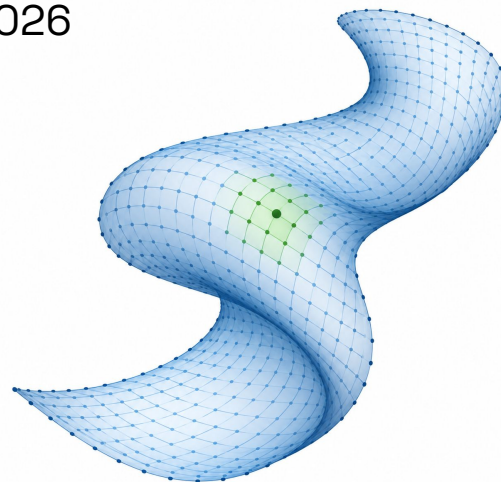


From Global to Local

Intrinsic Dimensionality as a lens on
Language Model Representation

Officine ItaliaNLP - 29 September 2026

Lucia Domenichelli + gpt-6
ItaliaNLP Lab, CNR-ILC



Talk outline

1. **Representation geometry in language models**

2. Intrinsic dimensionality and its estimators

3. From global to local: a lens on linguistic structure

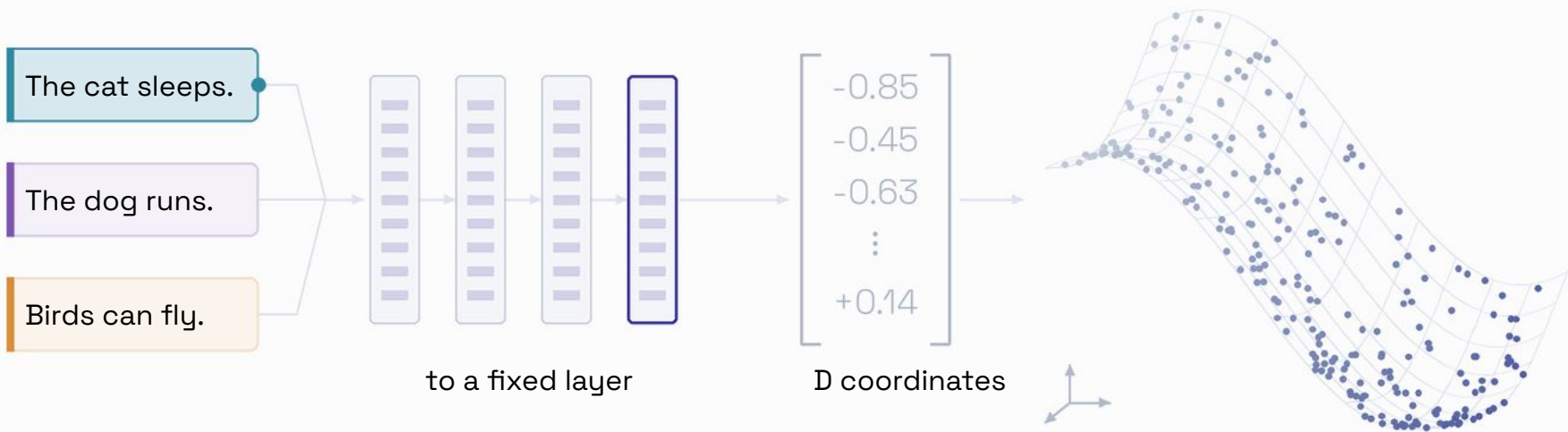
Representations as data in high dimensions

Text inputs

Model layers

Hidden state

Representation cloud



$$\mathbf{h}_\ell(x_i; \theta) \in \mathbb{R}^D, \quad \mathcal{H}_\ell = \left\{ \mathbf{h}_\ell(x_i; \theta) \right\}_{i=1}^n$$

Token and sentence representations

Sentence-BERT: Sentence Embeddings
using Siamese BERT-Networks
Reimers & Gurevych (2019) · EMNLP-IJCNLP

SimCSE: Simple Contrastive Learning
of Sentence Embeddings
Gao, Yao & Chen (2021) · EMNLP

How Contextual are Contextualized
Word Representations?
Ethayarajh (2019) · EMNLP-IJCNLP

SBERT-WK: A Sentence Embedding Method
by Dissecting BERT-based Word Models
Wang & Kuo (2020)

On the Sentence Embeddings from
Pre-trained Language Models
Li et al. (2020) · EMNLP

Ditto: A Simple and Efficient Approach
to Improve Sentence Embeddings
Chen et al. (2023) · EMNLP

Representation biases in
sentence transformers
Nikolaev & Padó (2023) · EACL

A Simple but Tough-to-Beat Baseline
for Sentence Embeddings
Arora, Liang & Ma (2017) · ICLR

Where Do LLMs Compose Meaning?
A Layerwise Analysis of
Compositional Robustness
Aljaafari, Carvalho & Freitas (2026) · EACL

What makes NN good learners?

Neural networks adapt to **structure** in the data

**Spectral statistics
of the weights**



**Representation
manifold**

Learning reshapes weights and the representations they produce.

Wang et al. (NeurIPS 2023) · Dandi et al. (AISTATS 2025)

How weights and representations connect

Input-data covariance shapes hidden-state covariance.

$$h_i = \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i)$$

Local approximation of the nonlinear model

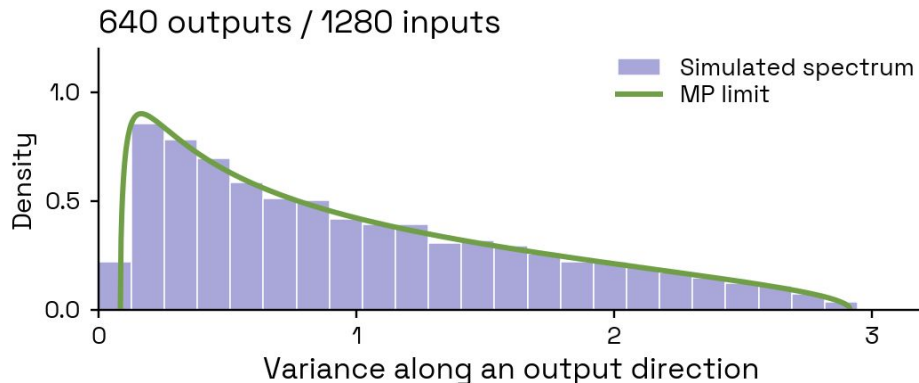
$$\Sigma_h \approx J \Sigma_{\text{data}} J^\top$$

Input-data covariance

J combines the local effects of weights and activations.

How weights and representations connect

Marchenko–Pastur theorem



$$W_{ij} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, d^{-1})$$
$$m, d \rightarrow \infty, \quad \frac{m}{d} \rightarrow \gamma \in (0, \infty)$$

Eigenvalue distribution converges to the MP law.

A broad spectrum can arise from randomness alone.

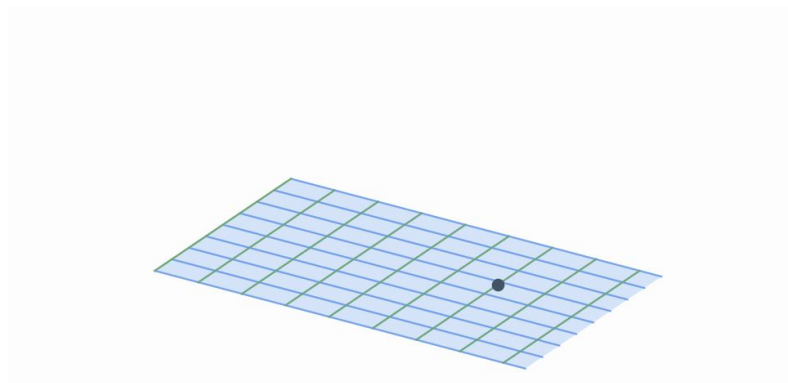
The manifold hypothesis

The data distribution concentrates on or near a smooth manifold with far fewer dimensions than the surrounding space.

Local degrees of freedom

$$\mathcal{M} \subset \mathbb{R}^D, \quad \dim \mathcal{M} = d \ll D$$

A small neighbourhood can be described by a few independent coordinates.



Goodfellow et al. (2016), §15.5 · Fefferman et al. (2016)

Why NOT a manifold?

Representations may concentrate near a union of manifolds, with different shapes or dimensions.

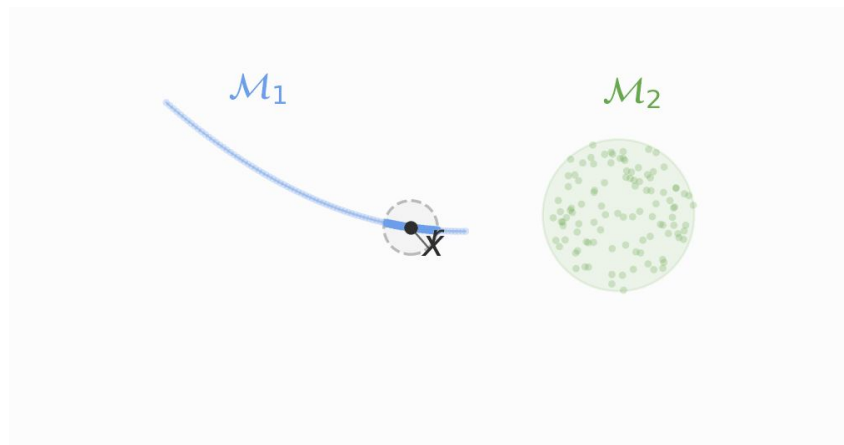
A union of components

Components can have different local dimensions.

One global ID may hide it.

$$\mathcal{S} = \bigcup_{c=1}^C \mathcal{M}_c$$

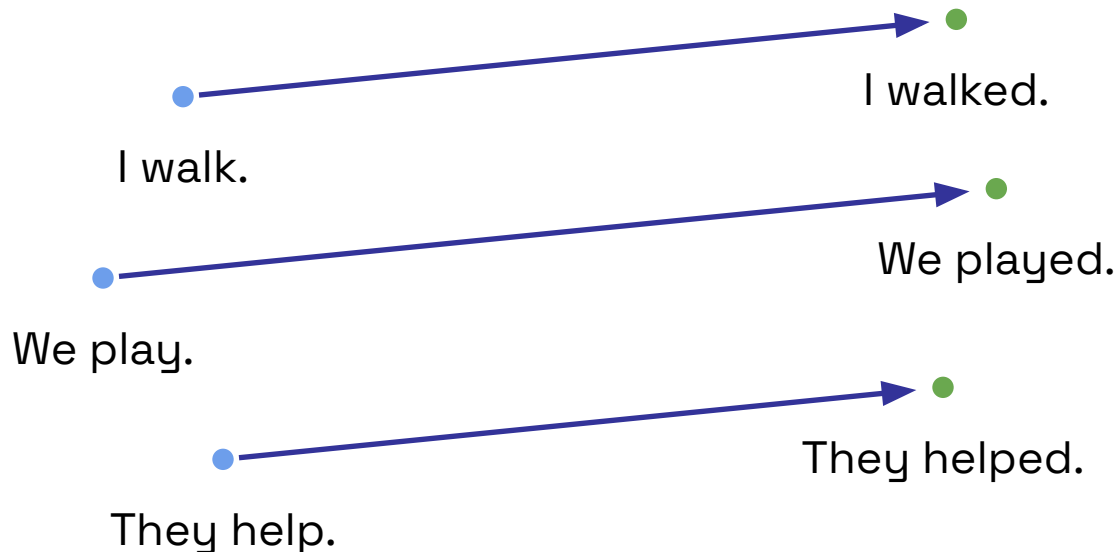
$$d_c = \dim \mathcal{M}_c$$



The linear representation hypothesis

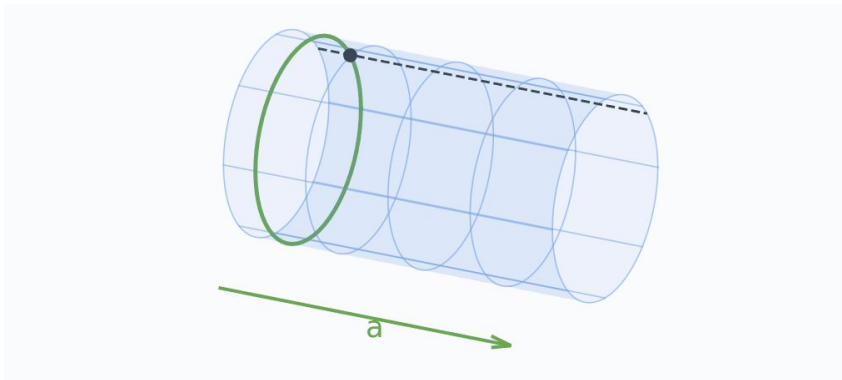
Concepts are encoded as directions in representation space.

Example: present to past



The two hypotheses can coexist

Curved support, two coordinates



$$\mathbf{h}(a, \theta) = (a, \cos \theta, \sin \theta, 0, \dots, 0)$$

$$\dim \mathcal{M} = 2 \ll D$$

One fixed feature direction

Read the feature with a linear map.

$$a = \mathbf{e}_1^\top \mathbf{h}$$

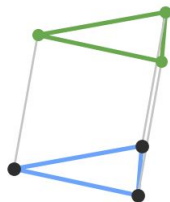
$$\mathbf{h}(a + \alpha, \theta) - \mathbf{h}(a, \theta) = \alpha \mathbf{e}_1$$

Shift along a fixed direction;
the angle stays unchanged.

Neither hypothesis proves the other.

Recurring structures in language models

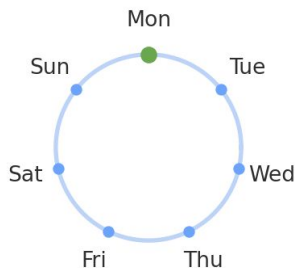
Prisms



Semantic
contrasts

Park et al. (2024)

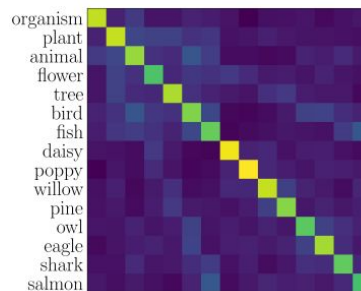
Loops



Days and months

Engels et al. (2025)

Hierarchies



Word
co-occurrence
Nava & Wyart (2026)

Talk outline

1. Representation geometry in language models

2. **Intrinsic dimensionality and its estimators**

3. From global to local: a lens on linguistic structure

What can we study about the geometry?

Intrinsic Dimension

Degrees
of freedom

Johnsson et al. (2015)
Erba et al. (2019)
Denti et al. (2022)
Cheng et al. (2025)

Curvature

Tangent planes
change orientation

Aamari & Levrard (2019)
Hosseini & Fedorenko (2023)

Topology

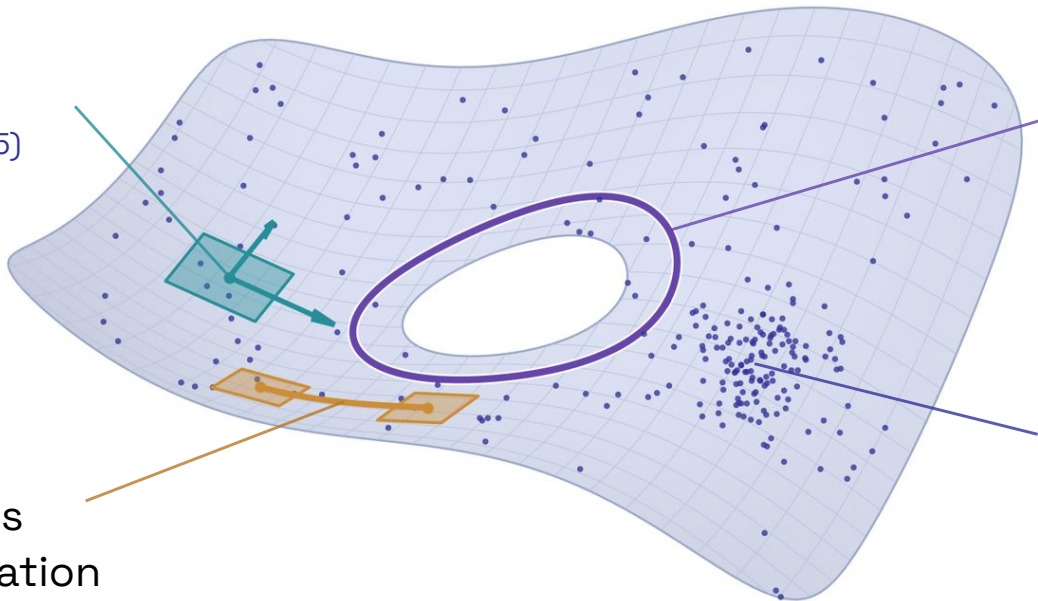
A loop that
cannot
shrink to a point

Edelsbrunner et al. (2002)
Gardinazzi et al. (2025)
Malhotra et al. (2026)

Sampling density

More samples in
some regions

Rodriguez et al. (2018)
Valeriani et al. (2023)



What can we study about the geometry?

Intrinsic Dimension

Degrees
of freedom

Johnsson et al. (2015)
Erba et al. (2019)
Denti et al. (2022)
Cheng et al. (2025)

Curvature

Tangent planes
change orientation

Aamari & Levrard (2019)
Hosseini & Fedorenko (2023)

Topology

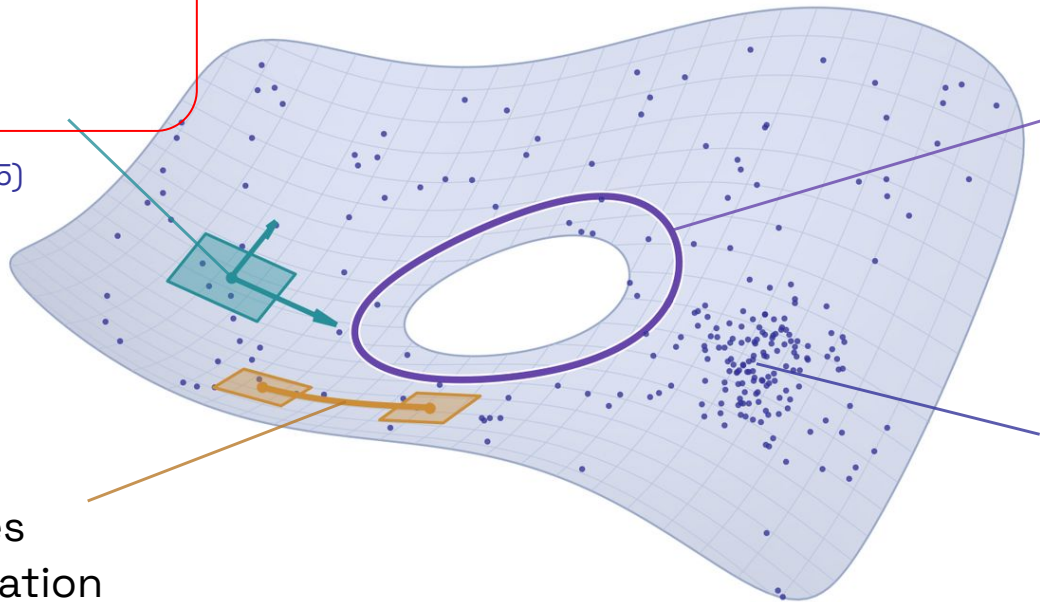
A loop that
cannot
shrink to a point

Edelsbrunner et al. (2002)
Gardinazzi et al. (2025)
Malhotra et al. (2026)

Sampling density

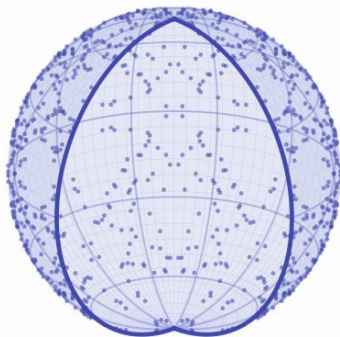
More samples in
some regions

Rodriguez et al. (2018)
Valeriani et al. (2023)



Why use intrinsic dimensionality?

Sphere surface (ID 2)



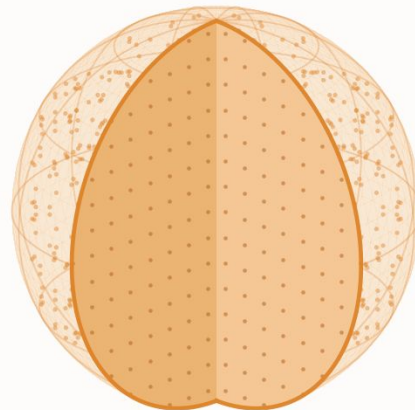
$$\|x\| = \sqrt{3}$$

Two local degrees of freedom.

Same global
covariance
 $\Sigma = I_3$

Covariance rank 3

Filled ball (ID 3)



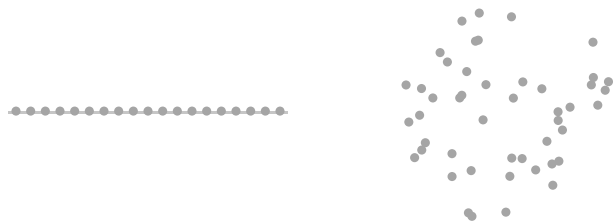
$$\|x\| \leq \sqrt{5}$$

Three local degrees of freedom.

From global to pointwise ID

Global ID

One summary for a dataset or layer



$$\mathcal{X} \mapsto \hat{d}_{\text{global}}$$

Pointwise ID

One local estimate per representation

$$d(\mathbf{x}_a) = 1$$

$$d(\mathbf{x}_b) = 2$$



$$\{\mathbf{x}_i \mapsto \hat{d}(\mathbf{x}_i; k)\}_{i=1}^n$$

k specifies the number of neighbours.

From global to pointwise ID

Global ID

One summary for a dataset or layer



$$\mathcal{X} \mapsto \hat{d}_{\text{global}}$$

Pointwise ID

One local estimate per representation

$$d(\mathbf{x}_a) = 1$$

$$d(\mathbf{x}_b) = 2$$

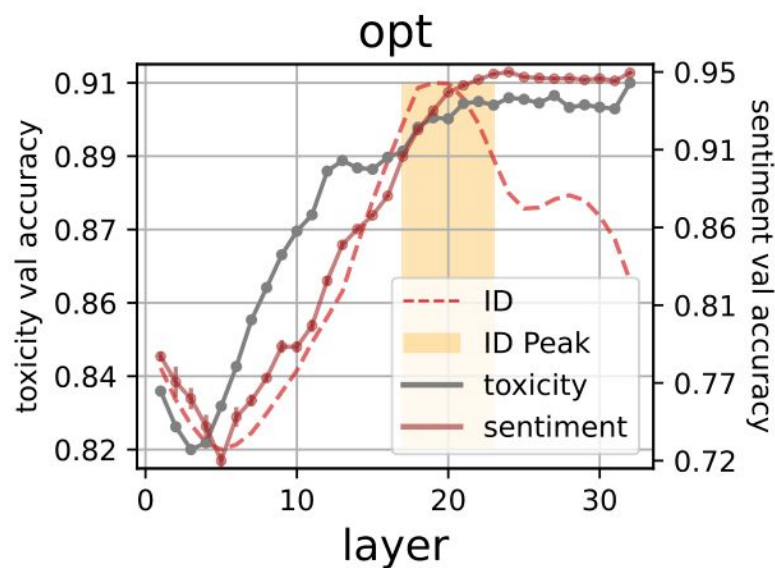
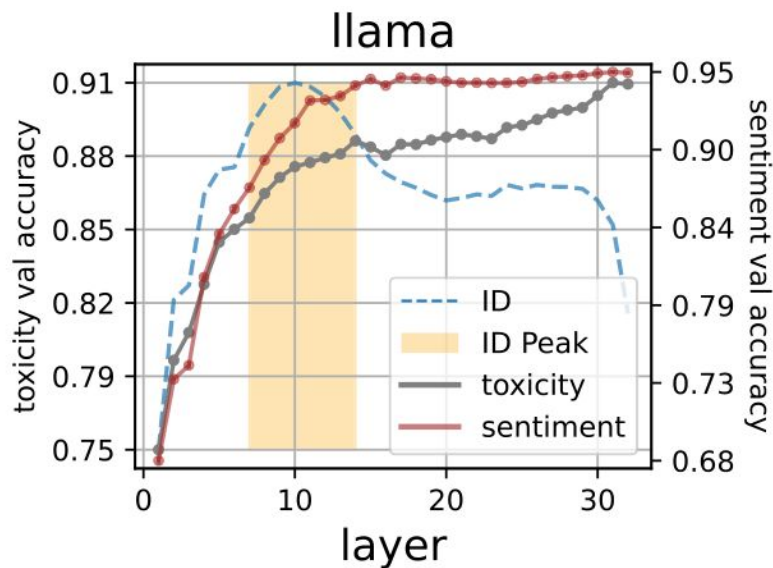


$$\{\mathbf{x}_i \mapsto \hat{d}(\mathbf{x}_i; k)\}_{i=1}^n$$

k specifies the number of neighbours.

ID helps locate useful representations

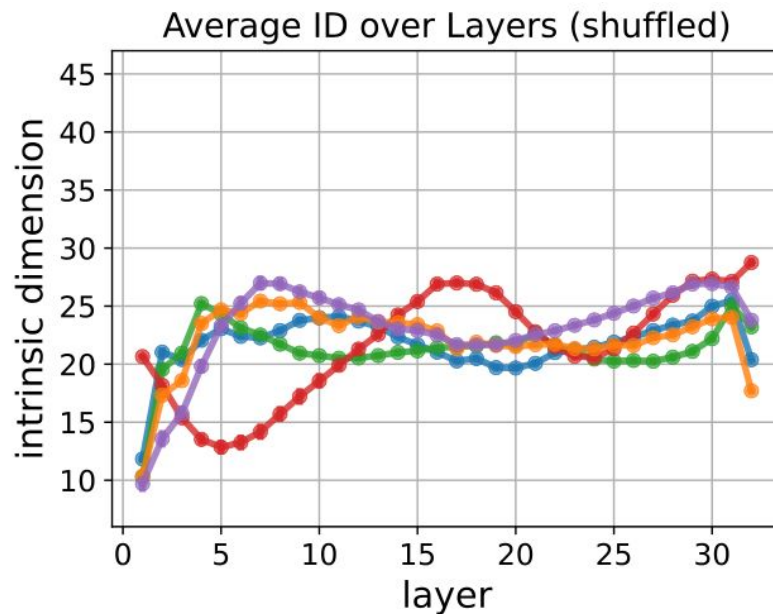
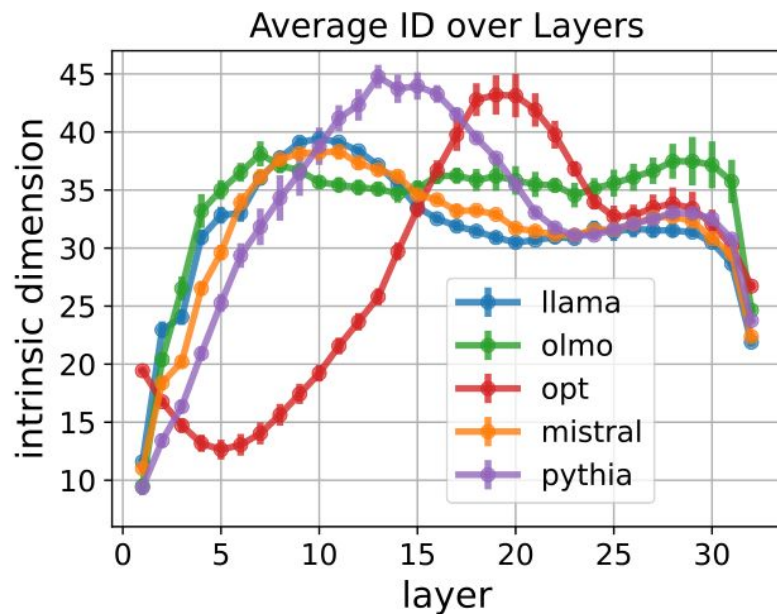
Compare layerwise ID with sentiment and toxicity classification accuracy.



Cheng et al., ICLR 2025 · Fig. 7 (Llama and OPT panels)

ID responds to linguistic structure

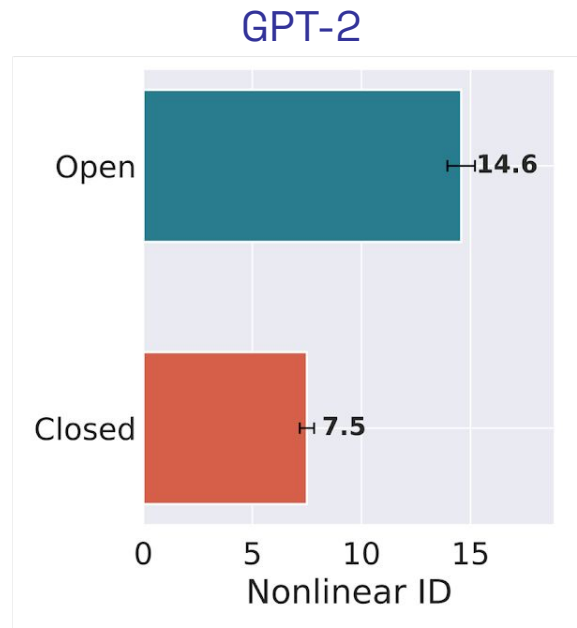
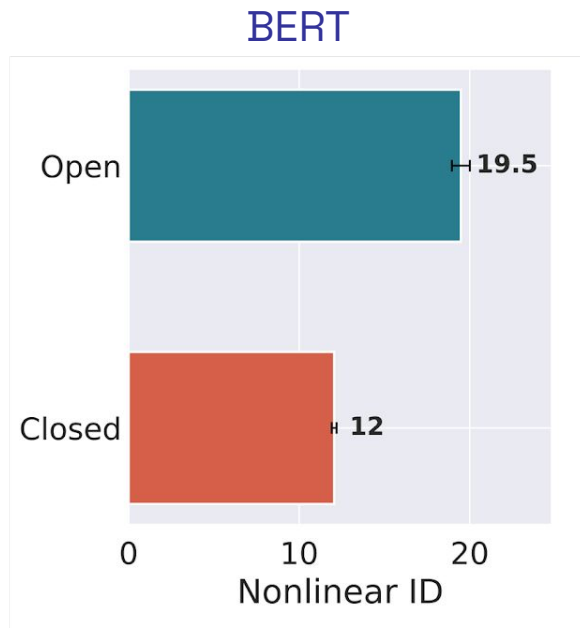
Shuffling tokens reduces the ID peak.



Cheng et al., ICLR 2025 · Fig. 1 (left & centre)

Intrinsic dimensionality by word class

Content words (open class) vs function words (closed class)



Domenichelli, Brunato & Dell'Orletta · CoNLL 2026 · Fig. 8, nonlinear-ID panels

From global to pointwise ID

One dataset-level ID can hide differences between neighbourhoods.

Global ID

One summary for a dataset or layer



$$\mathcal{X} \mapsto \hat{d}_{\text{global}}$$

Pointwise ID

One local estimate per representation

$$d(\mathbf{x}_a) = 1$$

$$d(\mathbf{x}_b) = 2$$

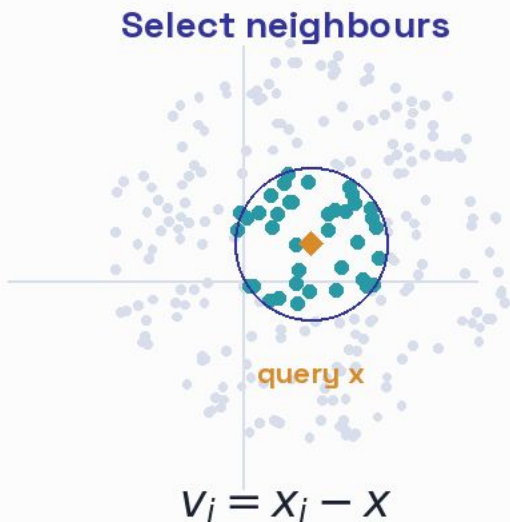


$$\{\mathbf{x}_i \mapsto \hat{d}(\mathbf{x}_i; k)\}_{i=1}^n$$

k specifies the number of neighbours.

One query, one neighbourhood, one estimate

Choose one representation and the reference points closest to it.



1. Pick a query x

The representation we want to describe.

2. Choose k neighbours

Here: 40 other points from one reference cloud.

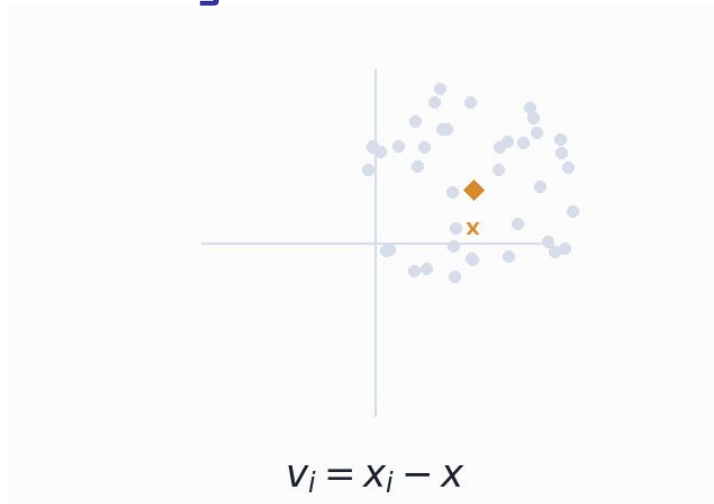
3. Centre on the query

Translate the neighbourhood so x is at the origin.

Two estimators summarize local geometry

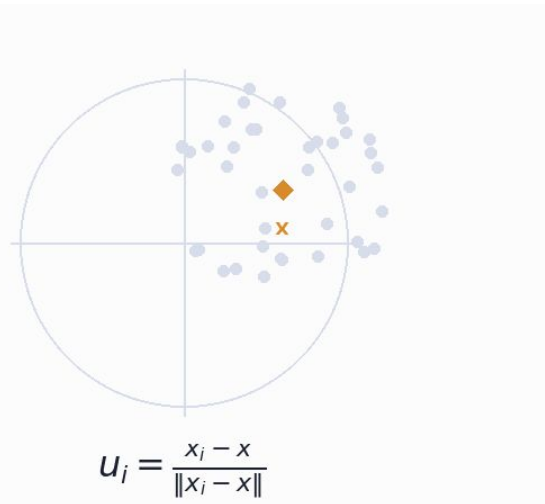
Same query-centred neighbourhood, different geometric summaries.

ESS: local triangles



Weighted mean of normalized areas

FCL: normalized directions

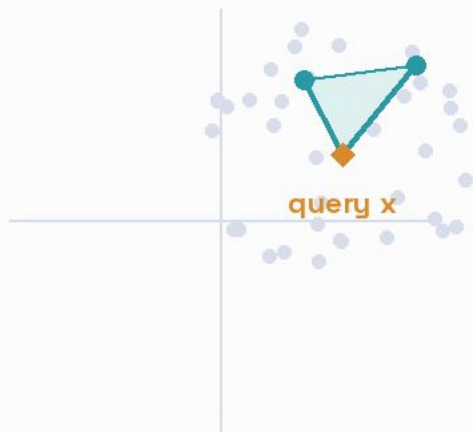


Curve of pairwise chord distances

ESS 1/3: Local triangles

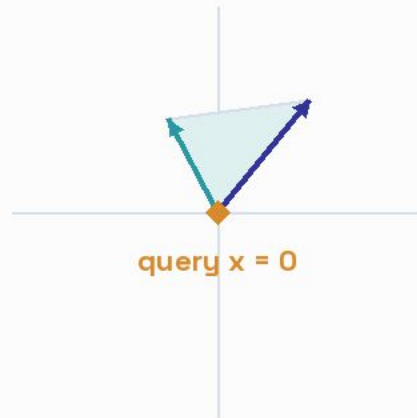
Expected Simplex Skewness: build triangles inside a small neighbourhood.

One query and 40 neighbours



$$V_i = X_i - X$$

Query + two neighbours



Every triangle starts at the query

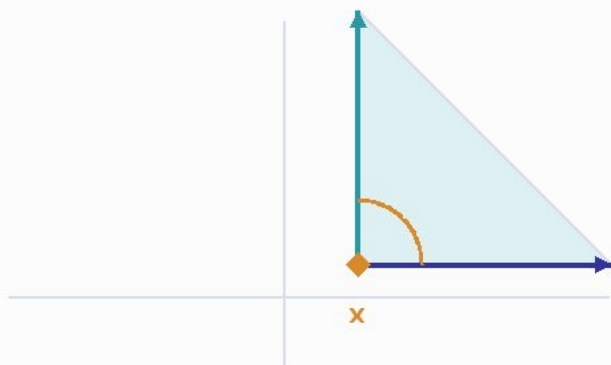
Use the query as the centre, then form triangles with pairs of neighbours.

Johnsson, Soneson & Fontes (2015) · query-centred illustration

ESS 2/3: Normalized area

A flat triangle scores 0; perpendicular edges score 1.

Same lengths, changing angle



Angle 90° • score 1.00



$$q_{ij} = \frac{2A_{ij}}{\|v_i\| \|v_j\|} = |\sin \theta_{ij}|$$

$$v_i = x_i - x$$

Weighted average across pairs

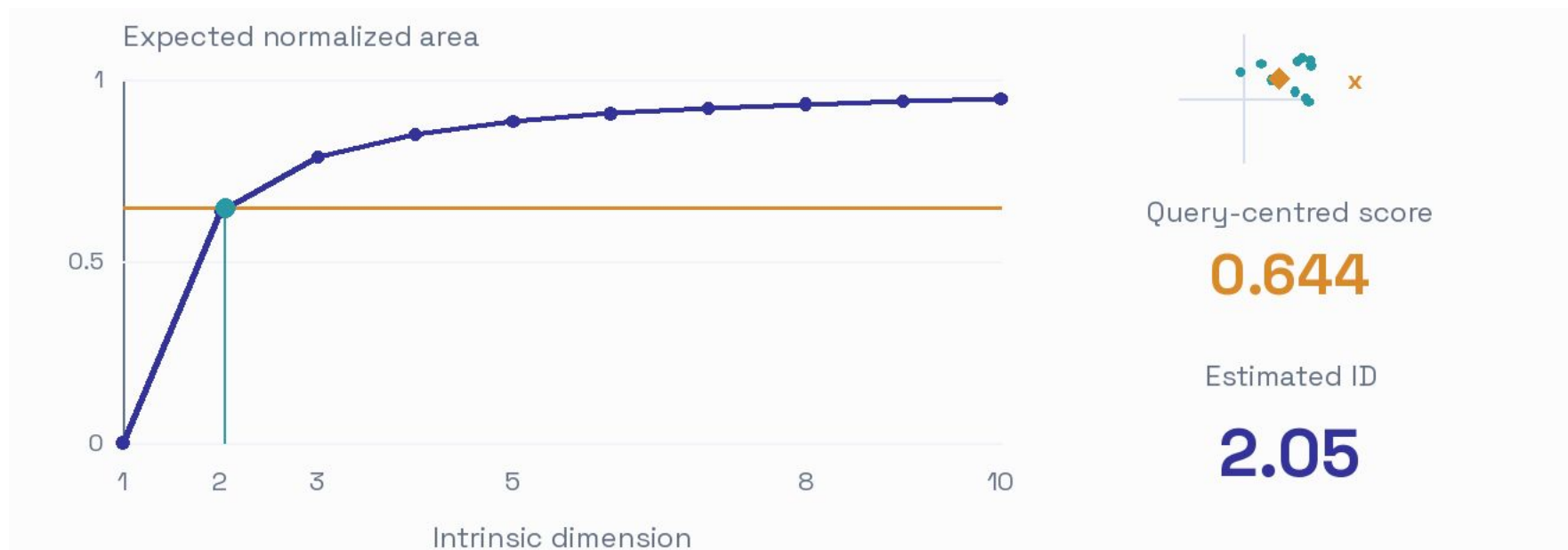
$$w_{ij} = \|v_i\| \|v_j\|$$

Higher-dimensional isotropic patches have larger typical scores.

Johnsson, Soneson & Fontes (2015) · query-centred illustration

ESS 3/3: From score to dimension

Compare the observed score with the isotropic reference curve.

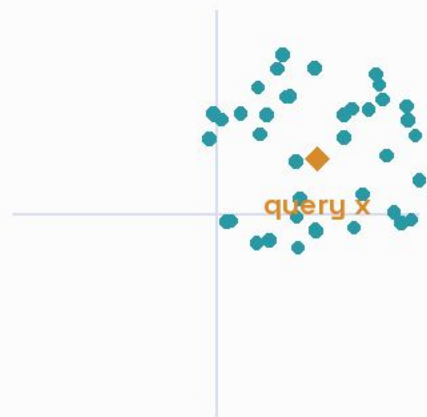


Same 2D patch, centred on the query: ESS $\approx$ 2.05.

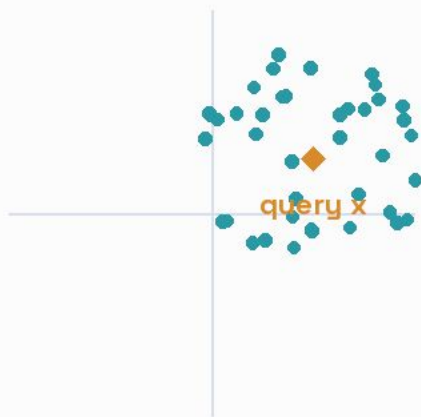
Johnsson, Soneson & Fontes (2015) · query-centred illustration

FCl: turn the same patch into directions

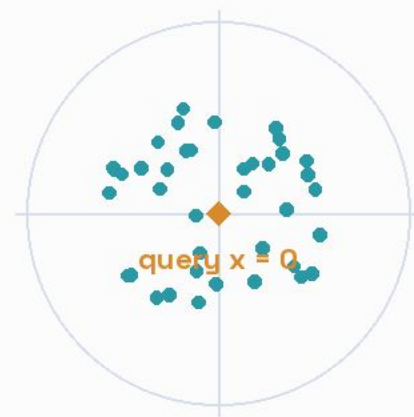
Original patch



Query at the origin

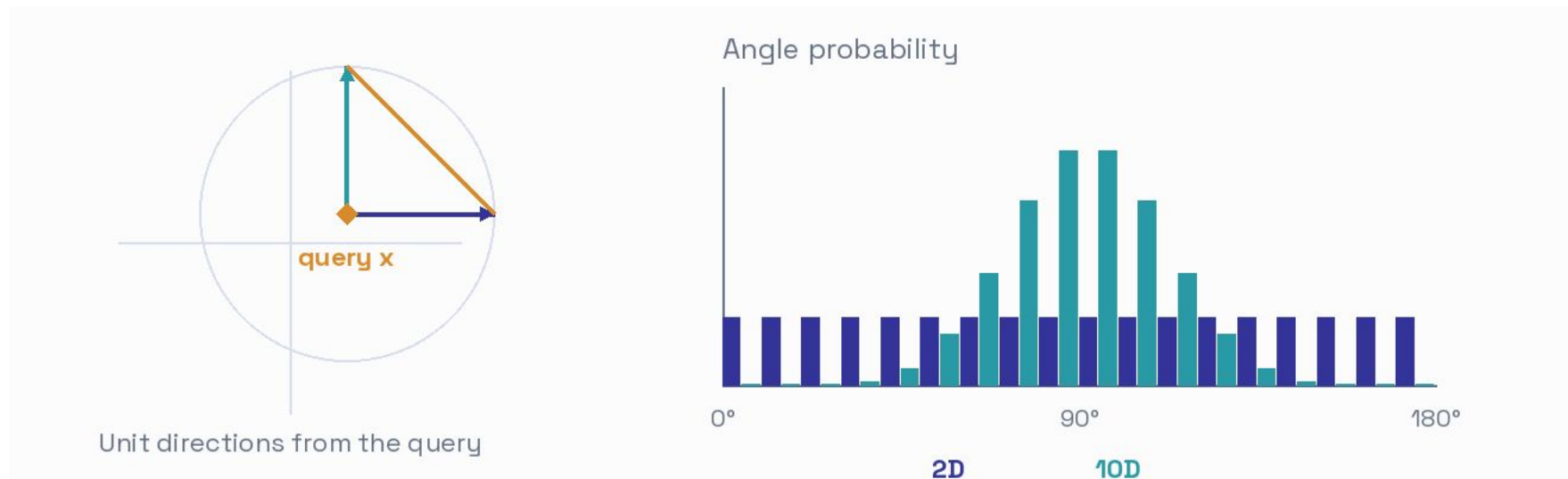


Directions on the sphere



FCl: understand the geometric signal

Higher dimensions make random directions nearly perpendicular.

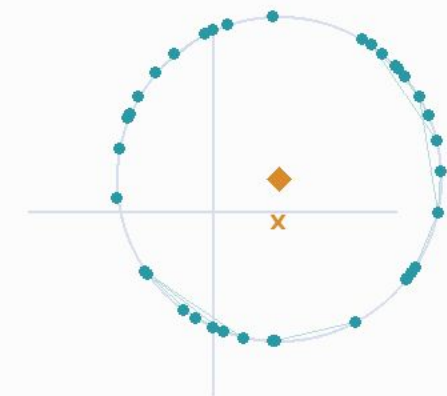


After normalization, nearly perpendicular directions have similar pair distances.

Spherical model · unit radius · bars show theoretical angle probabilities in 2D and 10D

FCl: measure the pair-distance curve

For one patch, vary the cutoff and count the pairs below it.



40 query-centred directions

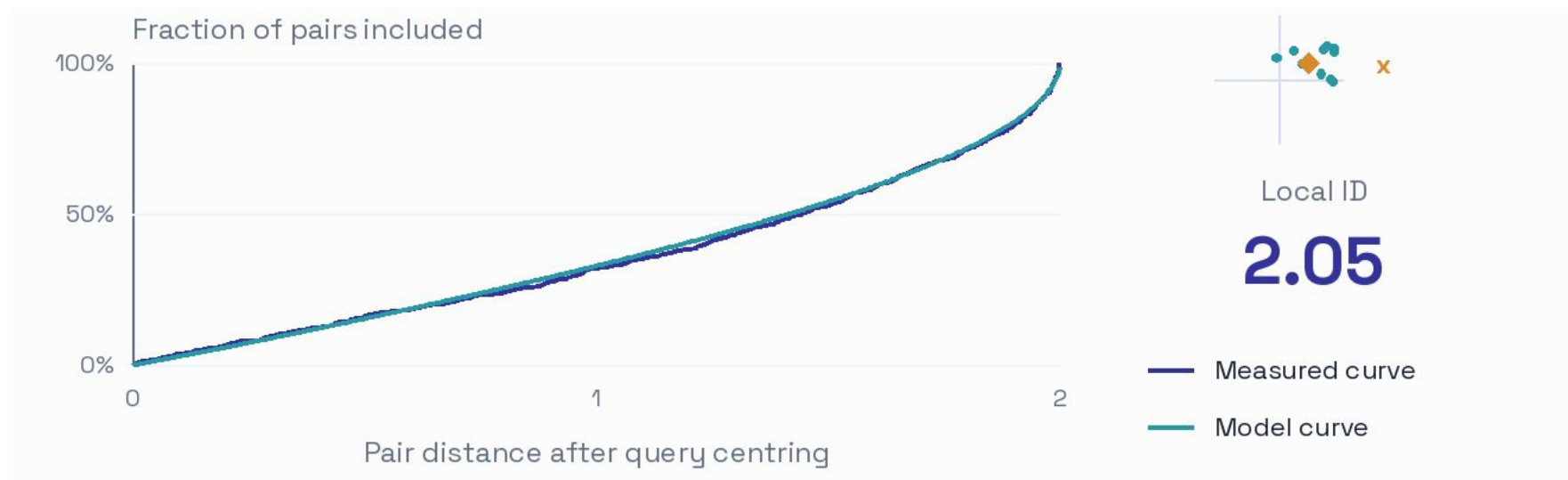


Cutoff 0.75: 181 / 780 pairs

The curve records the fraction of pairs included at each distance cutoff.

FCI: fit a dimension to the measured curve

Keep the data fixed; vary the model dimension to find the best fit.



Our study: linguistic variation in local ID

1. Can linguistic profiles predict pointwise ID?

2. Which properties matter at different layers?

3. When does this relationship emerge in training?

Pair each excerpt with two measurements

The same text provides a geometric target and an interpretable linguistic profile.

One text excerpt

English prose from The Pile

Geometry

Final non-padding-token state
One pointwise ID score

Language

43 linguistic descriptors:
+ 3 frequency controls

Experimental setup

Models

Pythia-6.9B

Qwen2.5-7B

Apertus-8B

Reference manifold

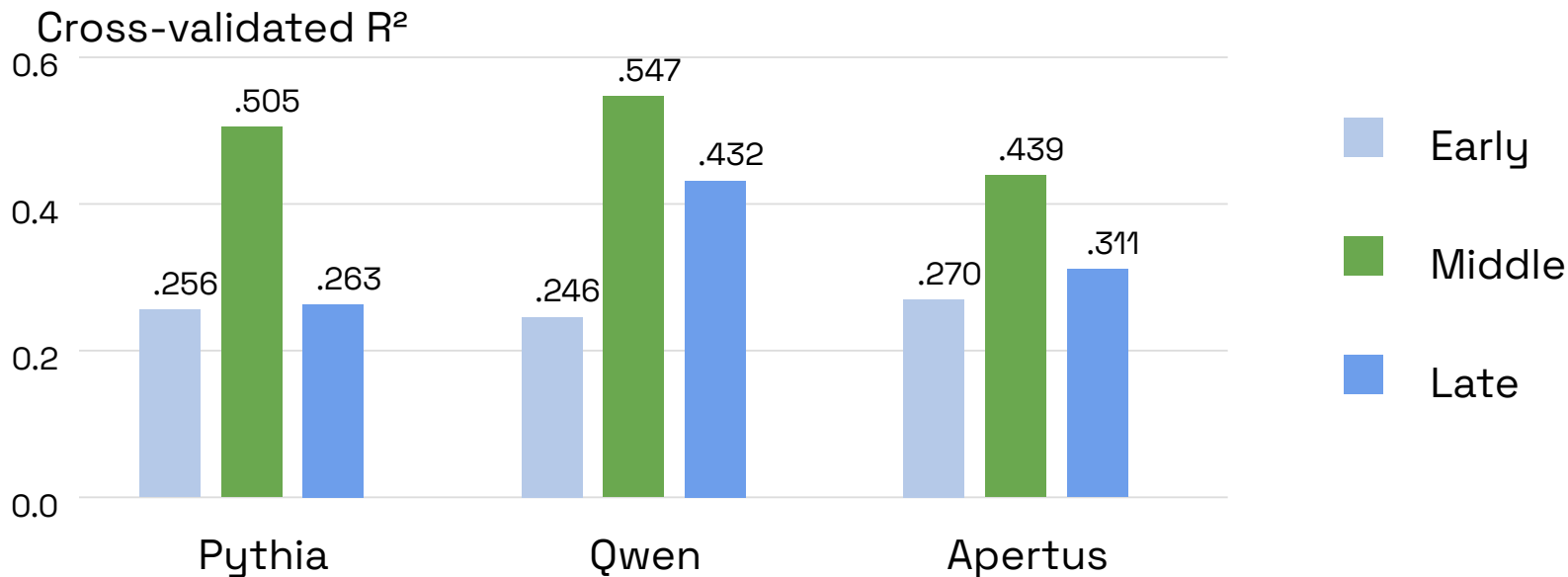
Different dimensions

Neighbourhood size

Different values of k

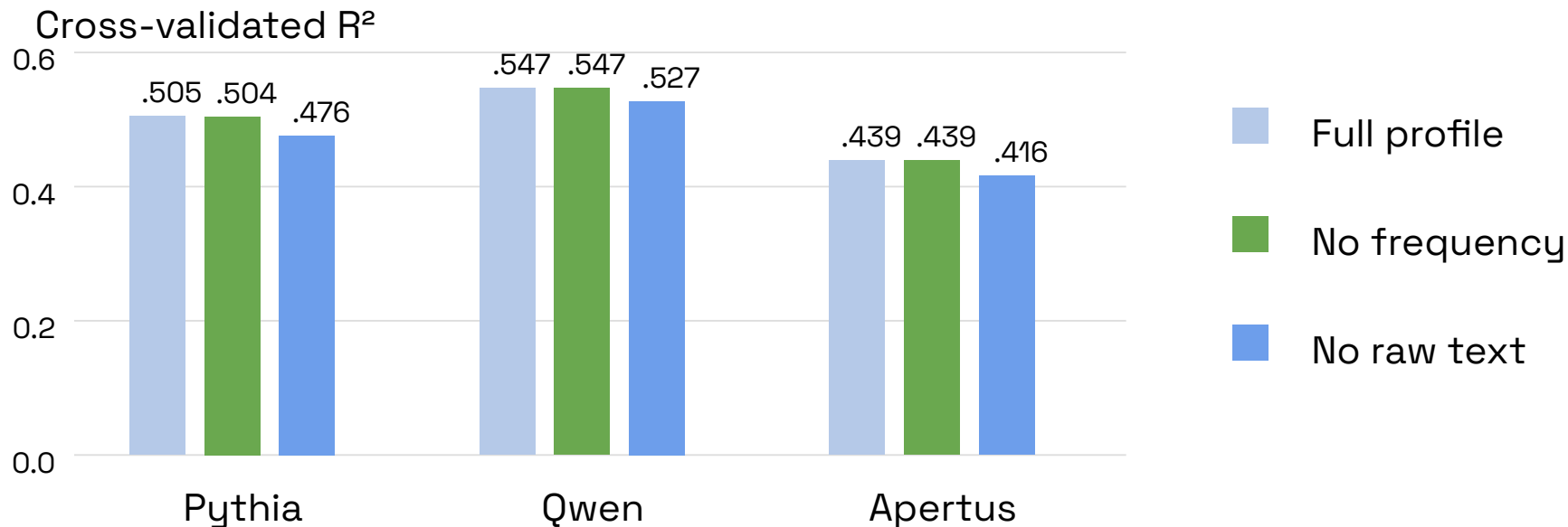
Prediction is strongest at intermediate depth

Linguistic profiles predict held-out excerpt geometry in all three models.

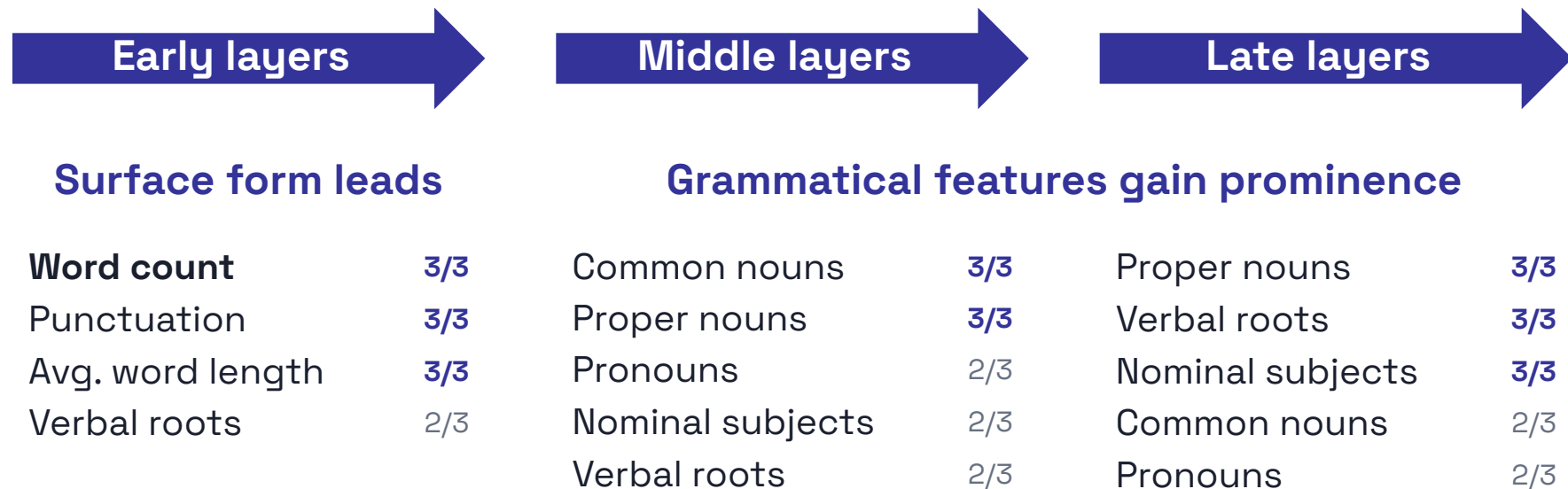


The signal survives basic controls

At intermediate depth, prediction remains strong after removing raw-text descriptors.



The linguistic signal changes with depth



Low ID, high ID

LOW ID

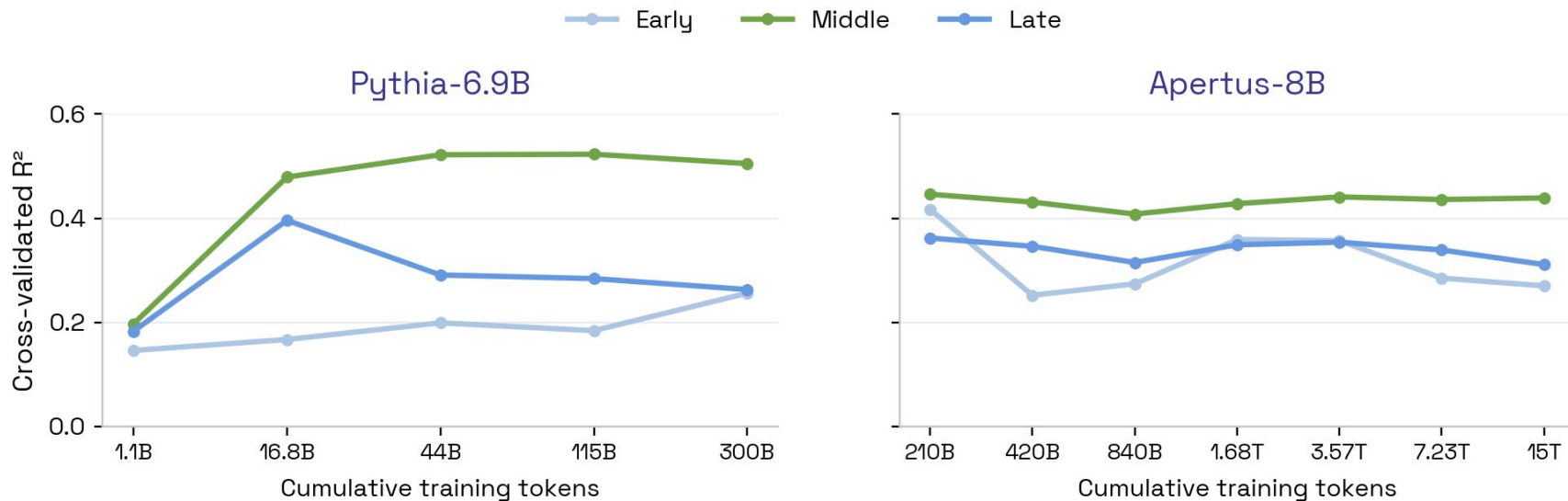
- What is the limit of the Death Eraser? The Death Note Wikia states that...
- Combining an ionic transition metal complex with a conjugated polymer for wide-range voltage-controlled light-emission color.
- Rubia fruticosa Rubia fruticosa is a species of flowering plant in the family Rubiaceae, native to the Canary Islands, Madeira and the Savage Islands.

HIGH ID

- When I heard that Children Of Bodom was putting out a new record, I was both excited and hesitant at the same time.
- In other news: I've re-discovered my love for Bismuth.
- Belostemma Belostemma is a genus of plant in the family Apocynaceae, first described as a genus in 1834. It is native to China, India, and Nepal.
- Too delicious! Lunch at Regine's is always a treat! Yesterday she whipped up a hearty three course meal that no one could resist.

The depth pattern develops during pretraining

The middle-layer advantage persists across later checkpoints in both models.



The End!



Thanks for your attention!

References · Representations

Arora, Liang & Ma (2017)

A Simple but Tough-to-Beat
Baseline for Sentence Embeddings

Reimers & Gurevych (2019)

Sentence-BERT: Sentence
Embeddings using Siamese
BERT-Networks

Ethayarajh (2019)

How Contextual are
Contextualized Word
Representations? Comparing the
Geometry of BERT, ELMo, and
GPT-2 Embeddings

Wang & Kuo (2020)

SBERT-WK: A Sentence Embedding
Method by Dissecting BERT-based
Word Models

Li et al. (2020)

On the Sentence Embeddings
from Pre-trained Language Models

Gao, Yao & Chen (2021)

SimCSE: Simple Contrastive
Learning of Sentence Embeddings

Chen et al. (2023)

Ditto: A Simple and Efficient
Approach to Improve Sentence
Embeddings

Nikolaev & Padó (2023)

Representation biases in sentence
transformers

Aljaafari, Carvalho & Freitas (2026)

Where Do LLMs Compose
Meaning? A Layerwise Analysis of
Compositional Robustness

Goodfellow, Bengio & Courville (2016)

Deep Learning

Fefferman, Mitter & Narayanan (2016)

Testing the manifold hypothesis

References · Geometry

Wang et al. (2023)

Spectral Evolution and Invariance
in Linear-width Neural Networks

Dandi et al. (2025)

A Random Matrix Theory
Perspective on the Spectrum of
Learned Features and Asymptotic
Generalization Capabilities

Marchenko & Pastur (1967)

Distribution of eigenvalues for
some sets of random matrices

Park, Choe & Veitch (2024)

The Linear Representation
Hypothesis and the Geometry of
Large Language Models

Engels et al. (2025)

Not All Language Model Features
Are One-Dimensionally Linear

Nava & Wyart (2026)

Hierarchical Concept Geometry in
Language Models Emerges from
Word Co-occurrence

Aamari & Levrard (2019)

Nonasymptotic rates for manifold,
tangent space and curvature
estimation

Hosseini & Fedorenko (2023)

Large language models implicitly
learn to straighten neural
sentence trajectories to construct
a predictive representation of
natural language

**Edelsbrunner, Letscher &
Zomorodian (2002)**

Topological Persistence and
Simplification

Gardinazzi et al. (2025)

Persistent Topological Features in
Large Language Models

Malhotra et al. (2026)

Tracking Representation Dynamics
in Large Language Models with
Persistent Homology

References · Intrinsic dimensionality

Johnsson, Soneson & Fontes (2015)

Low Bias Local Intrinsic Dimension Estimation from Expected Simplex Skewness

Erba, Gherardi & Rotondo (2019)

Intrinsic dimension estimation for locally undersampled data

Denti et al. (2022)

The generalized ratios intrinsic dimension estimator

Cheng et al. (2025)

Emergence of a High-Dimensional Abstraction Phase in Language Transformers

Rodriguez et al. (2018)

Computing the Free Energy without Collective Variables

Valeriani et al. (2023)

The geometry of hidden representations of large transformer models

Domenichelli et al. (2026)

Linguistic Profiling of Transformer Embedding Geometry

Current manuscript

Linguistic Features of Local Geometry in Language Models

Brunato et al. (2020)

Profiling-UD: a Tool for Linguistic Profiling of Texts