 classes): tests semantic polarity and compositional meaning.
- **Text Complexity** (Human Complexity Judgement [2], regression): gauges perceived difficulty, sensitive to sentence length, lexical choice, and syntactic depth.
- **POS Tagging** (235 classes): probes low-level morpho-syntactic knowledge.

`learning_rate`	`weight_decay`	`warmup_ratio`	`train_batch_size`	`eval_batch_size`
5×10^{-5}	0.01	0.01	16×2 GPUs	16×2 GPUs

Table 2: Finetuning hyperparameters.

2.1.1 Results

Figure 3 shows how the models perform on our three downstream benchmarks:

POS tagging (F1). F_1 climbs from 0.76 to 0.95 in the first 8 k steps, after which all curves plateau at 0.962–0.963. Length-based curricula reach the knee of the curve a few checkpoints *later* than the others, indicating a slight slow-down in convergence. Once the model has ingested roughly one epoch of data, corpus ordering is immaterial for coarse morpho-syntactic tagging.

Sentiment analysis (Accuracy). Accuracy jumps from $\sim$0.70 to $\sim$0.74 within the first 15 k optimisation steps and then drifts sideways. The curriculum, anti-curriculum and random traces remain co-extensive throughout training; their largest instantaneous separation (< 0.007) is well inside the run-to-run uncertainty band.

Hence sentence-level polarity, driven mainly by lexical cues, appears indifferent to how the pre-training corpus is ordered.

Text-complexity ranking (Spearman ρ). Rank correlation with human difficulty judgments rises from 0.79 to $\sim$0.84 over the first two epochs and then levels off. All seven schedules oscillate within the same ± 0.003 band; no sustained curriculum advantage emerges.

In sum, pre-training order yields at most *marginal* differences on end-to-end tasks. For token-level classification (POS) and lexically dominated sentence-classification (sentiment) the effect is statistically negligible after a single pass through the corpus, and even a task that explicitly targets higher-order structure (complexity ranking) shows no robust gain.

2.2 RQ2: Linguistic Knowledge

Probing Since the comprehensive SentEval suite of Conneau et al. 2018 [3], probing has become the standard microscope for inspecting neural representations: a lightweight model is trained *post-hoc* on frozen embeddings, and its accuracy is taken as evidence that the probed feature is already encoded. To respect the "simple-probe" doctrine of Alain and Bengio 2018 [1], we employ a linear Support Vector Regressor (`LinearSVR`) with a linear kernel from SCIKIT-LEARN. Sentence representations obtained by mean-pooling each layer token embeddings are rescaled to the interval $[0, 1]$ to remove scale biases. We report the mean performance over 5-folds (80 % train, 20 % test).

2.2.1 Results

Figure 4 zooms in on six representative UD-IT probes, chosen to cover surface statistics (N_TOKENS, CHAR_PER_TOK), lexico–morphology (UPOS_DIST_DET), syntax (AVG_LINKS_LEN), and clause–level phenomena (SUBJ_PREC, VERBS_MOOD_DIST_INF). Each panel plots probe performance over the 37 checkpoints for the three curricula (warm colours), their reversals (cooler tones) and an average of the three random ordering (their behaviour was equivalent). We find that:

Detailed probe trajectories (Layer 8). Figure **??** shows six representative probing tasks at the final encoder layer, plotted over 37 pre-training checkpoints for each ordering strategy (four curricula, their reversals, and random). The panels illustrate both surface-level and higher-order linguistic features:

- **Rapid convergence across all tasks.** Every probe climbs sharply within the first 10–15 k steps and then settles into a narrow band, indicating that the majority of linearly decodable information is acquired early.
- **Surface probes exhibit curriculum-specific offsets.** In CHAR_PER_TOK, the Gulpease curriculum (solid red) consistently outperforms all others by $\sim$0.10 ρ, reflecting its explicit bias toward character-level readability. Conversely, the N_TOKENS probe shows that the Sentence-Length curriculum (solid blue) maintains a slight advantage ($\approx$ 1.00 vs. 0.96) on raw token counts, whereas the other runs converge near 0.96–0.97.

- **Morpho-syntactic tagging gains are modest.** For UPOS_DIST_DET, all runs reach $\approx$ 0.88–0.90 ρ. Curricula (solid warm colours) hover about 0.02 points above their inverses (dashed cool tones) and the random baseline.
- **Clause-level phenomena remain sensitive to scheduling.** In both SUBJ_PREC and VERBS_MOOD_DIST_INF, curricula (solid lines) plateau at ≈ 0.73, anti-curricula at ≈ 0.70, and random at ≈ 0.75. These ~ 0.02–0.05 differences indicate that ordered exposure aids modeling of long-distance functional relations.
- **Epoch–aligned perturbations.** Minor dips and rebounds at $\sim$ 40 k and 80 k steps appear across several tasks (notably SUBJ_PREC), mirroring (we will see) the saw-tooth fluctuations in intrinsic dimensionality and IsoScore. These transient effects underscore the coupling between data scheduling and representation geometry.

Overall, while superficial features (token/character counts) are dominated by their respective curricula, deeper syntactic and clause-level probes doesn't exhibit measurable advantages in using them.

3 RQ3: Embeddings space

Geometry of the embedding space. Transformer models embed tokens in R^d, but empirical studies show that the resulting vectors rarely fill the ambient space; instead they collapse into a narrow cone, a phenomenon known as *anisotropy* [6]. This collapse, often labelled *representation degradation*, is suspected of obscuring fine-grained distinctions, yet theory links it to the inductive bias of stochastic gradient descent and sometimes to improved generalisation [8]. To examine whether curriculum learning alters this trade-off, we characterise each checkpoint with two complementary descriptors: (i) an *isotropy coefficient* given by the ratio between the smallest and largest eigenvalues of the embedding covariance matrix, (ii) and the *intrinsic dimension*, defined as the number of principal components needed to explain 99 % of the variance.

Isotropy A representation is *perfectly isotropic* when its variance is spread evenly across the d embedding dimensions, equivalently, when the covariance matrix is proportional to the identity, $\mathrm{Cov}(X) = \sigma^2 \mathbf{I}_d$. In such a space no direction is privileged, so cosine similarity reflects genuine semantic alignment rather than geometric bias. We quantify deviations from this ideal with an IsoScore [9]. Given a batch matrix $X \in R^{n \times d}$, we perform PCA, obtain the variances $\{\lambda_i\}_{i=1}^{d}$ along the principal components, and normalise them to a probability vector $\mathbf{v} = \left(\lambda_i / \sum_j \lambda_j\right)$. Let $\mathbf{u} = \frac{1}{d}\mathbf{1}$ be the uniform distribution that characterises perfect isotropy. The score

$$\mathrm{IsoScore}(X) = 1 - \|\mathbf{v} - \mathbf{u}\|_2$$

lies in $[0, 1]$, attaining 1 for an isotropic cloud and decreasing as variance collapses onto a subset of dimensions.

Intrinsic Dimensionality Although sentences are encoded as high-dimensional vectors, the underlying data often live on a much lower-dimensional manifold. Compressing the representation onto this manifold is not only computationally advantageous but can also sharpen the inductive bias of learning algorithms. The minimal number of coordinates required to retain task-relevant information is termed the *intrinsic dimension* (ID). We approximate the ID of each embedding cloud with a linear proxy: the number of

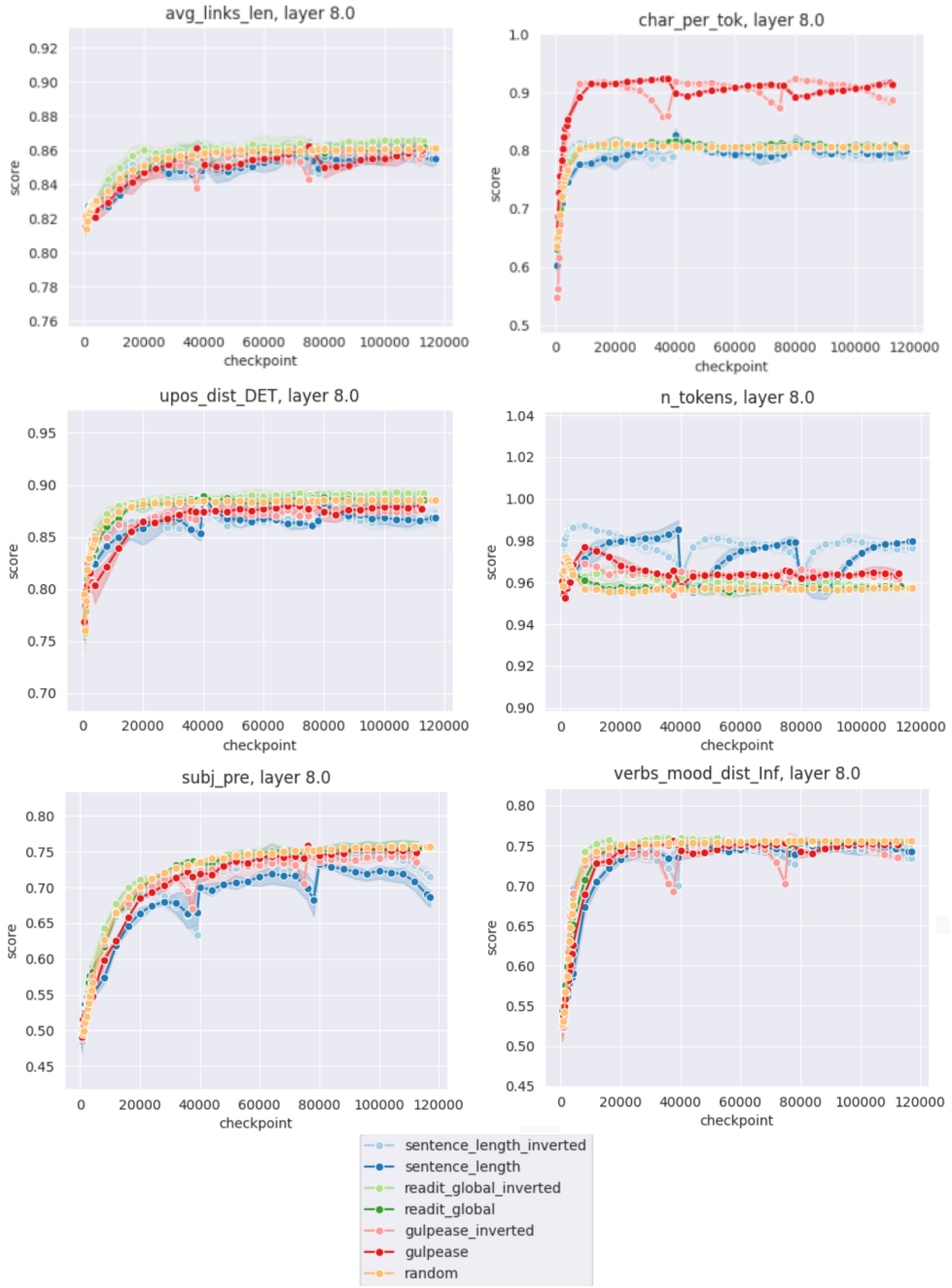

Figure 4: Some example plots of the probings performed, score used is Spearman ρ. Here all random models have been averaged.

principal components needed to account for 99 % of the variance. This estimate reflects the effective degrees of freedom used by the model; a smaller ID typically indicates a more compact and well-organised representation and has been shown to correlate with improved generalisation and interpretability.

3.0.1 Results

To probe how the geometry of the sentence embedding space evolves we performed an exploratory analysis on a fixed pool of the 20,000 held-out Wikipedia sentences from the pretraining test set.We fed each sentence through the network, collected the complete

stack of hidden states, and mean-pooled non-padding tokens. This yielded a tensor of shape (L, N, d) with L=8 layers, N=20 000 sentences, and d=768 dimensions.

Isotropy across training. Figure 5 complements the ID analysis with the *IsoScore*, a normalised measure of variance dispersion that equals 1 for perfectly isotropic clouds and 0 when variance is concentrated along a few axes [6]. Random ordering promotes a steady increase in isotropy, plateauing at 0.08–0.13 in the top layer. In stark contrast, every curriculum variant remains highly anisotropic, rarely exceeding 0.02 even after three epochs. IsoScore curves reproduce the same epoch-aligned saw-tooth pattern: replaying easy sentences briefly inflates isotropy, which then collapses again as sentence complexity ramps up.

Intrinsic dimensionality across training. Figure 6 plots the evolution of the *linear intrinsic dimensionality* (ID) of BERT-Medium at three depths (layers 1, 4, 8) over the 37 checkpoints that constitute the full pre-training run. All conditions exhibit the two-phase dynamics reported by [8]: an early "compression" stage during the first $\sim$5 k optimisation steps, followed by a slower expansion or stabilisation. The extent of this compression, however, is highly sensitive to data ordering. Models trained on randomly shuffled corpora rebound to values close to the ambient dimensionality of the hidden layer (e.g. $\mathrm{ID}_{\mathrm{L8}} \approx 388$ at the final checkpoint), whereas curriculum and anti-curriculum variants remain restricted to a substantially lower-rank manifold throughout training ($\mathrm{ID}_{\mathrm{L8}} \approx 310$, a $\sim$20% reduction). The trajectories display a distinctive saw-tooth profile: peaks occur at the start of every epoch when the data stream rewinds, and valleys emerge as sentence complexity increases. The separation between ordering regimes is greatest in the mid-stack (Layer 4), where syntactic abstractions are believed to crystallise [7]; here the curriculum runs stay up to 150 ID units below the shuffled baseline.

The tight coupling between ID and IsoScore trajectories (Spearman $\rho > 0.8$ across checkpoints) shows that curriculum learning simultaneously *compresses* the sub-space and *funnels* variance into a small set of privileged directions, producing a narrow, low-rank cone of activations. These geometric diagnostics reveal that ordering the training corpus by sentence complexity imposes a persistent structural bias on the representation space: the model operates with fewer effective degrees of freedom and allocates variance heavily along a limited set of axes. Such compactness can aid linear probes but it may also heighten vulnerability to domain shift and catastrophic interference during downstream fine-tuning.

3.1 Unsupervised analysis

As before, after extracting sentence representations, for every checkpoint (but last layer only) we:

1. **Reduced dimensionality with PCA.** We computed the full principal–component spectrum of the $N \times d$ matrix and used the *KneeLocator* algorithm to detect the inflection point of the cumulative explained variance ("scree plot"). The resulting dimension ranged from 25 to 60 across checkpoints and served as an information-preserving subspace for clustering.
2. **Discovered clusters with K-means.** On the PCA scores we ran K-means for $k \in \{2, \ldots, 15\}$ and plotted the inertia curve (elbow plot). The elbow was again

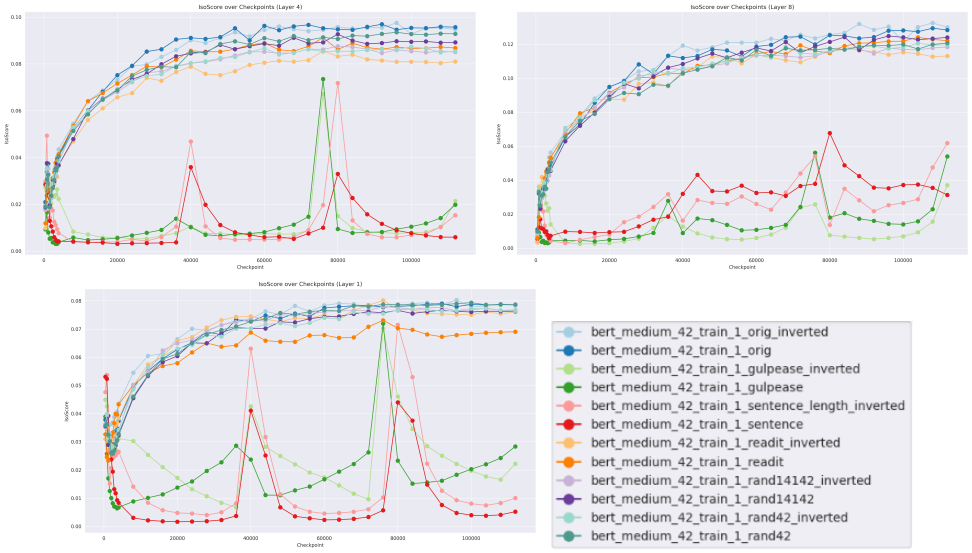

Figure 5: IsoScore for layers 1, 4, and 8.

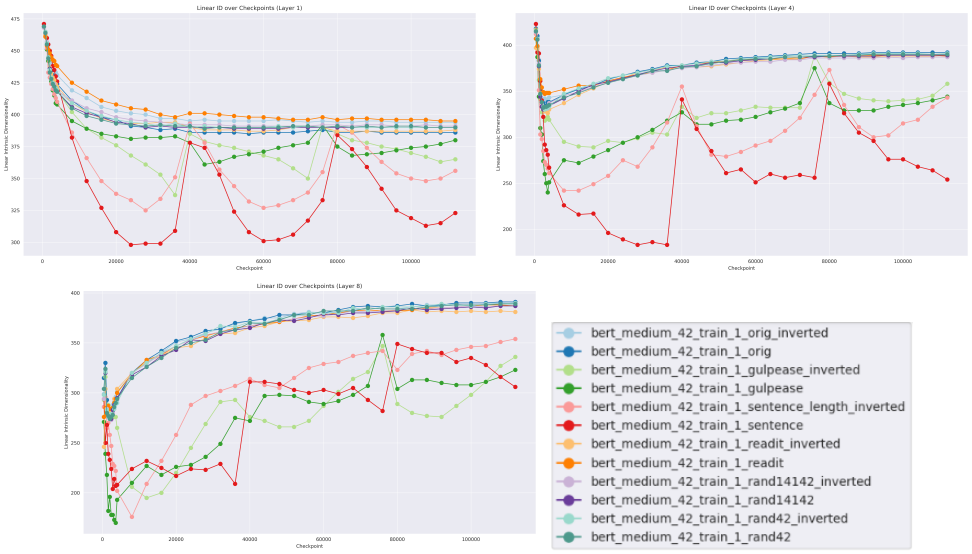

Figure 6: Linear ID for layers 1, 4, and 8.

localised with KNEED. The chosen k typically fluctuated between 6 and 12 as training progressed.

3. **Visualised trajectories via 3D t-SNE.** For qualitative insight we embedded the same PCA scores into R^3 with t-SNE (perplexity $= 5$, 1 000 iterations) and colour-coded points by their K-means label. To reveal temporal regularities, we linearly interpolated successive t-SNE frames and compiled an animated GIF that shows how clusters merge, split, and drift over the 37 checkpoints. Figure 8 shows the 2-D t-SNE at the first and last checkpoints for our three curricula and one random seed.
4. **Related clusters to surface features.** We asked whether cluster membership predicts sentence length, a crude proxy for syntactic complexity. For each checkpoint we recorded R^2, the proportion of length variance explained by the clustering.

3.1.1 Results

t-SNE clustering snapshots. Figure 8 presents 2D t-SNE projections of the final-layer embeddings at an early checkpoint (400 steps, left) and a late checkpoint (11 200 steps, right) for four ordering regimes: Sentence-Length (top row), Gulpease (second row), ReadIT (third row) and a Random shuffle (bottom row). At 400 steps all models exhibit fairly well-separated clusters. Sentence-Length displays clear radial bands corresponding to token count, while Gulpease and ReadIT each form tighter, semantics-driven groupings. By 11 200 steps the embedding clouds have expanded and overlapped, but the curriculum-trained models (rows 2–3) still retain pockets of colour coherence, indicating that the difficulty-driven ordering imprints a more durable cluster structure.

Length-dependency of the latent clusters under different curricula. Figure 9 reports, for four ordering regimes, `Gulpease` (upper-left), `ReadIT` (upper-centre), `Sentence-Length` (upper-right) and a random shuffle (lower-left), how much of the variance in *sentence length* can be explained (R^2) by the K-means clusters extracted from the representation space of the final Transformer layer.

- **Gulpease curriculum.** The readability-based schedule begins with a strong length imprint ($R^2 \approx 0.80$) but this influence decays rapidly after the first 10 k updates and stabilises below 0.30. The curriculum therefore drives the model to organise sentences along syntactic or lexical cues more quickly than along raw length.
- **ReadIT curriculum.** A similar pattern holds, albeit more abruptly: after an initial spike the length signal collapses to $R^2 < 0.25$ and never recovers, indicating that the richer ReadIT difficulty metric steers clustering away from superficial size even earlier in training.
- **Sentence-length curriculum.** Here the correlation remains high for most of training, oscillating between 0.60 and 0.80 with clear saw-tooth resets at epoch restarts (vertical dashed lines). Because sentence length is itself the ordering principle, clusters keep reflecting that dimension; the network therefore retains a "length axis" well into the later checkpoints.
- **Random ordering.** After a brief compression phase R^2 settles in the 0.20–0.35 range, with minor epoch-aligned bumps. With no systematic exposure schedule, length exerts only a modest and transient pull on the representation geometry.

Overall, the figure shows that *how* a curriculum is defined dictates whether sentence length remains a primary organising factor in the latent space: schedules based on lexical–syntactic difficulty (Gulpease, ReadIT) quickly down-weight length, whereas an explicit length curriculum preserves it, and random order falls in between.

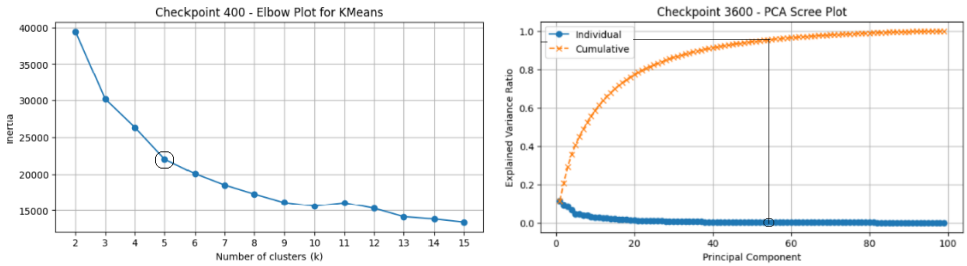

Figure 7: Example elbow & scree plot for two of our models's last layer.

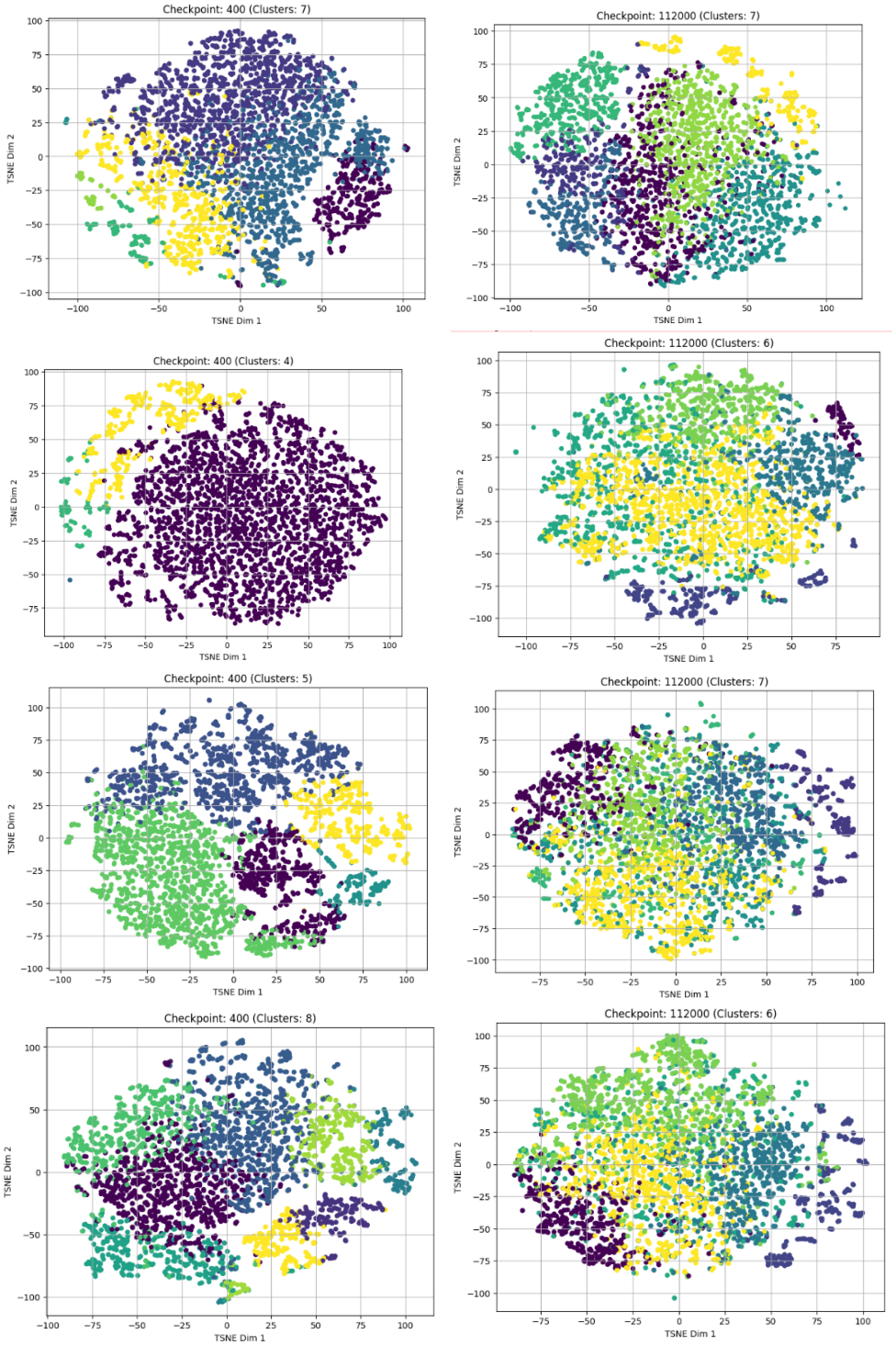

Figure 8: From top to bottom: cluster from the reprepresentations extracted from model Sentence Length, Gulpease, ReadIT and one of the random seeds.

4 Conclusion

Curriculum learning leaves a much deeper footprint on *how* a transformer encodes language than on *what* it ultimately achieves on standard benchmarks. Ordering the Italian Wikipedia corpus from easy to hard (according to length, Gulpease or ReadIT scores) compresses the representation manifold (ID $\downarrow 20$ %) and drives variance into a narrow

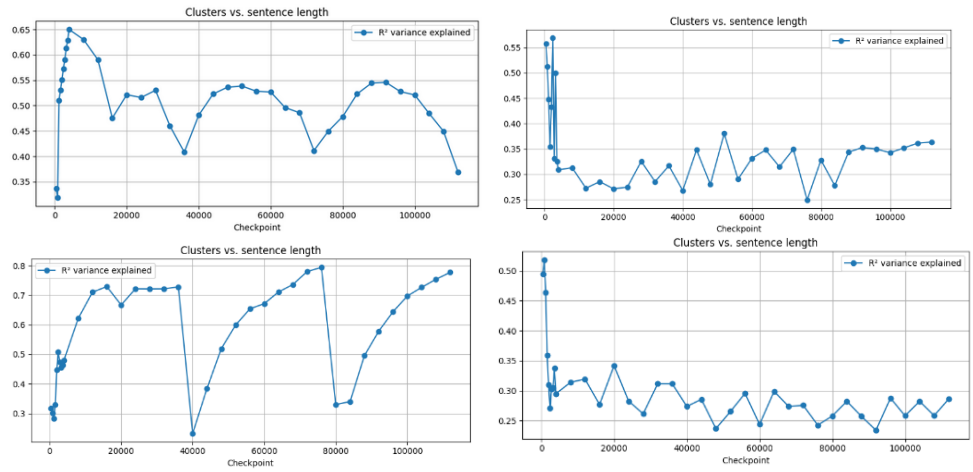

Figure 9: R^2 scores on Gulpease, ReadIT, Sentence Length and one of the random seeds.

set of axes (IsoScore ≤ 0.02), yet translates into only marginal gains on end-to-end tasks and surface-level probes. The same curricula, however, *do* make syntactic and clause-level information more linearly extractable and accelerate its emergence in early training. Taken together, the results indicate that progressive exposure shapes the geometry of the embedding space in a way that is favourable for diagnostic analysis without necessarily improving headline accuracy. We therefore argue that representation-centric metrics should accompany task scores whenever the efficacy of a curriculum is assessed.

5 Future work

- **Scale.** Repeat the study with BERT-Base and BERT-Large to test whether geometric effects persist at larger capacity and whether they interact with deeper layer hierarchies.
- **Multilingual curricula.** Apply the same ordering strategies to multilingual corpora to examine whether language typology modulates the impact of curriculum learning.
- **Word-level geometry.** Extend the isotropy and ID analysis to token and subtoken representations to see if compression propagates upward from lexical to sentential space.
- **Non-linear diagnostics.** Incorporate manifold learning (e.g. UMAP, diffusion maps) and topological data analysis to capture structure that linear PCA may miss.
- **Temporal dynamics.** Use functional data analysis or recurrent probing to model how individual sentences migrate through the representation space during training.
- **Continual-learning robustness.** Fine-tune curriculum- and randomly-trained checkpoints on a sequence of tasks to quantify resistance to catastrophic forgetting; the compressed cones we observe may be more brittle.
- **Compression for deployment.** Investigate whether the low-rank structure induced by curricula makes pruning or low-precision quantisation more effective without extra distillation steps.

References

[1] Guillaume Alain and Yoshua Bengio. *Understanding intermediate layers using linear classifier probes*. 2018. arXiv: 1610.01644 [stat.ML]. URL: https://arxiv.org/abs/1610.01644.

[2] Dominique Brunato et al. "Is this Sentence Difficult? Do you Agree?" In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2690–2699. DOI: 10.18653/v1/D18-1289. URL: https://aclanthology.org/D18-1289/.

[3] Alexis Conneau et al. "What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2126–2136. DOI: 10.18653/v1/P18-1198. URL: https://aclanthology.org/P18-1198/.

[4] Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. "READ–IT: Assessing Readability of Italian Texts with a View to Text Simplification". In: *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*. Ed. by Norman Alm. Edinburgh, Scotland, UK: Association for Computational Linguistics, July 2011, pp. 73–83. URL: https://aclanthology.org/W11-2308/.

[5] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL]. URL: https://arxiv.org/abs/1810.04805.

[6] Kawin Ethayarajh. "How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 55–65. DOI: 10.18653/v1/D19-1006. URL: https://aclanthology.org/D19-1006/.

[7] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. "What Does BERT Learn about the Structure of Language?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3651–3657. DOI: 10.18653/v1/P19-1356. URL: https://aclanthology.org/P19-1356/.

[8] Vardan Papyan, X. Y. Han, and David L. Donoho. "Prevalence of neural collapse during the terminal phase of deep learning training". In: *Proceedings of the National Academy of Sciences* 117.40 (Sept. 2020), pp. 24652–24663. ISSN: 1091-6490. DOI: 10.1073/pnas.2015509117. URL: http://dx.doi.org/10.1073/pnas.2015509117.

[9] William Rudman et al. "IsoScore: Measuring the Uniformity of Embedding Space Utilization". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022, pp. 3325–3339. DOI: 10.18653/v1/2022.findings-acl.262. URL: http://dx.doi.org/10.18653/v1/2022.findings-acl.262.

[10] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.