Shaping Representation Space: Geometric Diagnostics in Transformers

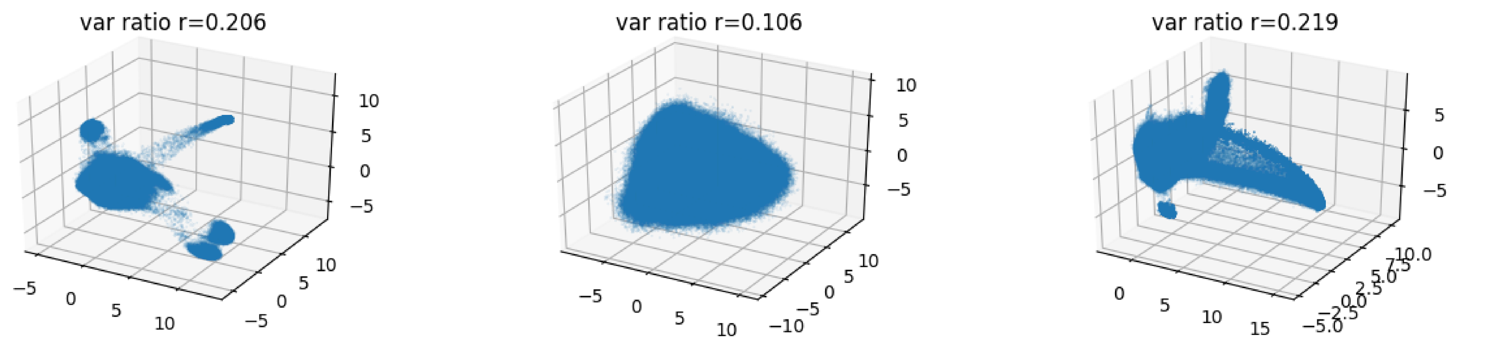
Lucia Domenichelli[1,2]

[1]ItaliaNLP Lab @ CNR-ILC Pisa

[2]University of Pisa

Introduction to the representation space of LLMs

- **High-dimensional data.** Embeddings form curved, twisted hypersurfaces with loops, pockets and bottlenecks.
- **Narrow-Cone Hypothesis** (Ethayarajh, 2019). Embeddings occupy a tight cone—highly anisotropic, not uniformly spread.
- **Manifold Hypothesis** (Clayton, 2015). Although embedded in high D, data lie on a much lower-dimensional manifold.



Research questions

(i) **Linear-ID (99%)**: $\mathrm{LinearID}_{0.99} = \min\{k : \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{j=1}^{D}\lambda_j} \geq 0.99\}$

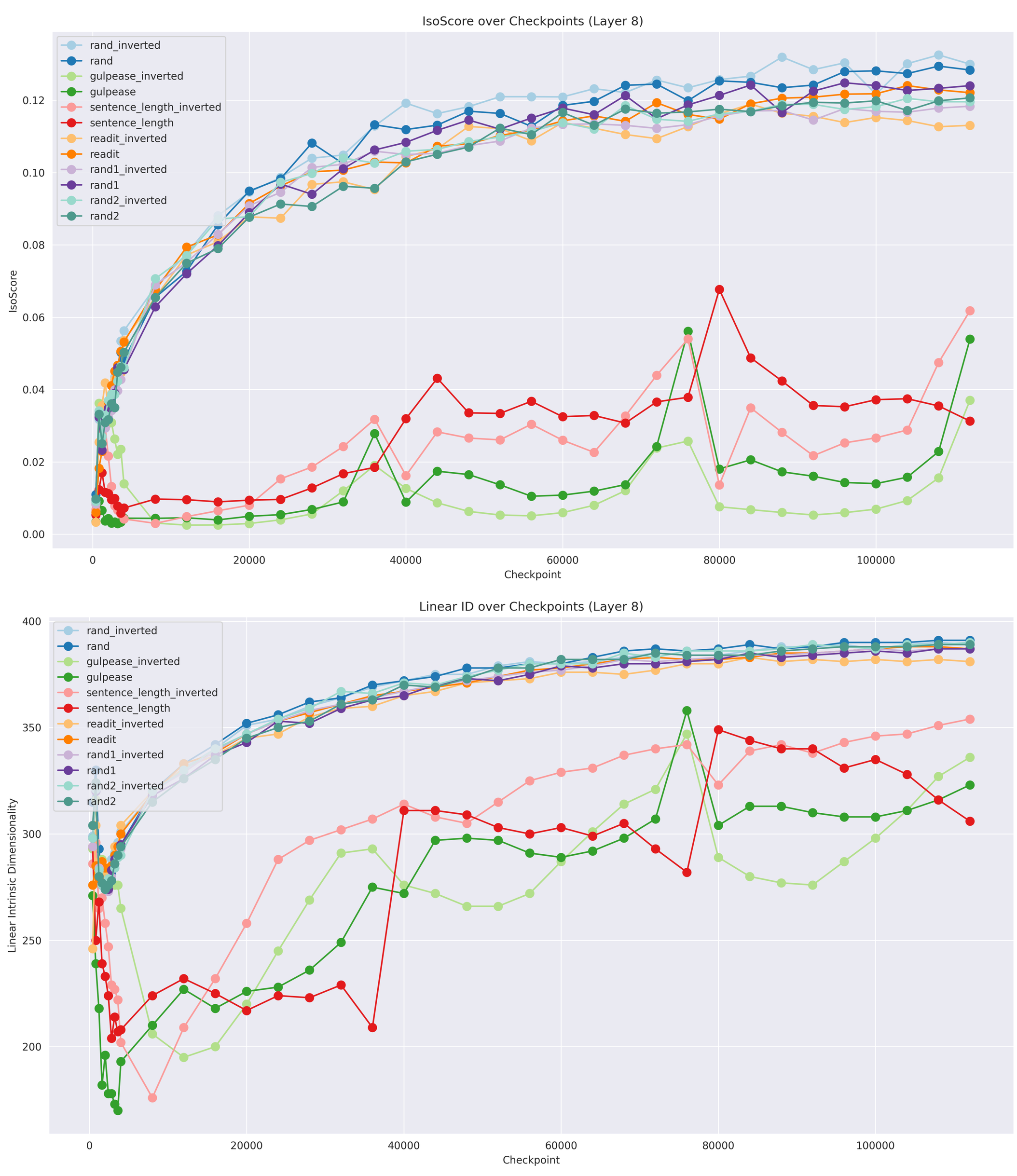
(ii) **IsoScore**: $\tilde{\lambda}_i = \lambda_i \frac{\sqrt{D}}{\|\boldsymbol{\lambda}\|_2}, \quad \mathrm{IsoScore} = 1 - \frac{\|\tilde{\boldsymbol{\lambda}} - \mathbf{1}\|_2}{\sqrt{2\,(D - \sqrt{D})}}$

Eigenvalues $\lambda_1 \geq \ldots \geq \lambda_D$ are taken from the covariance of centred embeddings.

What happens during pre-training?

Curriculum learning

Over checkpoints



Over layers

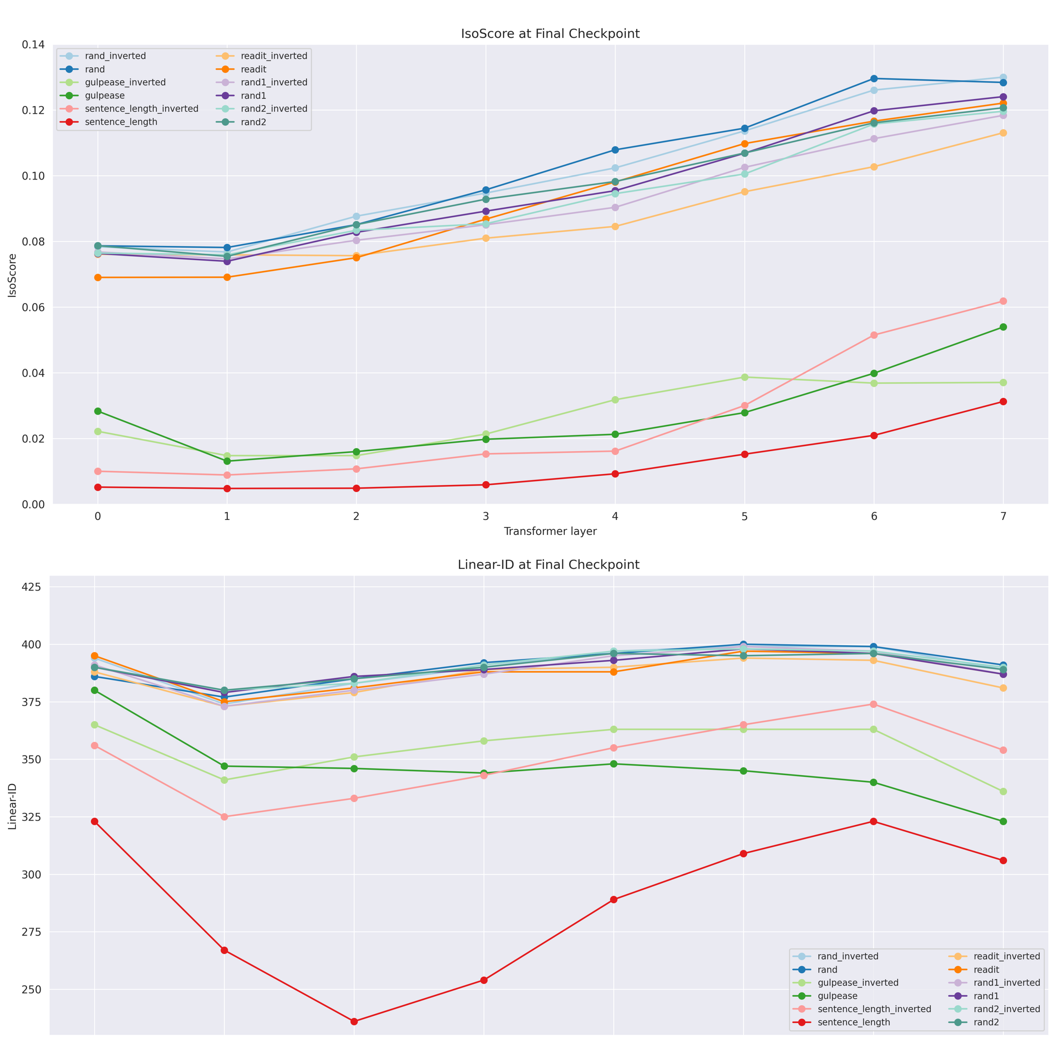


Figure 2: IsoScore and Linear-ID across checkpoint (top) and at the final checkpoint (bottom) for BERT-medium.

What happens after fine-tuning?

Eye-tracking

English

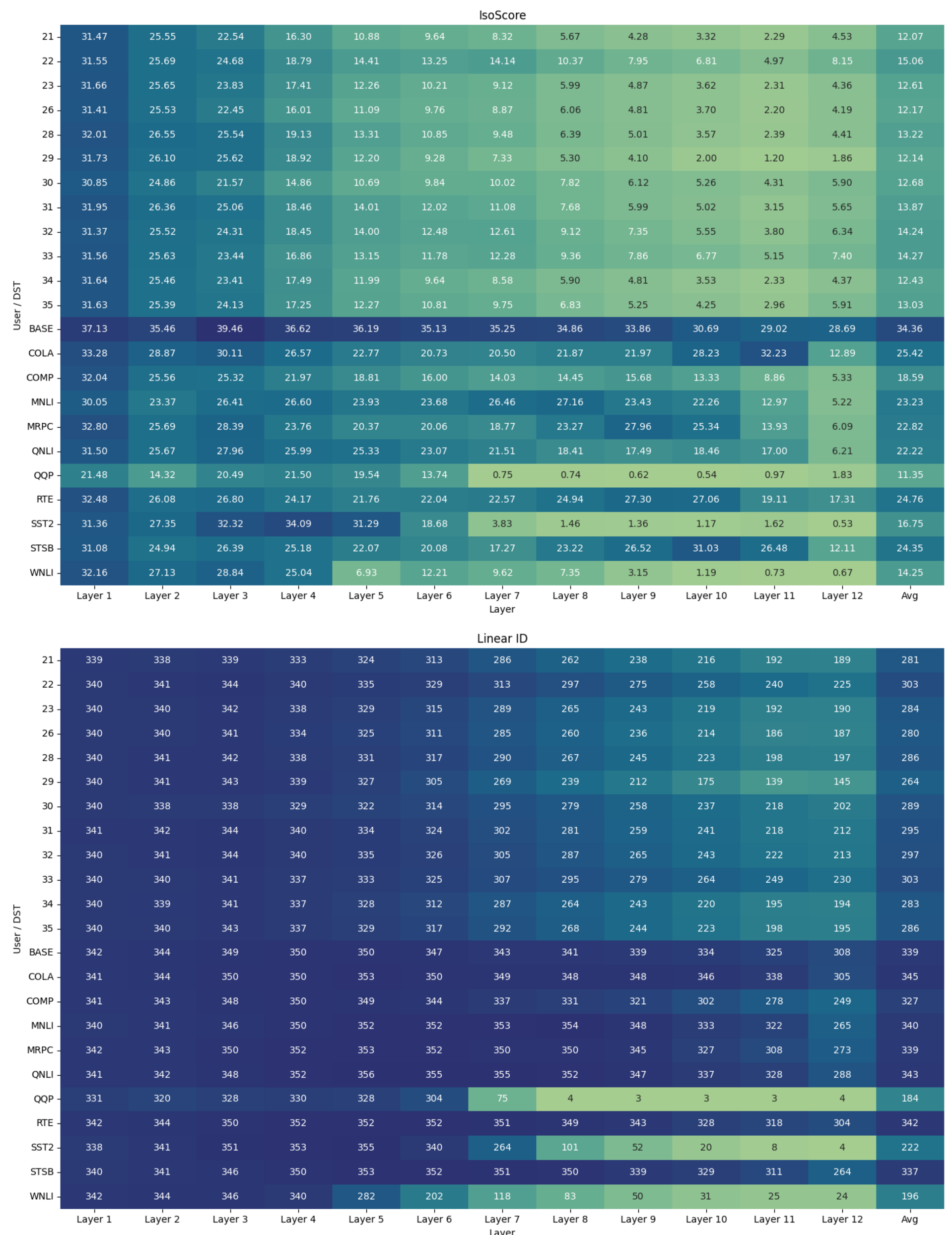


Figure 3: IsoScore and Linear-ID across layers on diverse fine-tuning tasks.

Italian

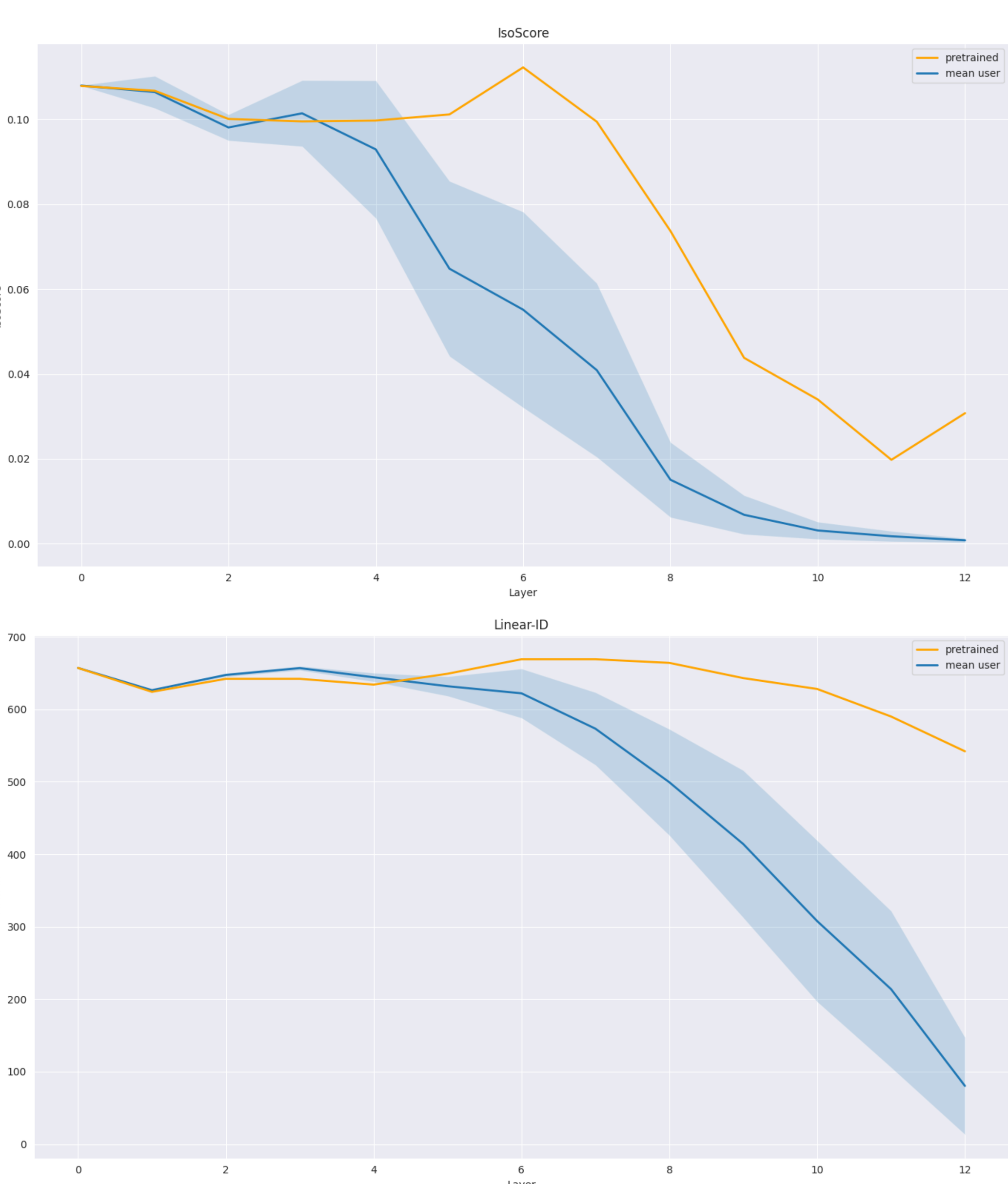
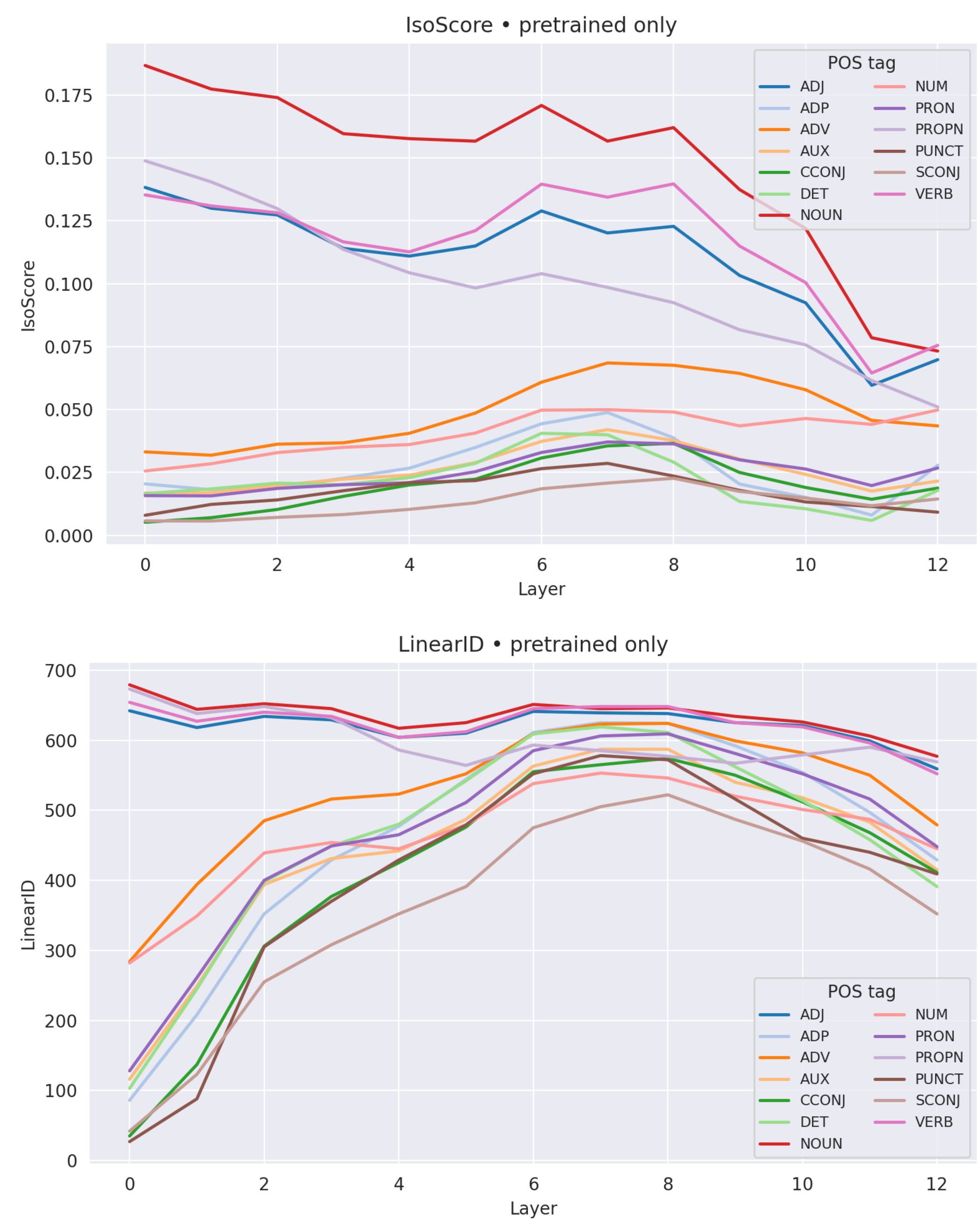


Figure 4: IsoScore and Linear-ID across layers for baseline XLM-RoBERTa and eye-tracking-fine-tuned model.

What happens to sub-spaces?

Part-of-speech



Head distance

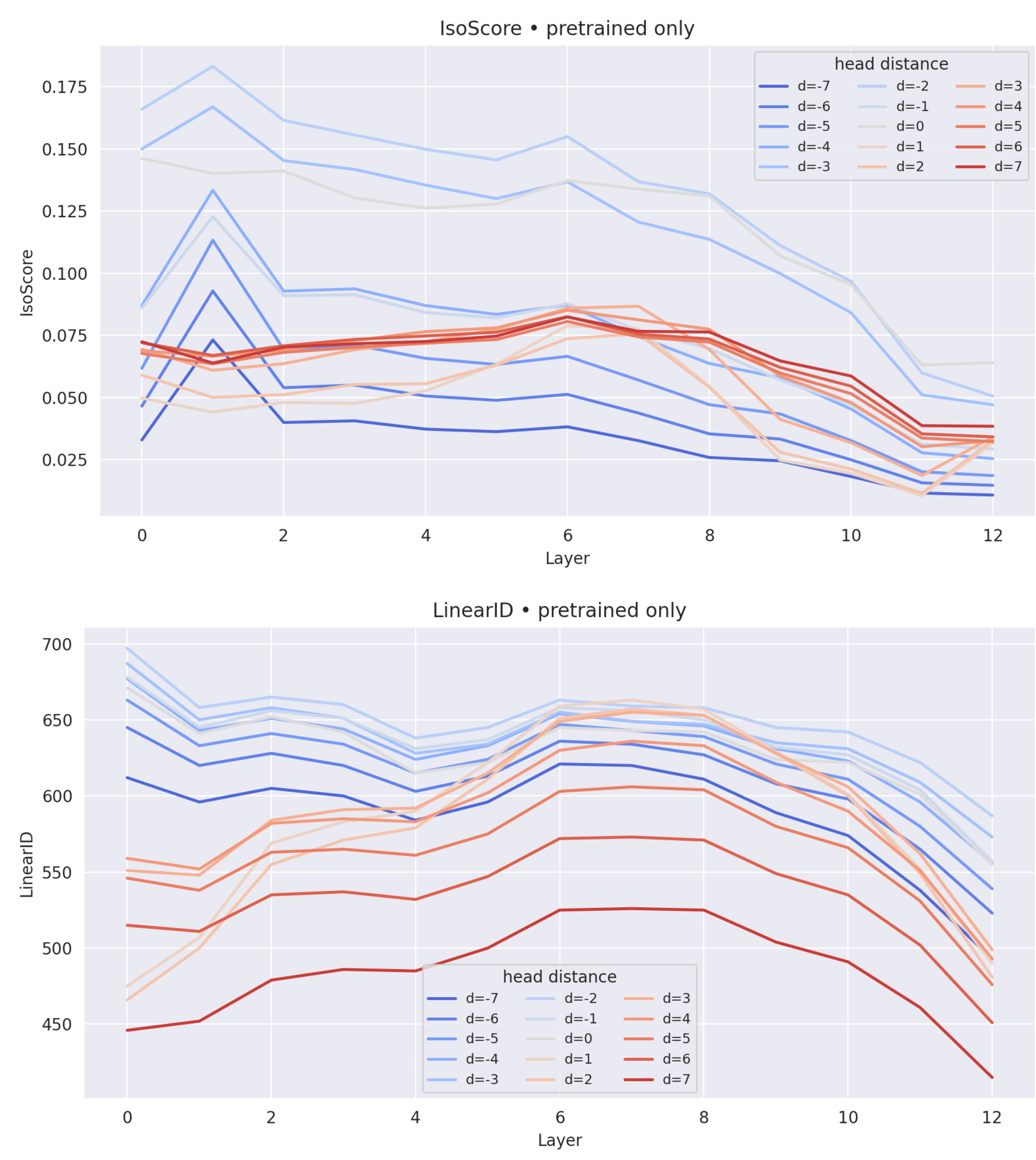


Figure 5: IsoScore and Linear-ID across layers for POS and head–dependent distance subspaces.

Lectures on Computational Linguistics 2025, 18–20 June 2025, Milan, Italy