Investigating Representation Geometry in Neural Language Models

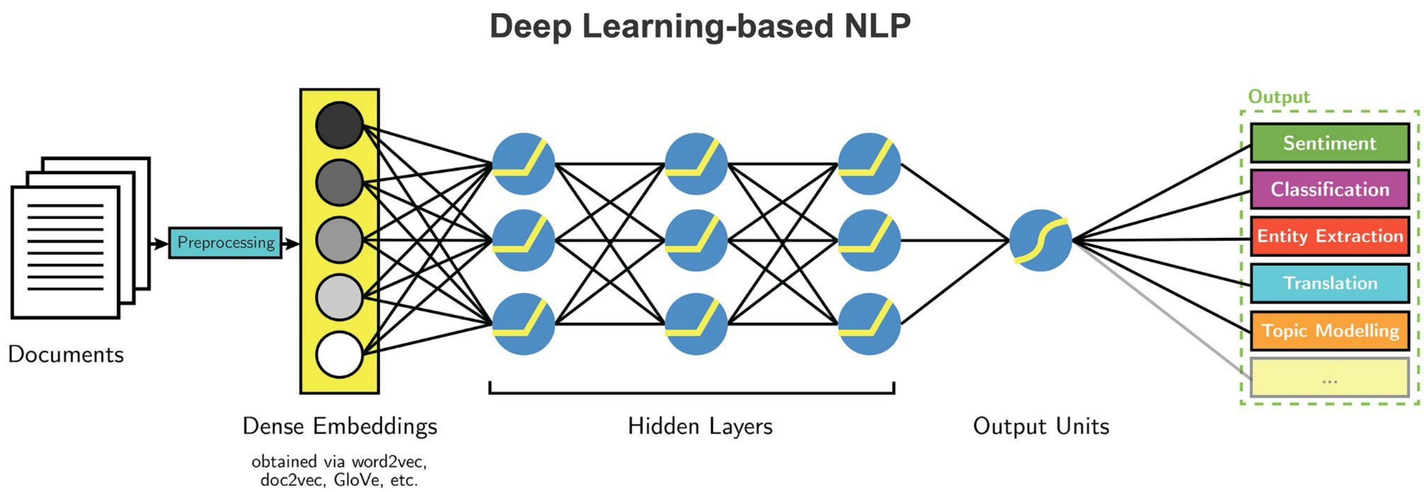
Lucia Domenichelli - National PhD in Artificial Intelligence

TABLE OF CONTENTS

1

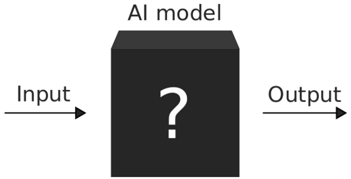
PhD project overview

Some background...



Problems 🚩

- ❑ Efficiency and usage of resources
- ❑ Black boxes



Interpretability

Probing

Mech Int

...

Representation space

Research questions ❓

- ❑ How does representation geometry form and evolve across layers, training objectives, model sizes, and curricula, and can early-stage geometry predict eventual performance or learning speed?
- ❑ How do adaptation choices, such as fine-tuning strategies reshape geometry? Are these changes durable or task-specific, and do they translate into accuracy gains?
- ❑ Do geometric properties reliably encode and distinguish linguistic features across languages, and how do these properties correlate with (or predict changes in) downstream behavior?

Research questions ❓

- How does representation geometry form and evolve across layers, training objectives, model sizes, and curricula, and can early-stage geometry predict eventual performance or learning speed?
- How do adaptation choices, such as fine-tuning strategies reshape geometry? Are these changes durable or task-specific, and do they translate into accuracy gains?
- Do geometric properties reliably encode and distinguish linguistic features across languages, and how do these properties correlate with (or predict changes in) downstream behavior?

Why understand LLMs?

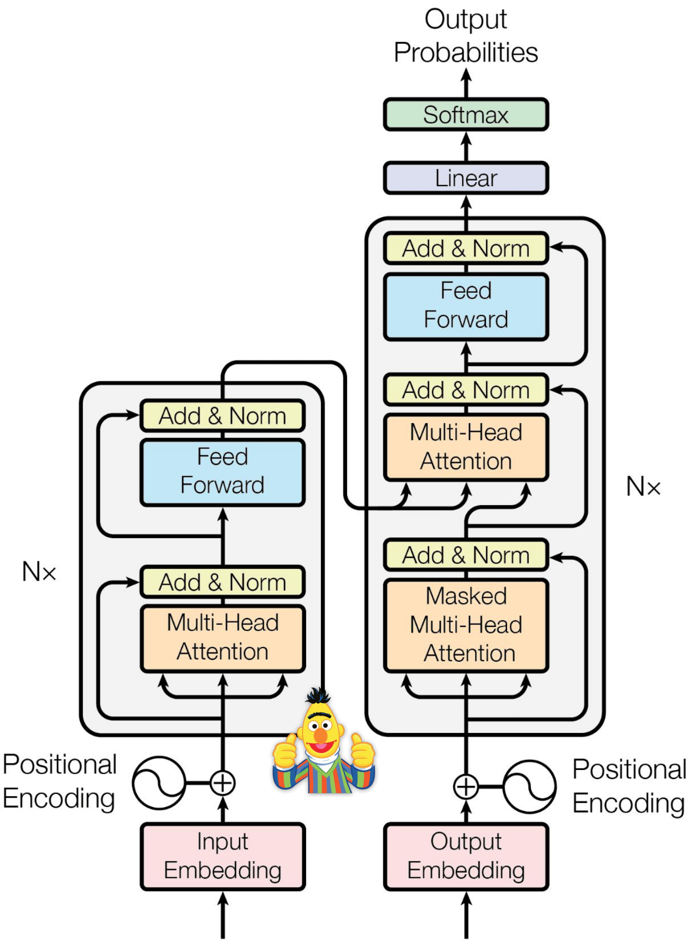
- They are everywhere
- Their training doesn't lend itself towards trust:
  - Unsupervised pretraining
  - Supervised finetuning
  - RLHF
- An understandable fear of hallucinations and malicious outputs.

2

State of the Art

The Transformer model

- ❏ Transformers Models have become ubiquitous in Natural Language Processing
- ❏ Pretraining + finetuning
- ❏ Still the backbone!

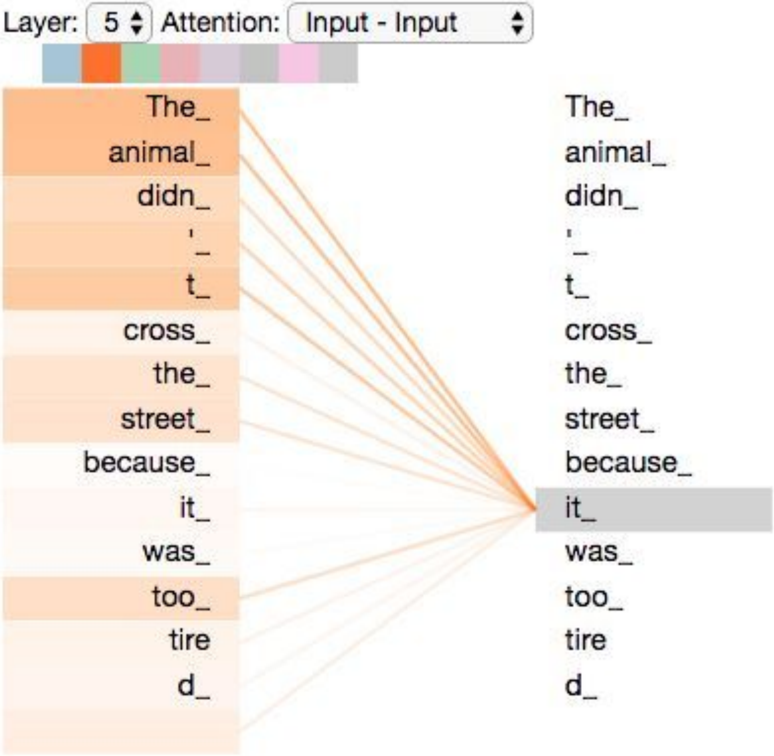


Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Attention is all you need!

- Attention is the method that allows the model to "attend" to different positions of the input sequence to compute a representation of that sequence.

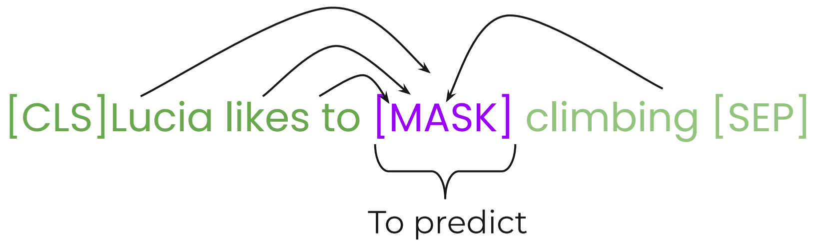
$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$



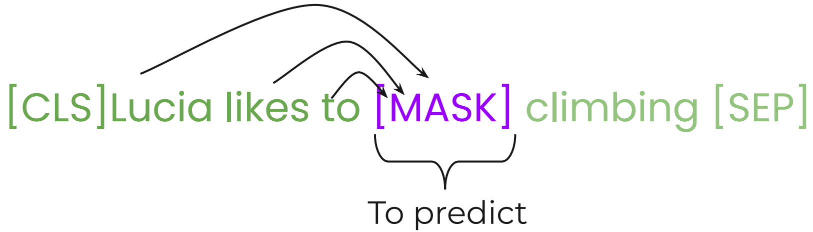
Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Pretraining

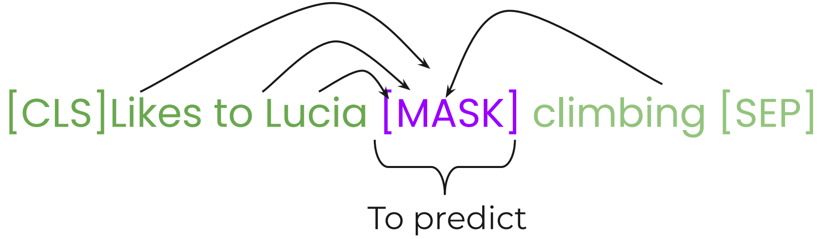
- Masked language modeling (Encoders)



- Causal language modeling (Decoders)

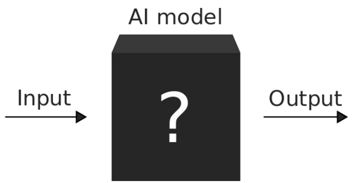


- Denoising autoencoders (Encoder+decoder)



2.1

Interpretability



The Case of Interpretability

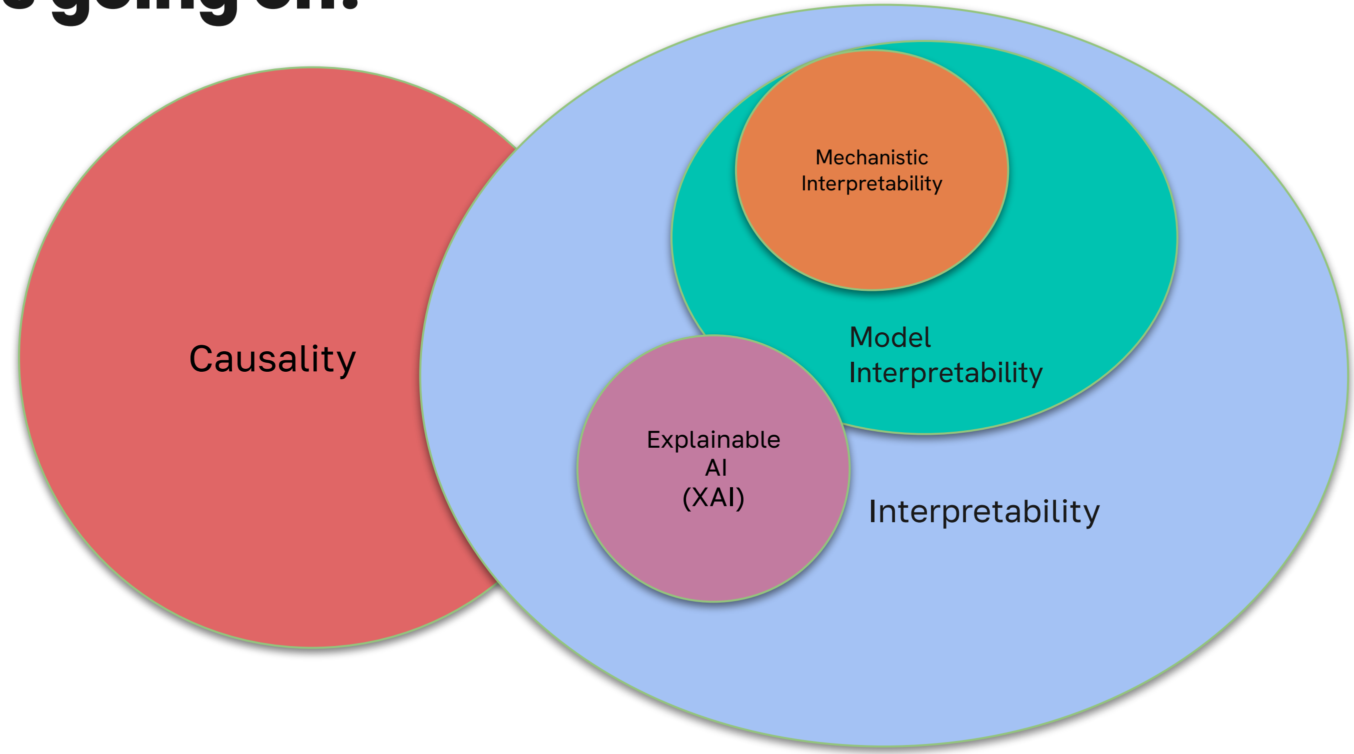
- ❑ The development of powerful state-of-the-art NLMs comes at the cost of interpretability, since complex NN models offer little transparency about their inner workings and their abilities.

Objectives:

- ❑ Understand the nature of AI systems → be faithful to what influences the AI decisional process.

- ❑ Empower AI system users → derive actionable useful insights from AI choices

What's going on?



3.1

■ Static approaches

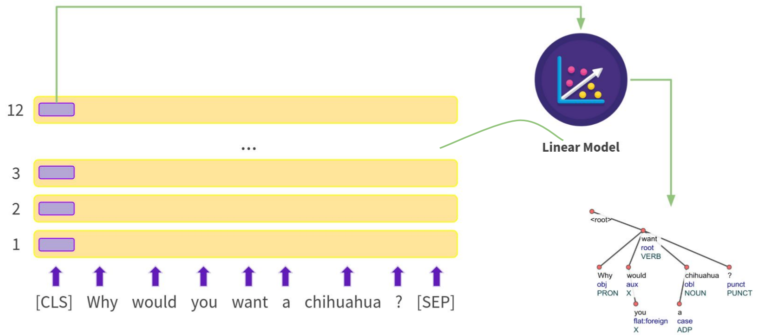
Probings

- Core idea: use supervised models (the probes) to determine what is latently encoded in the hid

- Representations ha … osyntactic and semantic informati

- A very powerful pro … in the target model (but rather in your probe) → Control tasks

Defined this way, probes are correlative not causative!



[Img:https://github.com/gsarti/lcl23-xnlm-lab/blob/main/notebooks/1.2_Probing.ipynb]

Linguistic probings

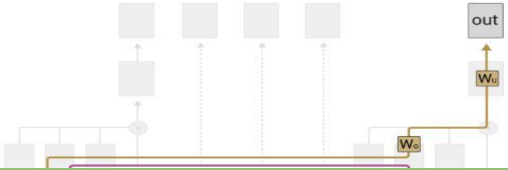
Layerwise ρ scores for the 68 linguistic features.

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	B
char_per_tok	0.46	0.44	0.44	0.4	0.4	0.4	0.38	0.35	0.34	0.32	0.33	0.32	0.032
sent_length	0.99	0.99	0.98	0.98	0.98	0.97	0.97	0.97	0.97	0.96	0.96	0.95	1
ttr_form	0.8	0.8	0.81	0.81	0.81	0.8	0.8	0.78	0.78	0.75	0.72	0.71	0.2
ttr_lemma	0.79	0.79	0.8	0.8	0.8	0.8	0.8	0.78	0.78	0.75	0.72	0.71	0.26
lexical_density	0.79	0.81	0.81	0.81	0.8	0.79	0.79	0.78	0.77	0.74	0.72	0.72	0.18
upos_dist_ADJ	0.67	0.69	0.68	0.67	0.66	0.65	0.64	0.63	0.63	0.61	0.6	0.58	0.27
upos_dist_ADP	0.86	0.86	0.86	0.84	0.83	0.81	0.78	0.76	0.75	0.72	0.7	0.69	0.46
upos_dist_ADV	0.68	0.7	0.67	0.64	0.62	0.61	0.61	0.6	0.59	0.57	0.55	0.54	0.28
upos_dist_AUX	0.81	0.84	0.84	0.84	0.82	0.82	0.82	0.8	0.8	0.79	0.77	0.77	0.25
upos_dist_CCONJ	0.86	0.86	0.85	0.83	0.81	0.8	0.77	0.74	0.74	0.71	0.67	0.66	0.44
upos_dist_DET	0.89	0.9	0.89	0.87	0.85	0.84	0.83	0.81	0.79	0.77	0.73	0.74	0.42
upos_dist_NUM	0.63	0.63	0.62	0.6	0.58	0.58	0.57	0.56	0.55	0.54	0.53	0.53	0.18
upos_dist_PART	0.7	0.71	0.7	0.69	0.66	0.64	0.63	0.61	0.6	0.58	0.57	0.57	0.35
upos_dist_PRON	0.87	0.88	0.88	0.88	0.88	0.87	0.87	0.87	0.86	0.85	0.84	0.83	0.22
upos_dist_PROPN	0.63	0.63	0.64	0.65	0.66	0.67	0.67	0.67	0.67	0.66	0.65	0.65	0.083
upos_dist_SCONJ	0.58	0.58	0.57	0.57	0.55	0.56	0.55	0.55	0.55	0.55	0.53	0.52	0.39
upos_dist_VERB	0.77	0.79	0.8	0.8	0.81	0.81	0.81	0.8	0.79	0.78	0.77	0.76	0.25
xpos_dist_,	0.73	0.72	0.7	0.7	0.69	0.67	0.65	0.62	0.62	0.59	0.56	0.58	0.36
xpos_dist_.	0.75	0.76	0.77	0.81	0.81	0.8	0.81	0.8	0.79	0.78	0.76	0.73	0.26
xpos_dist_NN	0.6	0.61	0.63	0.64	0.64	0.64	0.64	0.63	0.62	0.6	0.58	0.58	0.1
xpos_dist_NNS	0.58	0.6	0.63	0.63	0.63	0.63	0.63	0.61	0.61	0.58	0.55	0.54	0.3
xpos_dist_RB	0.67	0.68	0.66	0.63	0.62	0.62	0.62	0.61	0.6	0.58	0.56	0.56	0.23
xpos_dist_TO	0.63	0.63	0.62	0.6	0.57	0.55	0.53	0.5	0.49	0.48	0.47	0.47	0.32
xpos_dist_VB	0.68	0.69	0.69	0.68	0.68	0.68	0.69	0.68	0.68	0.68	0.67	0.67	0.21
xpos_dist_VBD	0.64	0.66	0.68	0.68	0.68	0.68	0.68	0.67	0.67	0.68	0.68	0.68	0.25
xpos_dist_VBN	0.51	0.54	0.53	0.54	0.52	0.51	0.49	0.48	0.47	0.46	0.45	0.45	0.3
xpos_dist_VBP	0.61	0.63	0.63	0.64	0.63	0.63	0.63	0.62	0.63	0.62	0.61	0.62	0.17
xpos_dist_VBZ	0.64	0.67	0.67	0.69	0.67	0.67	0.66	0.65	0.64	0.63	0.62	0.63	0.19
aux_form_dist_Fin	0.74	0.77	0.76	0.76	0.75	0.74	0.73	0.72	0.71	0.72	0.71	0.71	0.42
aux_mood_dist_Ind	0.76	0.78	0.78	0.78	0.77	0.76	0.76	0.75	0.74	0.74	0.73	0.73	0.42
aux_Sing+3	0.7	0.71	0.71	0.7	0.69	0.68	0.67	0.65	0.64	0.64	0.63	0.63	0.27
aux_tense_dist_Pres	0.72	0.74	0.73	0.75	0.74	0.73	0.73	0.72	0.72	0.71	0.7	0.71	0.3
avg_links_len	0.82	0.83	0.83	0.82	0.83	0.83	0.84	0.83	0.83	0.82	0.8	0.8	0.79
avg_prep_chain_len	0.74	0.74	0.74	0.73	0.72	0.71	0.7	0.69	0.68	0.67	0.65	0.65	0.54

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	B
avg_subord_chain_len	0.8	0.81	0.82	0.81	0.81	0.81	0.8	0.8	0.8	0.79	0.78	0.77	0.66
avg_token_per_clause	0.76	0.77	0.78	0.77	0.77	0.77	0.77	0.77	0.77	0.76	0.75	0.75	0.62
avg_verb_edges	0.72	0.73	0.74	0.74	0.75	0.75	0.75	0.74	0.75	0.75	0.75	0.74	0.6
dep_dist_advcl	0.55	0.56	0.55	0.54	0.53	0.54	0.53	0.54	0.53	0.53	0.52	0.51	0.4
dep_dist_advmod	0.72	0.73	0.71	0.68	0.66	0.66	0.66	0.65	0.64	0.62	0.6	0.6	0.28
dep_dist_amod	0.63	0.64	0.64	0.64	0.64	0.63	0.64	0.63	0.62	0.59	0.57	0.55	0.36
dep_dist_aux	0.69	0.72	0.72	0.73	0.72	0.71	0.71	0.69	0.69	0.68	0.66	0.67	0.29
dep_dist_case	0.85	0.85	0.85	0.83	0.83	0.81	0.79	0.77	0.76	0.74	0.72	0.71	0.47
dep_dist_cc	0.85	0.85	0.85	0.83	0.81	0.8	0.77	0.74	0.74	0.71	0.68	0.66	0.44
dep_dist_compound	0.5	0.52	0.55	0.57	0.58	0.58	0.56	0.56	0.56	0.54	0.53	0.52	0.27
dep_dist_conj	0.82	0.82	0.81	0.8	0.78	0.77	0.75	0.74	0.74	0.72	0.69	0.68	0.47
dep_dist_cop	0.62	0.63	0.63	0.64	0.62	0.62	0.61	0.6	0.59	0.58	0.57	0.57	0.19
dep_dist_det	0.9	0.9	0.9	0.88	0.86	0.85	0.84	0.81	0.8	0.77	0.74	0.74	0.42
dep_dist_mark	0.72	0.72	0.72	0.71	0.7	0.69	0.69	0.68	0.68	0.66	0.65	0.64	0.45
dep_dist_nmod	0.69	0.7	0.69	0.67	0.66	0.66	0.64	0.63	0.63	0.61	0.59	0.6	0.45
dep_dist_nmod:poss	0.64	0.67	0.67	0.65	0.63	0.63	0.62	0.59	0.58	0.55	0.53	0.52	0.29
dep_dist_nsubj	0.79	0.81	0.82	0.83	0.84	0.85	0.85	0.85	0.85	0.85	0.84	0.83	0.2
dep_dist_obj	0.66	0.69	0.69	0.7	0.71	0.71	0.72	0.71	0.69	0.68	0.67	0.66	0.29
dep_dist_obl	0.66	0.67	0.67	0.66	0.65	0.64	0.62	0.61	0.61	0.59	0.57	0.56	0.43
dep_dist_punct	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.83	0.83	0.81	0.78	0.77	0.14
max_links_len	0.89	0.9	0.89	0.89	0.88	0.88	0.88	0.88	0.87	0.87	0.86	0.85	0.91
obj_post	0.69	0.71	0.72	0.73	0.73	0.74	0.74	0.73	0.72	0.72	0.71	0.7	0.47
parse_depth	0.91	0.91	0.91	0.91	0.91	0.91	0.9	0.9	0.9	0.9	0.9	0.89	0.89
prep_dist_1	0.63	0.63	0.63	0.61	0.61	0.6	0.59	0.59	0.58	0.57	0.56	0.55	0.47
principal_prop_dist	0.63	0.66	0.68	0.68	0.7	0.72	0.73	0.73	0.74	0.73	0.71	0.7	0.066
subj_pre	0.7	0.72	0.72	0.72	0.72	0.73	0.73	0.73	0.74	0.74	0.74	0.73	0.55
subordinate_dist_1	0.55	0.56	0.56	0.55	0.56	0.56	0.56	0.56	0.56	0.55	0.54	0.54	0.49
subordinate_post	0.7	0.71	0.72	0.71	0.72	0.73	0.73	0.73	0.72	0.72	0.7	0.7	0.55
subord_prop_dist	0.76	0.77	0.78	0.77	0.77	0.77	0.77	0.76	0.76	0.76	0.74	0.74	0.62
verbal_arity_2	0.41	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.42	0.41	0.25
verbal_arity_3	0.41	0.42	0.42	0.42	0.42	0.42	0.43	0.42	0.42	0.41	0.41	0.4	0.35
verbal_arity_4	0.44	0.44	0.44	0.44	0.44	0.44	0.45	0.45	0.44	0.44	0.44	0.44	0.41
verbal_heads_dist	0.9	0.91	0.91	0.9	0.9	0.9	0.89	0.89	0.89	0.88	0.87	0.87	0.79
verbal_root_perc	0.64	0.66	0.66	0.67	0.68	0.68	0.69	0.69	0.69	0.69	0.69	0.69	0.51

[Linguistic Profiling of a Neural Language Model](https://aclanthology.org/2020.coling-main.65/) (Miaschi et al., COLING 2020)

Mechanistic Interpretability



Difficult to apply on larger models!

SUPERPOSITION

The phenomenon where individual neurons represent multiple, overlapping features rather than a single, distinct concept.

CIRCUITS

The specific arrangements of neurons and attention heads within transformer models that collaboratively perform distinct computational tasks or operations, enabling structured processing and representation of information.

INDUCTION HEADS

Attention heads in transformer models specialized in detecting and copying repeated token patterns, enabling the model to generalize by inducing simple structural rules from prior context.

Elhage, et al., "A Mathematical Framework for Transformer Circuits", Transformer Circuits Thread, 2021.

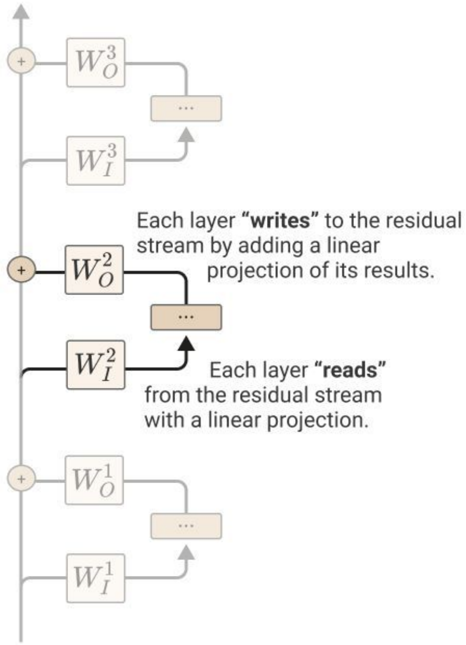
Mechanistic Interpretability – The residual stream

- At any point during the forward pass, the residual stream is simply the sum of the activations of all prior (Attention+MLP) layers along with the initial embedding.

- Attention heads use their **Wv** and **Wo** matrices to **read** and **write** from the **residual stream**.

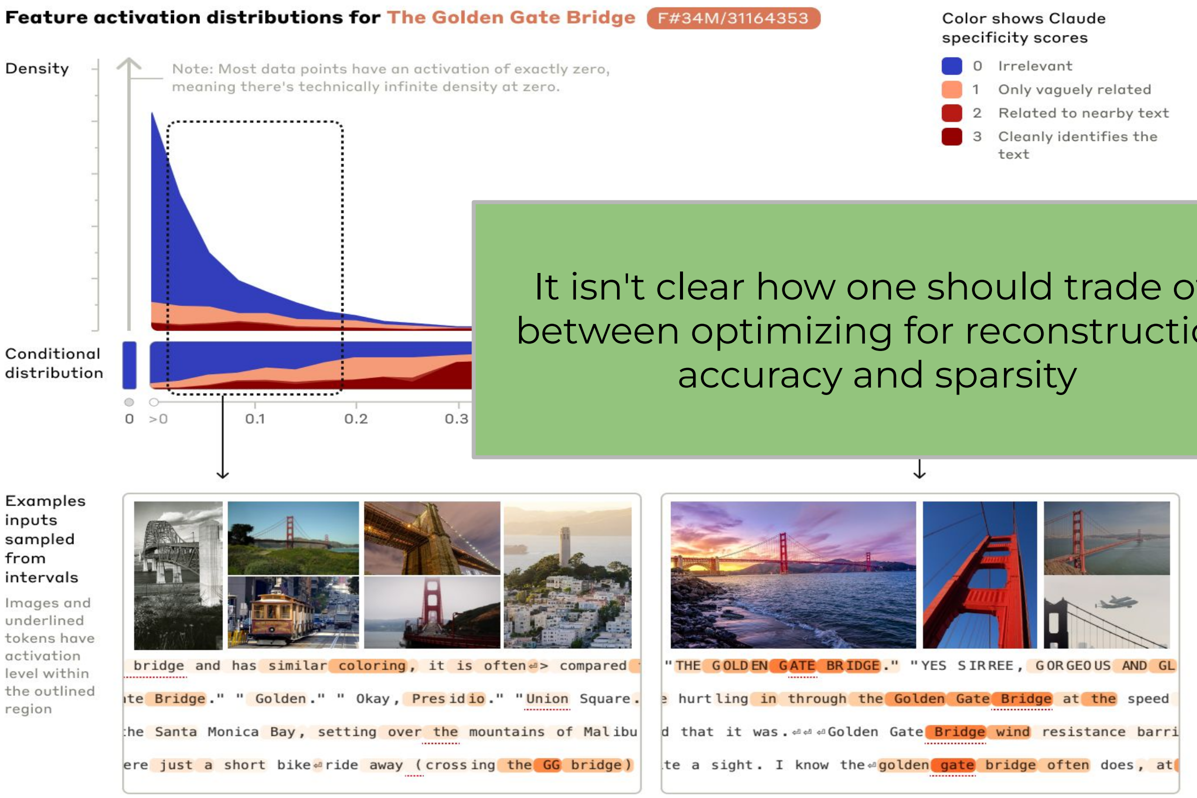
- These matrices help understand which portions of the residual stream individual attention heads modify, as well as which portions they use to perform this modification.

The residual stream is modified by a sequence of MLP and attention layers “reading from” and “writing to” it with linear operations.



Elhage, et al., "A Mathematical Framework for Transformer Circuits", Transformer Circuits Thread, 2021.

Mechanistic Interpretability



It isn't clear how one should trade off between optimizing for reconstruction accuracy and sparsity

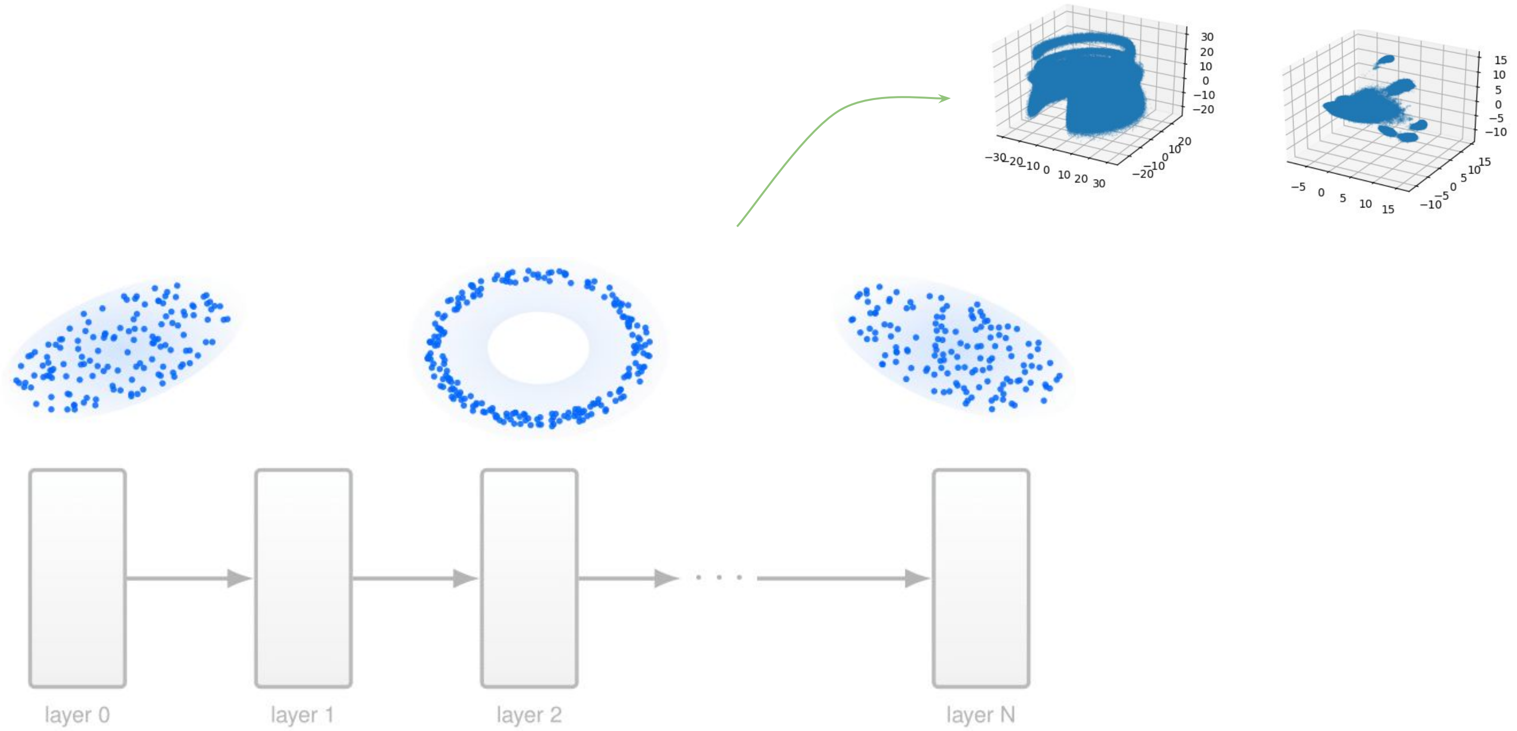
Using SAE!

Templeton, et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet", Transformer Circuits Thread, 2024.

2.2

Our approach

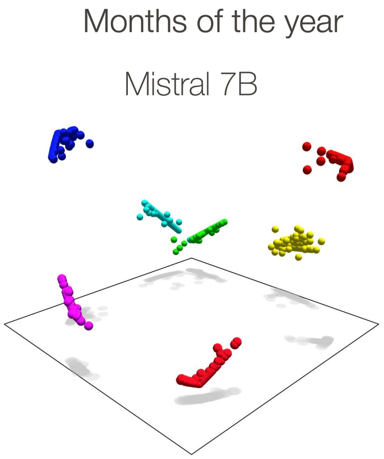
The geometry of Large Language Models

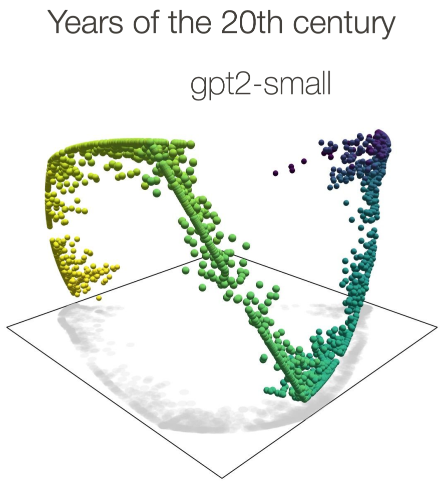


The basics

- **Narrow Cone Hypothesis (Ethayarajh 2019)** → Models are not isotropic, i.e., they do not uniformly utilize the embedding space.

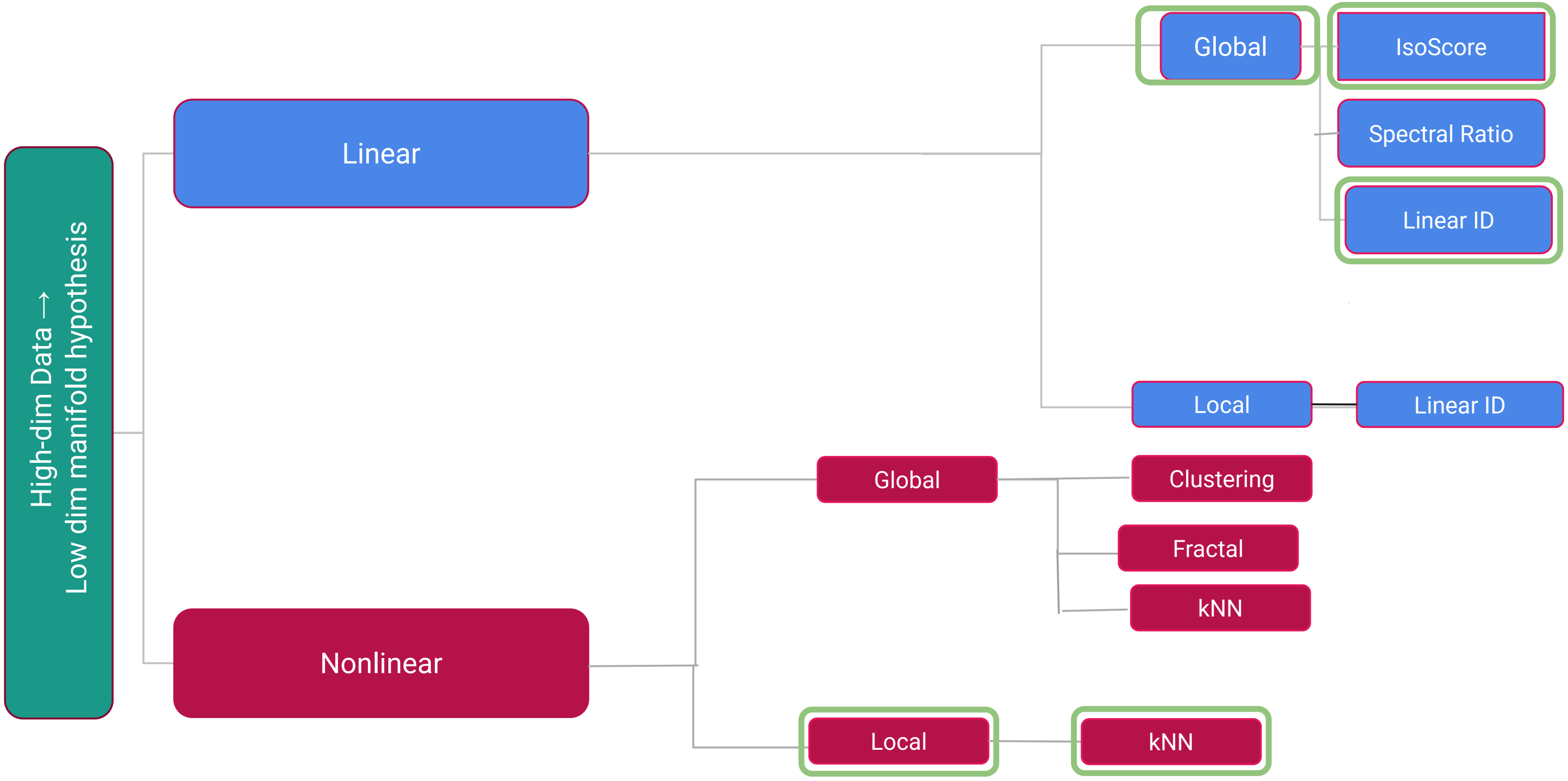
- **Manifold Hypothesis (Clayton, 2015)** → High Dimensional data lies on a manifold of much lower dimensionality than the number of features.





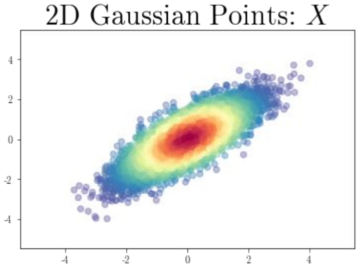
Modell, Alexander, Patrick Rubin-Delanchy, and Nick Whiteley. "The Origins of Representation Manifolds in Large Language Models." *arXiv preprint arXiv:2505.18235* (2025).

Taxonomy 📚

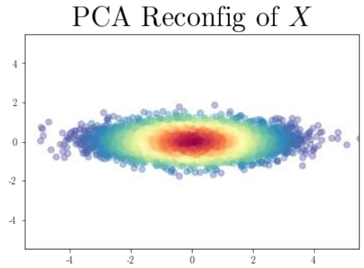


High-dim Data →
Low dim manifold hypothesis
Linear
Global
IsoScore
Spectral Ratio
Linear ID
Local
Linear ID
Nonlinear
Global
Clustering
Fractal
kNN
Local
kNN

IsoScore



1) Point cloud X in R^2.



2) Project X using PCA to get X^{PCA}.

$$\begin{pmatrix} 1.80 & 0.00 \\ 0.00 & 0.20 \end{pmatrix}$$

3) Compute covariance of X^{PCA}.

$$\frac{\sqrt{2}}{\|\begin{pmatrix}1.80 & 0.20\end{pmatrix}\|} \cdot \begin{pmatrix}1.80 & 0.20\end{pmatrix}$$

4) Normalize the diagonal of X^{PCA} to have the same norm as (1,1) to get V^{PCA}.

$$\frac{\|\begin{pmatrix}1.41 & 0.16\end{pmatrix} - \begin{pmatrix}1 & 1\end{pmatrix}\|}{\sqrt{2(2-\sqrt{2})}}$$

5) Calculate the Euclidean distance between V^{PCA} and (1,1) then normalize.

$$0.22$$

6) Linearly rescale to be in the interval [0,1].

Rudman, William, et al. "IsoScore: Measuring the Uniformity of Embedding Space Utilization." *Findings of the Association for Computational Linguistics: ACL 2022.*

Linear ID

1) Point cloud *X* in R^2.

2) *Standardize features and perform PCA*

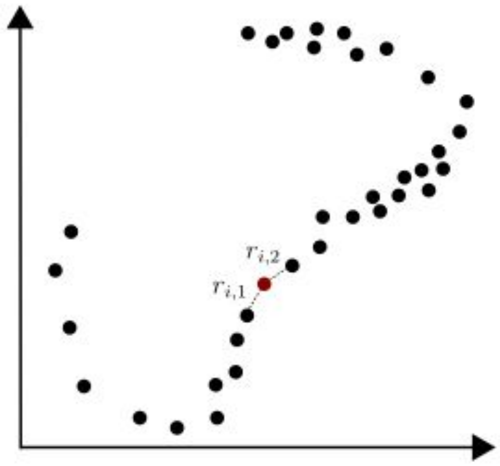
3) *Compute cumulative sums, normalizes by total* $\tilde{S}(d) = \frac{S(d)}{S(D)} = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i}$

4) Finds smallest d such that:

This first two metrics only capture up to 2º order statistics of a point cloud.

Many variants exist

Nonlinear metric: TwoNN



1) Point cloud *X* in *R*^{2}. *xi be uniformly sampled on a manifold with intrinsic dimension d*

2) Compute $$\mu_i = \frac{r_{i,2}}{r_{i,1}}$$

3) The the probability distribution of μ_i is $p(\mu_i|d) = \frac{d}{\mu_i^{d+1}}$ *where d is the Intrinsic Dimensionality*

4) Infer ID from the empirical probability distribution.

5) Repeat the calculation selecting a fraction of points at random. This gives the ID as a function of the scale.

→ Many caveats!

Ansuini, Alessio, et al. "Intrinsic dimension of data representations in deep neural networks." *Advances in Neural Information Processing Systems* 32 (2019).

Some works 📄

Geometric Signatures of Compositionality Across a Language Model's Lifetime (Jin Hwa Lee, Thomas Jiralerspong, Lei Yu, Yoshua Bengio, Emily Cheng**)**

- Nonlinear ID tends to capture deeper semantic / compositional structure
- Linear dimensionality tends to correlate more with superficial / input complexity

The [quality$_1$.ADJ] [nationality$_1$.ADJ] [job$_1$.N] [action$_1$.V] the [size$_1$.ADJ] [texture.ADJ] [color.ADJ] [animal.N] then [action$_2$.V] the [size$_2$.ADJ] [quality$_2$.ADJ] [nationality$_2$.ADJ] [job$_2$.N].

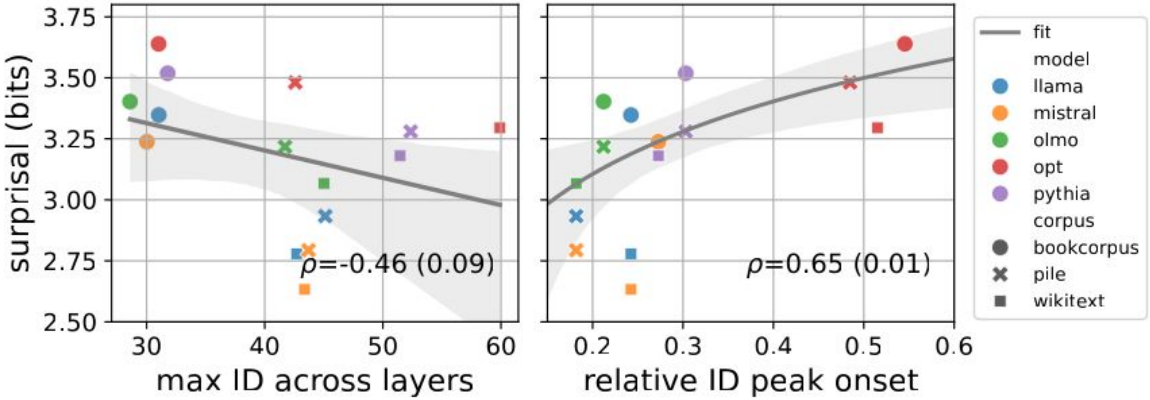
Sentences 17 tokens long, 12 semantic categories and uniformly sample a 50-word vocabulary for each category, categories' vocabularies are disjoint. During data generation, k-grams are independently sampled, which constrains the degrees of freedom in each sentence.

Some works

Emergence of a High-Dimensional Abstraction Phase in Language Transformers (Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, Marco Baroni**)**

- Across multiple pre-trained transformer LMs and diverse datasets, there is a **phase (layer region)** in which the representations reach a peak in intrinsic dimensionality: this corresponds to the first *full linguistic abstraction* of the input

-The **earlier onset** of this high-dimensional abstraction phase correlates with **better language modeling performance**. In other words, models that “reach abstraction earlier” tend to be stronger.



(Left): Surprisal negatively correlates to maximum ID with Spearman ρ = −0.46, p = 0.09, meaning that higher ID indicates better LM performance. (Right): Surprisal positively correlates to ID peak onset, ρ = 0.65, p = 0.01, meaning that an earlier ID peak indicates better LM performance.

Some works

The Representation Landscape of Few-Shot Learning and Fine-Tuning in Large Language Models (Diego Doimo, Alessandro Serra, Alessio Ansuini, Alberto Cazzaniga**)**

- Both ICL and SFT show a *sharp transition* around the middle layers: before that, the geometry / landscape is relatively smooth or semantically organized; after, it shifts to more task-specific clustering.

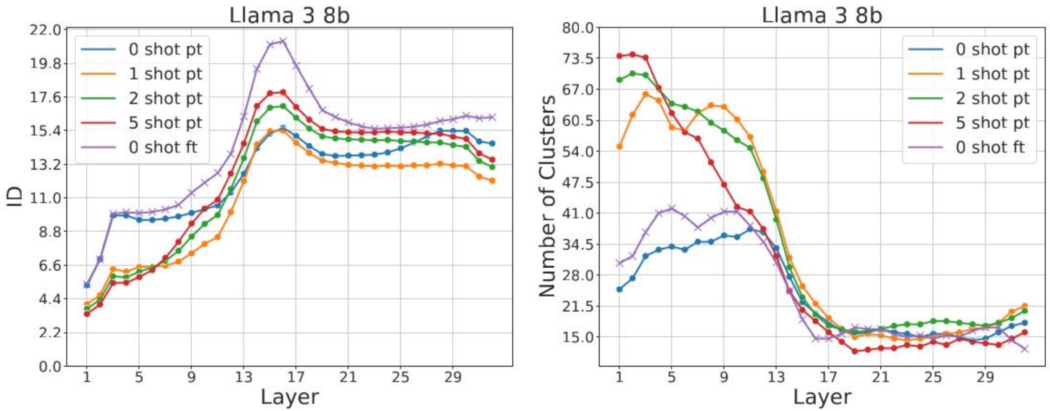
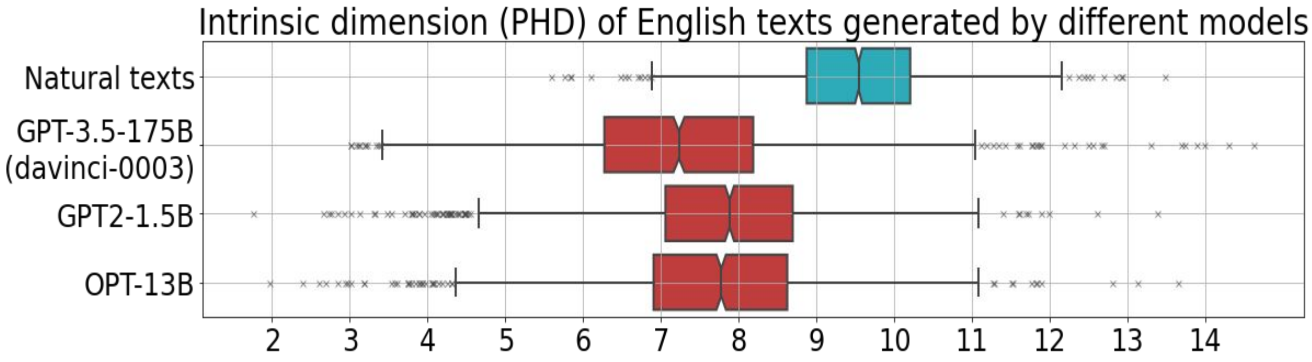


Figure shows the ID (left), the number of density peaks (center), and the fraction of core points (right) for the last-token representation of Llama3-8b for an increasing number of few-shots and fine-tuned models. The three quantities change in the proximity of layer 17 in a two-phased fashion.

Some works 📄

Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts (Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, Evgeny Burnaev**)**

- For several alphabetic languages, **human-written texts** tend to have an average ID ≈ 9, while **AI-generated texts** have ~1.5 lower ID on average



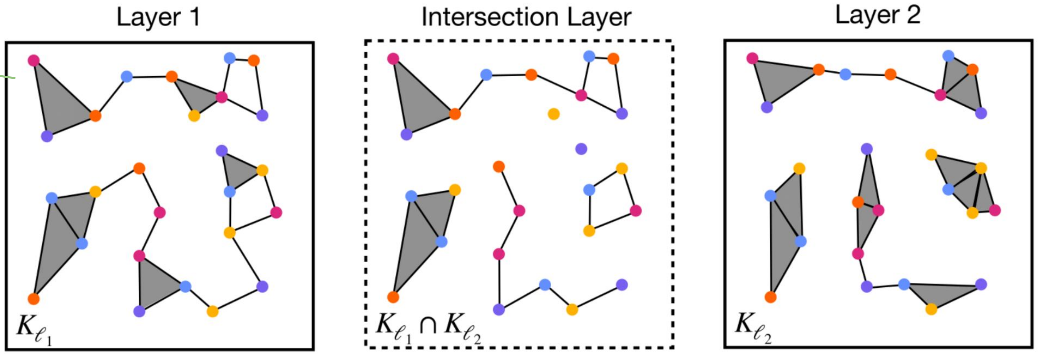
Boxplots of PHD distributions for different generative models in comparison to human-written text on Wikipedia data. Embeddings are obtained from RoBERTa-base.

2.3

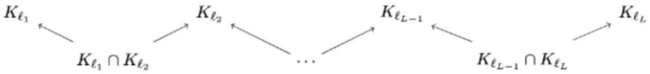
■ Dynamic approaches

Topological Data Analysis

Simplex



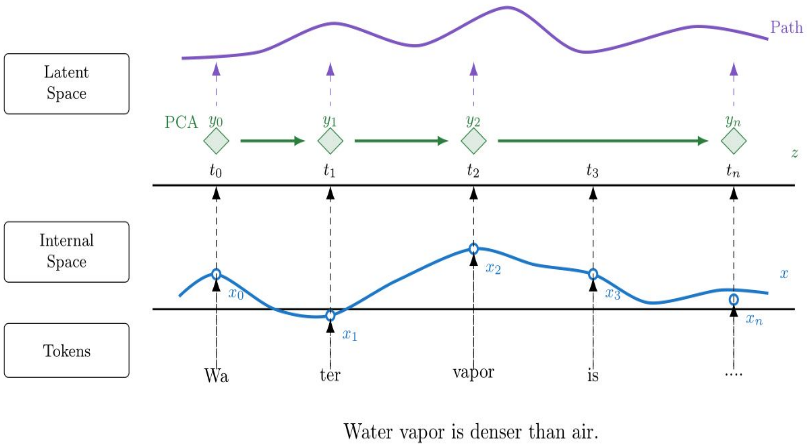
Third step: Intersection Layers



Persistent homology → Zigzag persistent homology

Gardinazzi, Yuri, et al. "Persistent topological features in large language models." *arXiv preprint arXiv:2410.11042* (2024).

Continuous dynamical system



ODE solvers

$$z(t) = z(0) + \int_0^t f(s, z(s); \theta_f) ds$$

$$\text{with} \quad z(0) = h(\boldsymbol{y}; \theta_h),$$

A NN is just a function!

Since we have a continuous-time system, standard backpropagation cannot be directly applied

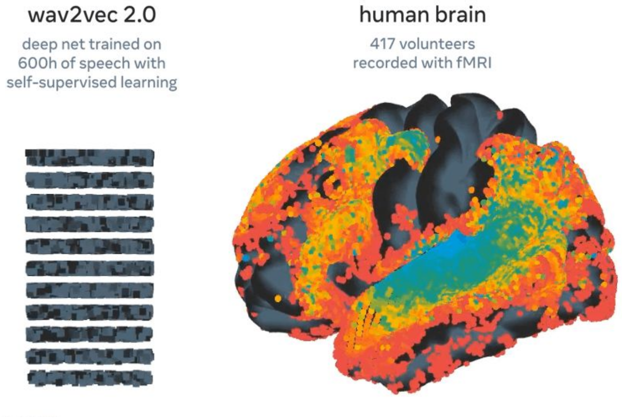
Adjoint method $a(t) = \frac{\partial L}{\partial z(t)}.$

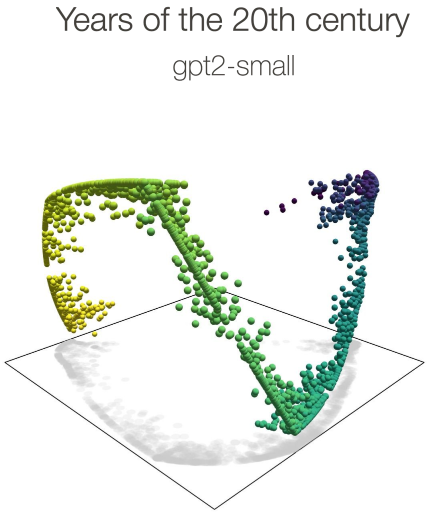
3

What we have done

Neural manifold

Analogies?





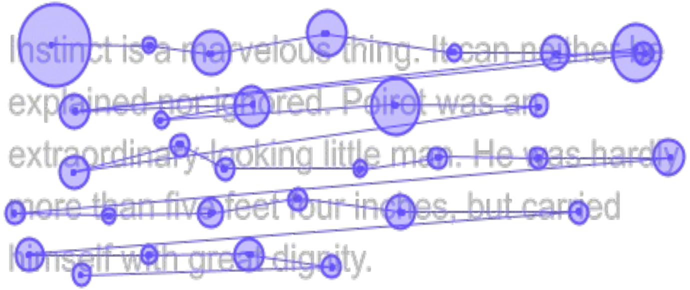
1 - From Human Reading to NLM Understanding: Evaluating the Role of Eye-Tracking Data in Encoder-Based Models

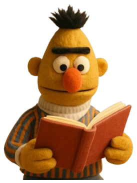
- *"How does human-like learning (e.g., eye-tracking) affect geometry and attention?"*
- *"How do fine-tuning strategies reshape geometry, and are these changes permanent or task-specific?"*

"*From Human Reading to NLM Understanding: Evaluating the Role of Eye-Tracking Data in Encoder-Based Models*". In Proceedings of the Association for Computational Linguistics: ACL 2025. Dini L., Domenichelli L, Brunato D., Dell'Orletta F. (2025).

Eye-tracking data

We used the English section of the **GECO** corpus, that contains eye-tracking data for **12 users** reading a novel by Agatha Christie. **<u>We treat users separately!</u>**





WORD	FFD	TRT	FRNF	NFIX	FRD
The	95	381	1	2	95
intense	54	828	1	3	54
interest	333	565	1	2	333
aroused	78	428	1	3	78
in	154	154	1	1	154
the	165	165	1	1	165

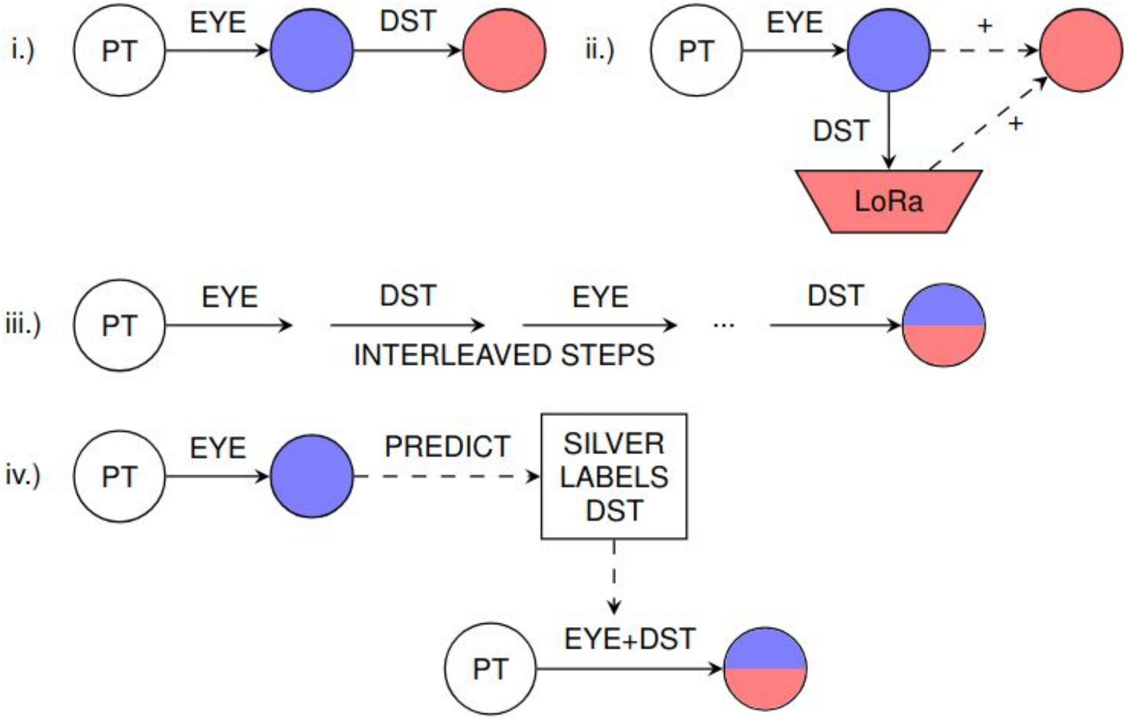
Injection strategies

i.) Intermediate finetuning

ii.) Finetuning with LoRa adapters

iii.) Multi-task finetuning with interleaved steps

iv.) Multi-task finetuning with eye-tracking silver labels



Results – 1

Fine-tuning	Downstream Task										
	COLA	COMP	MNLI M/MM	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	0.56	0.90	0.88 / 0.88	0.90	0.93	0.90	0.70	0.92	0.91	0.56	0.82
INT-LAST3	0.25	0.88	0.70 / 0.71	0.80	0.82	0.81	0.54	0.88	0.81	0.56	0.71
INT-LAST2	0.15	0.85	0.62 / 0.64	0.77	0.75	0.77	0.53	0.86	0.74	0.56	0.66
INT-CLF	0.00	0.70	0.43 / 0.44	0.75	0.61	0.61	0.50	0.76	0.12	0.56	0.50
LORA	0.41	0.87	0.85 / 0.85	0.80	0.91	0.86	0.49	0.93	0.88	0.55	0.76
MT-IL	0.53	0.91	0.83 / 0.83	0.90	0.92	0.88	0.75	0.93	0.90	0.52	0.81
MT-SILV	0.51	0.91	0.88 / 0.87	0.88	0.93	0.90	0.60	0.93	0.91	0.50	0.76
DST-ONLY	0.60	0.91	0.88 / 0.88	0.90	0.93	0.90	0.77	0.93	0.90	0.56	0.83

Intermediate full finetuning (**INT-FULL**) and the two Multi-task approaches, specially the one with interleaved steps (**MT-IL**) generally preserve performances on downstream-tasks.

Results – 2

Fine-tuning	Attention correlation (last layer)										
	COLA	COMP	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	0.19	**0.35**	0.05	0.16	0.06	0.08	0.12	0.04	0.09	0.18	0.13
INT-LAST3	**0.29**	**0.28**	0.24	**0.31**	0.16	**0.29**	0.26	0.21	0.20	0.23	0.25
INT-LAST2	**0.28**	0.26	0.19	**0.28**	**0.30**	0.24	**0.28**	**0.29**	**0.31**	**0.28**	0.27
INT-CLF	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**	**0.29**
LORA	**0.27**	0.22	0.13	0.20	0.13	0.20	**0.32**	0.16	0.21	**0.30**	0.21
MT-IL	0.26	0.23	0.22	**0.27**	0.16	0.21	**0.27**	0.20	**0.27**	**0.28**	0.24
MT-SILV	0.25	0.11	**0.28**	0.15	**0.31**	0.23	**0.33**	**0.31**	0.14	**0.27**	0.24
DST-ONLY	0.06	0.08	0.05	0.01	0.07	0.03	0.02	0.07	0.11	0.12	0.08

Overall all methods increase correlation coefficients, specially intermediate finetuning (excluding INT-FULL), followed by multitask approaches.

Results – 3

F-T	Linear ID										
	COLA	COMP	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	**127**	**89**	191	185	242	11	161	**4**	**32**	127	**117**
INT-LAST3	173	135	194	162	**148**	154	**154**	92	142	154	151
INT-LAST2	162	148	166	160	160	153	157	142	158	158	157
INT-CLF	160	160	160	160	160	160	160	160	160	160	160
LORA	184	144	310	**158**	279	256	166	202	146	163	201
MT-IL	232	154	**110**	179	228	88	155	251	228	152	178
MT-SILV	249	209	233	268	251	207	206	221	264	209	232
DST-ONLY	289	249	249	249	249	**3**	278	**4**	249	**16**	186
BASE	297										–
EYE-ONLY	160										–

F-T	IsoScore* $\times 10^3$										
	COLA	COMP	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STSB	WNLI	AVG
INT-FULL	**0.74**	**1.19**	2.75	15.59	3.03	**0.35**	5.95	0.88	**0.71**	9.74	4.09
INT-LAST3	7.92	2.96	7.40	6.79	**2.10**	3.53	4.36	**0.69**	3.46	5.00	4.42
INT-LAST2	5.89	3.78	7.45	5.05	5.58	4.08	4.35	3.24	5.77	4.69	4.99
INT-CLF	4.99	4.99	4.99	4.99	4.99	4.99	4.99	4.99	4.99	4.99	4.99
LORA	11.26	5.36	30.23	8.47	11.34	28.62	6.01	9.99	2.72	5.27	11.93
MT-IL	4.34	5.02	**1.39**	**2.71**	4.06	1.07	**2.52**	5.83	4.66	3.69	**3.53**
MT-SILV	17.38	10.76	8.28	12.00	11.89	6.57	10.14	11.26	21.56	11.97	12.18
DST-ONLY	6.53	35.94	4.58	15.08	4.69	0.40	28.03	1.17	11.14	**0.27**	10.78
BASE	28.69										–
EYE-ONLY	4.97										–

In most tasks, we observe that eye-tracking data injection yields larger reductions in isotropy and intrinsic dimensionality than standard finetuning, yet preserves downstream performance.

2 - The Role of Eye-Tracking Data in Encoder-Based Models: an In-depth Linguistic Analysis

“Can geometry capture and distinguish linguistic features, and is this consistent across languages?”

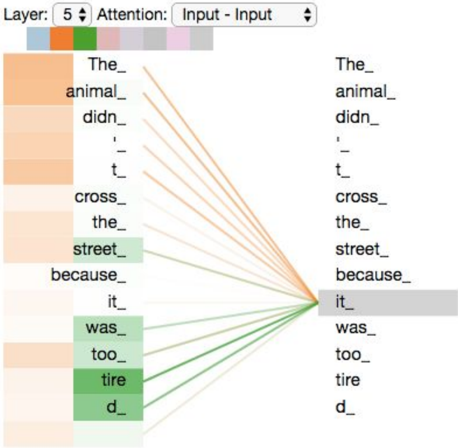
“*The Role of Eye-Tracking Data in Encoder-Based Models: an In-depth Linguistic Analysis*”. In Proceedings of the Italian Association for Computational Linguistics: CLIC-it 2025. Domenichelli L, Dini L., Brunato D., Dell’Orletta F. (2025).

Motivations

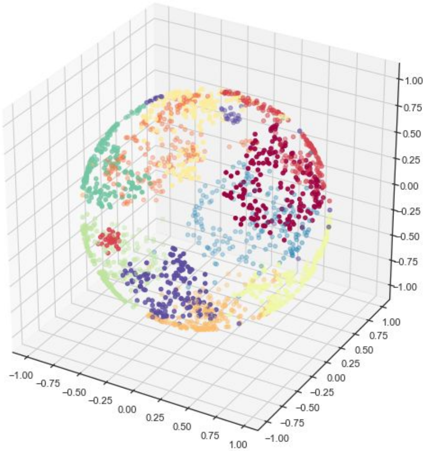
Why eye-tracking in NLP?

- Neural Language Models are powerful but hard to interpret.
- Cognitive signals (like eye-tracking data) offer insight into human language processing.
- **Goal**: study the impact of eye-tracking data injection on the way NLMs build words representations.

Attention patterns



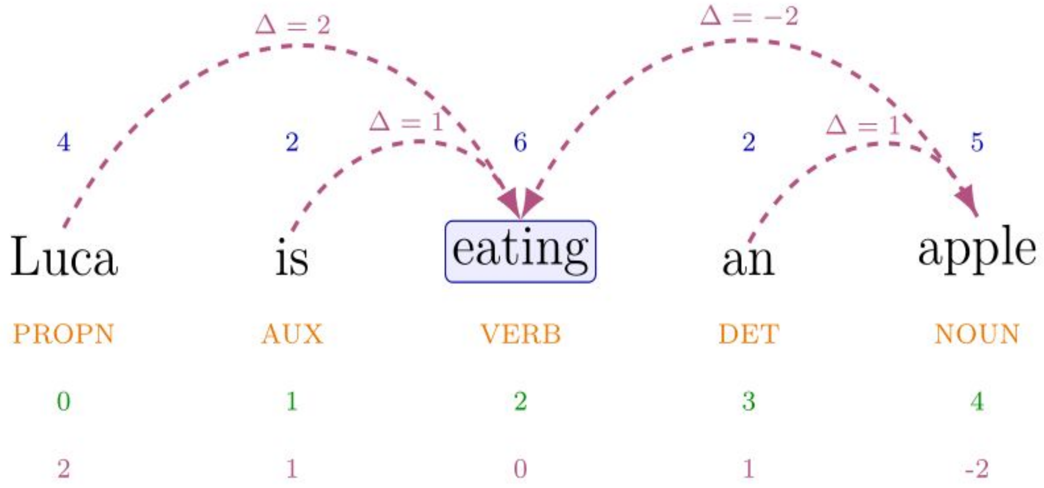
Embedding space



Linguistically informed approach

To enable a more fine-grained analysis of how ET fine-tuning affects word representations, we condition our evaluation on linguistic features extracted from UD treebanks:

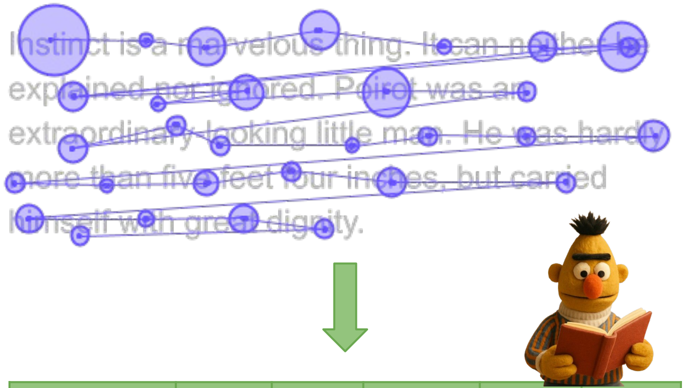
- Word length
- Part-of-Speech
- Position in sentence
- Distance from syntactic head

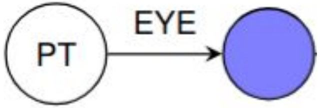


Eye-tracking data

Eye-tracking data are measurement of eye-movements, in this case collected during **reading**.

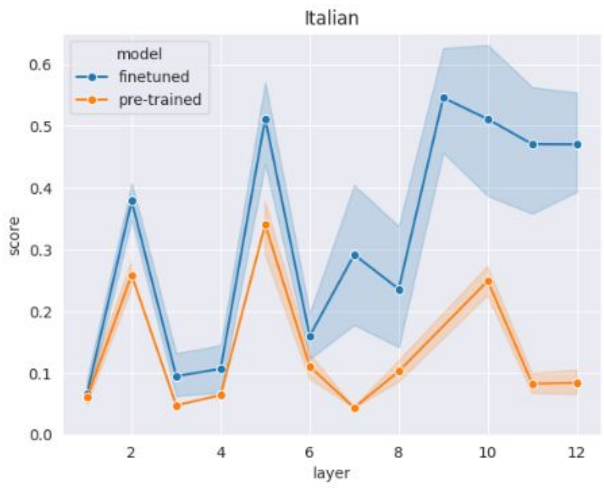
We used the **English** and **Italian** sections of the MECO eye-tracking corpus.





WORD	FFD	TRT	FRNF	NFIX	FRD
The	95	381	1	2	95
intense	54	828	1	3	54
interest	333	565	1	2	333
aroused	78	428	1	3	78
in	154	154	1	1	154
the	165	165	1	1	165

Attention correlation



As shown by literature, fine-tuning on eye-tracking increases correlations between model attention (attention weights) and human attention (TRT).

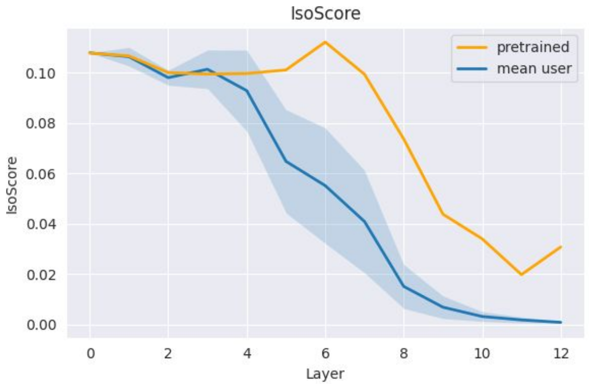
Representation space

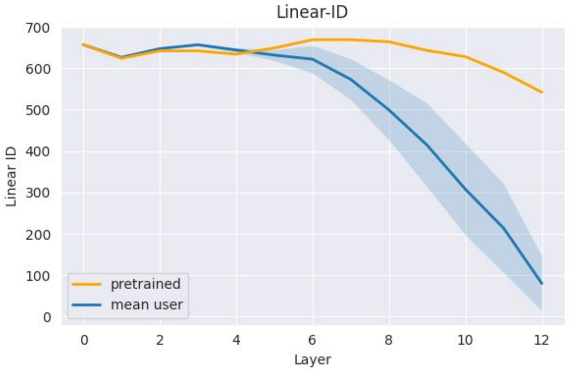
What we know already:

- Embeddings from NLM tend to be anisotropic
- They need way less dimension that their ambient space

After ET fine-tuning...

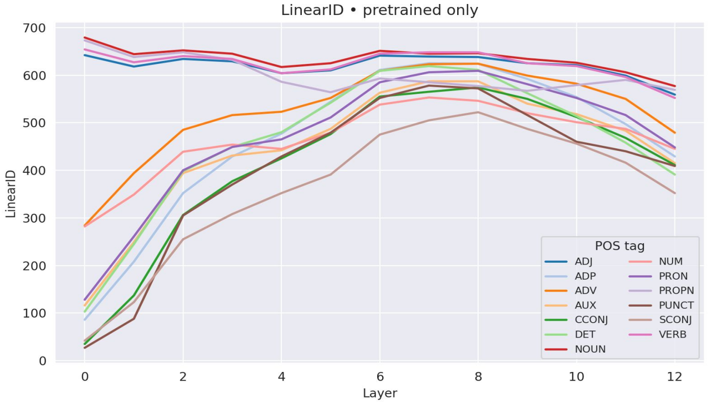
- Representations become even more anisotropic as depth increases!
- Representations have less degrees of freedom as depth increase!

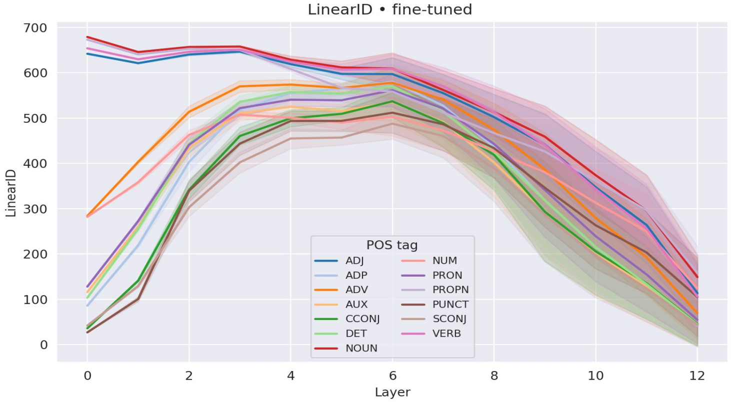




In line with previous studies!

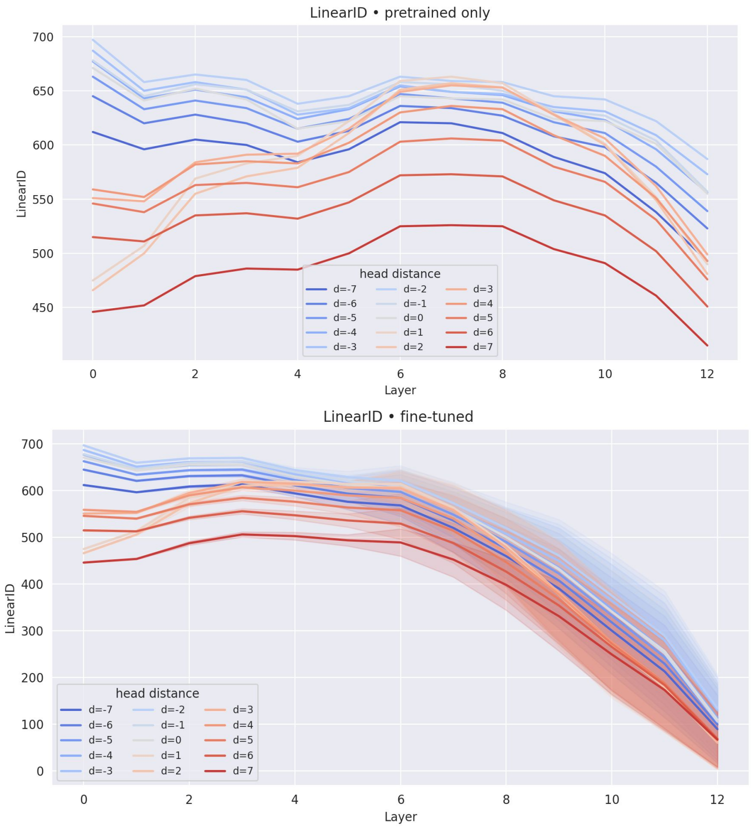
Representation space – POS classes





- Fine-tuning further **anisotropize** and **reduces dimensions**.
- **Content words** (NOUN, PROPN,VERB) tend to **require more dimensions** uniformly in space.

Representation space – Head Distance classes



- Notable asymmetry **already in the pre-trained model** based on the position of the dependent.
- Related to closed functional words but also other unknown effects.

Some statistics

POS	tokens	TTR	Top UD relation label	Top UD relation share	Head dist. val	Head dist. share	Span length val	Span length share	Head arity val	Head arity share	Left of head share	examples
NOUN	21.38	15.90	nmod	29.67	3.00	29.29	5.00	17.12	2.00	39.57	17.64	anni , presidente , parte
ADP	16.40	0.28	case	91.07	2.00	51.39	2.00	39.60	0.00	99.64	98.39	di , in , a
PUNCT	11.94	0.08	punct	100.00	1.00	25.21	1.00	99.74	0.00	100.00	21.25	, , ., "
DET	10.82	0.55	det	92.27	1.00	85.33	2.00	61.64	0.00	99.92	99.71	il, la , un
VERB	9.09	33.21	root	38.23	0.00	38.23	8.00	17.80	3.00	26.88	4.58	ha , è , hanno
ADJ	7.14	27.09	amod	84.40	1.00	75.81	8.00	18.30	0.00	75.56	26.53	primo , prima , nuovo
PROPN	5.23	38.22	nmod	37.92	1.00	33.56	6.00	20.66	1.00	37.53	13.22	italia , shakespeare , balzac
AUX	4.23	1.91	aux	52.11	1.00	70.09	1.00	28.38	0.00	99.91	95.11	è , sono , ha
ADV	4.12	5.47	advmod	91.00	1.00	48.10	3.00	31.74	0.00	84.86	79.64	non , più, anche
PRON	3.65	1.65	nsubj	29.29	1.00	42.91	3.00	38.66	0.00	68.27	78.19	che, si, chi
CCONJ	2.97	0.49	cc	99.92	1.00	37.50	1.00	82.18	0.00	99.99	99.91	e, o, ma
NUM	1.76	23.56	nummod	67.70	1.00	46.54	4.00	28.27	0.00	64.92	50.24	due, 1, tre
SCONJ	1.11	1.60	mark	91.71	1.00	20.29	3.00	41.87	0.00	99.47	92.20	che , se ,

ITALIAN

Different statistics, similar scores

POS	tokens	TTR	Top UD relation label	Top UD relation share	Head dist. val	Head dist. share	Span length val	Span length share	Head arity val	Head arity share	Left of head share	examples
NOUN	17.24	20.54	obj	20.23	2.00	23.75	4.00	18.33	2.00	27.81	29.76	time, people, way
PUNCT	11.35	0.42	punct	100.00	1.00	26.80	1.00	96.14	0.00	100.00	34.27	. , , , "
VERB	11.34	16.01	root	30.91	0.00	30.91	4.00	27.70	3.00	27.54	8.40	have, get, know
PRON	9.37	0.76	nsubj	55.28	1.00	50.13	2.00	26.52	0.00	89.92	79.57	i, you, it
ADP	9.01	0.66	case	92.33	1.00	39.82	2.00	64.49	0.00	99.09	92.48	of, in, to
DET	8.27	0.22	det	96.52	1.00	57.66	3.00	60.41	0.00	98.18	98.53	the, a, this
ADJ	6.49	17.05	amod	68.24	1.00	56.63	4.00	22.45	0.00	65.24	70.36	other, new, good
AUX	5.91	0.78	aux	51.22	1.00	47.30	2.00	31.25	0.00	98.13	96.61	is, was, be
PROPN	5.75	33.99	compound	20.15	1.00	35.26	5.00	17.75	0.00	46.26	42.67	bush, us, al
ADV	5.09	7.22	advmod	93.20	1.00	54.05	4.00	36.17	0.00	85.62	69.22	so, when, just
CCONJ	3.40	0.32	cc	98.25	1.00	43.53	3.00	86.27	0.00	99.26	99.49	and, but, or
PART	2.18	0.17	mark	76.42	1.00	82.29	2.00	76.85	0.00	99.15	98.51	to, not, too
SCONJ	1.96	1.63	mark	98.43	2.00	24.30	2.00	40.58	0.00	98.56	98.51	that, if, as

ENGLISH

4

Ongoing work

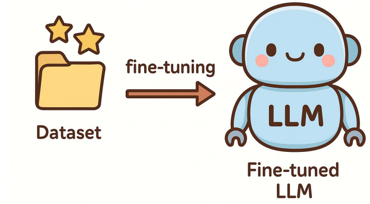
The impact of data ordering on the pretraining phase

"Does curriculum learning reshape geometry, and can early geometry predict performance?"

Evaluation

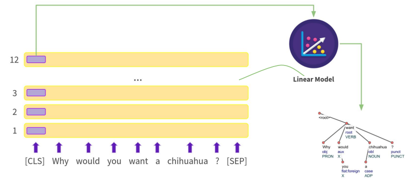
Evaluation strategies

Fine-tuning



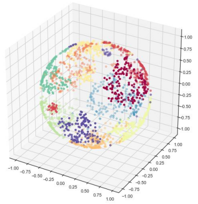
Complexity, POS-tagging, Sentiment Analysis

Probing



Profiling-UD

Representation space



Isoscore (Isotropy)

Linear ID (@99%)

10.000 sentences not seen already

Ordering strategies

Ordering strategies:

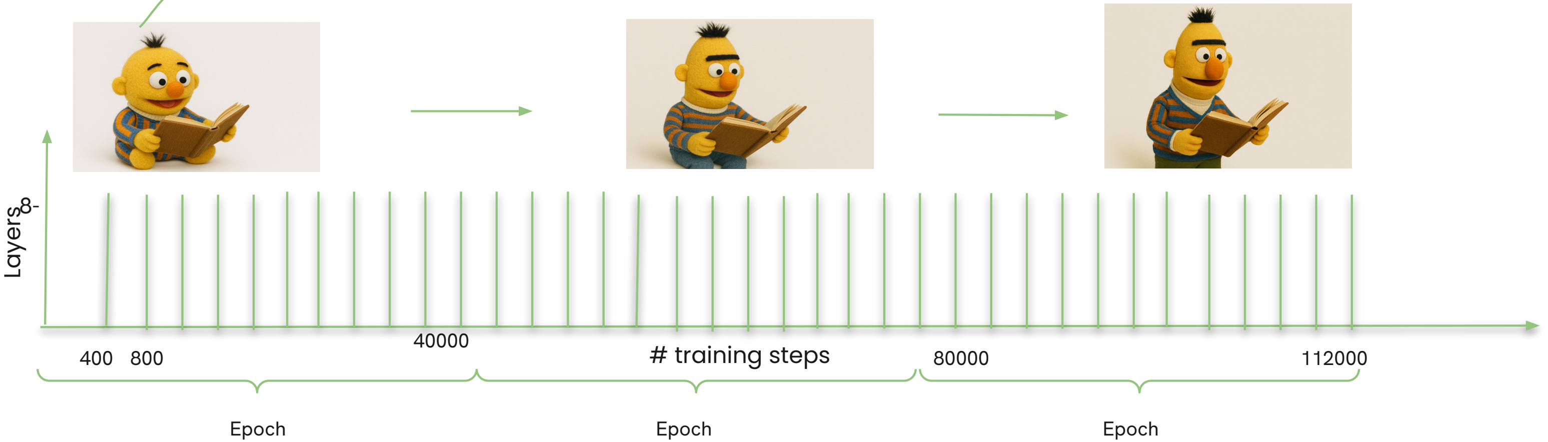
- Sentences’ length
- Gulpease
- ReadIT
- 3 Random orderings

Gulpease, ReadIT } Linguistically Motivated ordering

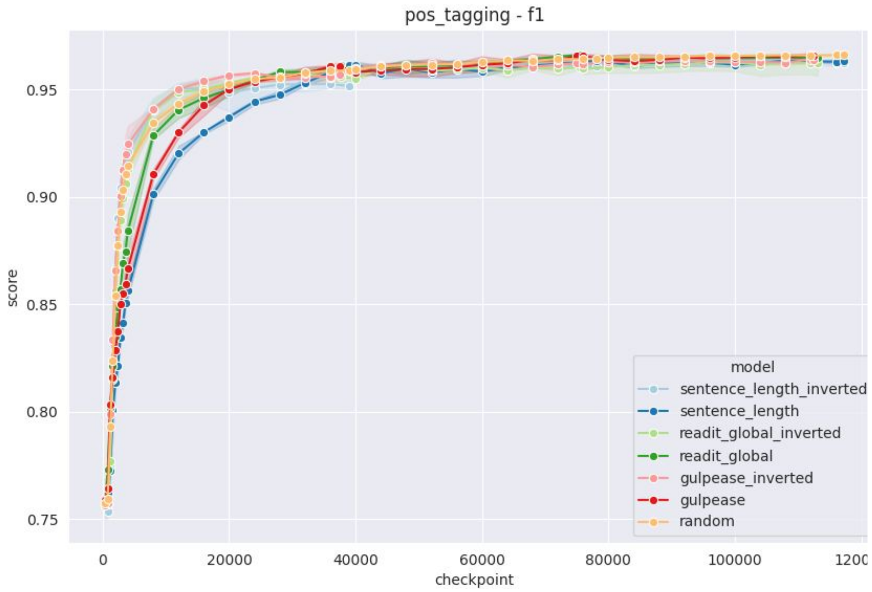
... and anti-curriculum!

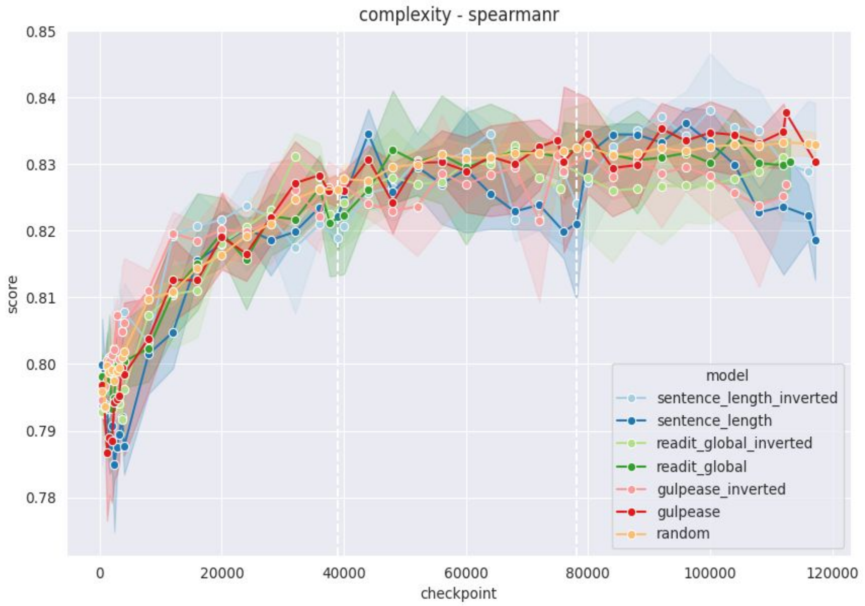
BERT-medium

10000000 sentences from Wikipedia.



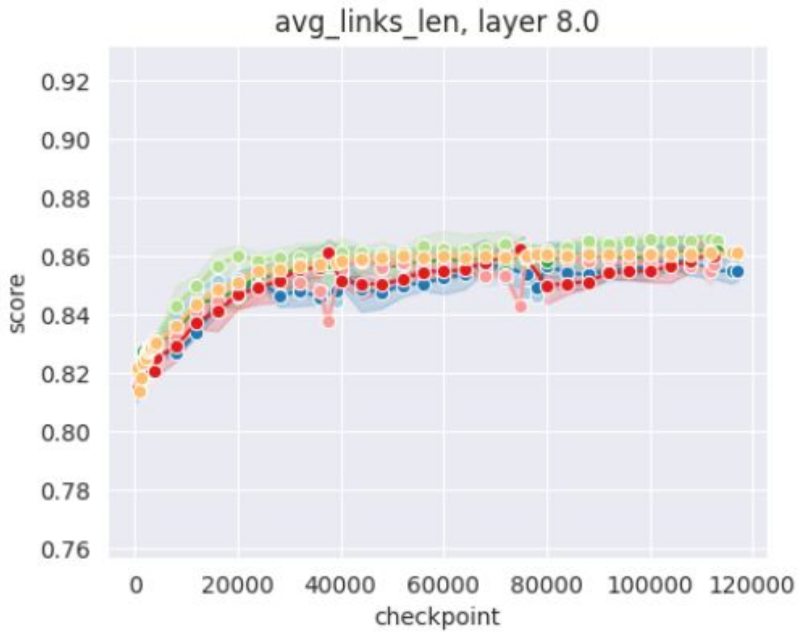
Finetuning 🎯

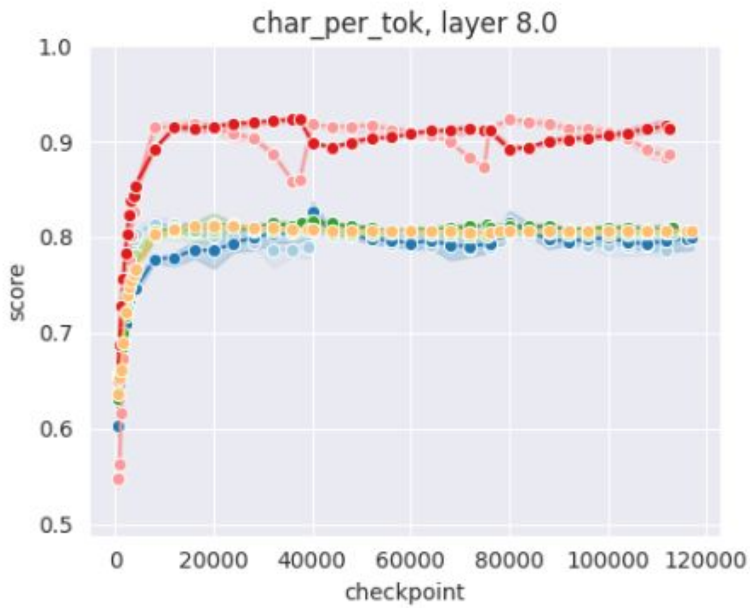




Qualitatively similar, only difference is the “speed of convergence”.

Probing 🎯

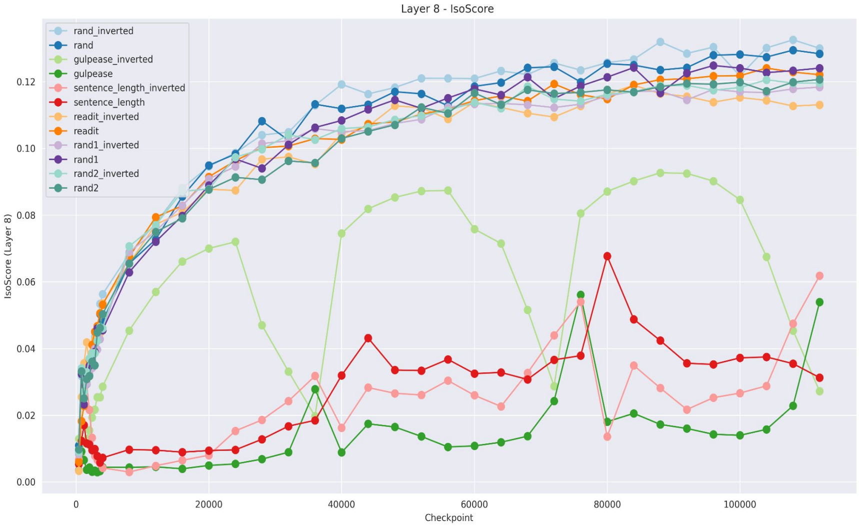




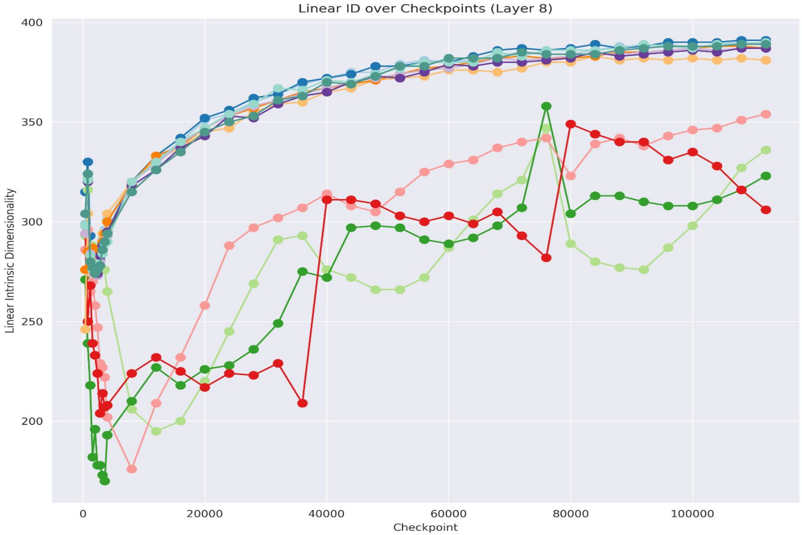
Most features have plots similar to the one on the left, with some exceptions...

Representation space 🎯

Isotropy



Linear ID



Qualitatively different behaviour between models with curriculum-ordered and not ordered data!

Future work 🔭

- How will this research field evolve? ⟶ Unknown, the field is blooming now!

🤓 About us....

- Keep working on our curriculum learning project.

- Expand our second paper and work on a "Linguistic Profiling of Geometric Spaces" ✍️.

- Visiting period abroad while working on my PhD proposal💃.

Carried out activities 💪

Passed exams:

- Topological Data Analysis (16 CFU)
- Statistical Learning and Large Data 1 & 2 (40 CFU) → 80 CFU TOT ✅ EXAMS DONE!
- Predictive Models for Time Series Data (24 CFU)

Carried out activities 💪

Passed exams:

- Topological Data Analysis (16 CFU)
- Statistical Learning and Large Data 1 & 2 (40 CFU)
- Predictive Models for Time Series Data (24 CFU)

→ 80 CFU TOT ✅ EXAMS DONE!

PhD Schools:

- HPLT & NLPL 2025 Winter School in Oslo 🇳🇴
- Learning over Time Spring School in Siena 🏇
- AI & Society 2025 Summer School in Pisa

→ 60 H TOT ✅ MAYBE DONE!

Research seminars attended:

- Lectures on Computational Linguistics in Milan

Carried out activities 💪

Passed exams:

- Topological Data Analysis (16 CFU)
- Statistical Learning and Large Data 1 & 2 (40 CFU)
- Predictive Models for Time Series Data (24 CFU)

→ 80 CFU TOT ✅ EXAMS DONE!

PhD Schools:

- HPLT & NLPL 2025 Winter School in Oslo 🇳🇴
- Learning over Time Spring School in Siena 🏇
- AI & Society 2025 Summer School in Pisa 

→ 60 H TOT ✅ MAYBE DONE!

Research seminars attended:

- Lectures on Computational Linguistics in Milan 

Written papers:

- "*From Human Reading to NLM Understanding: Evaluating the Role of Eye-Tracking Data in Encoder-Based Models*". In Proceedings of the Association for Computational Linguistics: ACL 2025. Dini L., Domenichelli L, Brunato D., Dell'Orletta F. (2025).
- "*The Role of Eye-Tracking Data in Encoder-Based Models: an In-depth Linguistic Analysis*". In Proceedings of the Italian Association for Computational Linguistics: CLIC-it 2025. Domenichelli L, Dini L., Brunato D., Dell'Orletta F. (2025).

Questions?

Thank you for your attention!

References 📚

- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). “Attention is all you need”. In: Advances in neural information processing systems 30.
- Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2020). “A Primer in BERTology: What We Know About How BERT Works”. In: Transactions of the ACL.
- Miaschi, Alessio and Felice Dell’Orletta (2020). “Contextual and non-contextual word embeddings: an in-depth linguistic investigation”. In: Proceedings of the 5th Workshop on Representation Learning for NLP, pp. 110-119
- Hewitt, John and Christopher D. Manning (June 2019). “A Structural Probe for Finding Syntax in Word Representations”. In: 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). Ed. by Jill
- Rudman, William, Nate Gillman, Taylor Rayne, and Carsten Eickhoff (May 2022a). “IsoScore: Measuring the Uniformity of Embedding Space Utilization”. In: Findings of the Association for Computational Linguistics: ACL 2022. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 3325–3339.
- Team, Anthropic Interpretability (2024). The Engineering Challenges of Scaling Interpretability. Anthropic Research Blog. https://www.anthropic.com/research/engineering-challenges-interpretability
- Bengio, Yoshua, J´erˆome Louradour, Ronan Collobert, and Jason Weston (2009). “Curriculum learning”. In: Proceedings of the 26th annual international conference on machine learning, pp. 41–48.
- Cai, Xingyu, Jiaji Huang, Yuchen Bian, and Kenneth Church (2021). “Isotropy in the contextual embedding space: Clusters and manifolds”. In: International conference on learning representations.
- Ansuini, Alessio, Alessandro Laio, Jakob H Macke, and Davide Zoccolan (2019). “Intrinsic dimension of data representations in deep neural networks”. In: Advances in Neural Information Processing Systems 32
- Cheng, Emily, et al. "Emergence of a High-Dimensional Abstraction Phase in Language Transformers." *The Thirteenth International Conference on Learning Representations*.
- Modell, Alexander, Patrick Rubin-Delanchy, and Nick Whiteley. "The Origins of Representation Manifolds in Large Language Models." *arXiv e-prints* (2025): arXiv-2505.
- Rajaee, Sara and Mohammad Taher Pilehvar (2021). “How Does Fine-tuning Affect the Geometry of Embedding Space? A Case Study on Isotropy”. In: Findings of EMNLP.
- Doimo, Diego, et al. "The representation landscape of few-shot learning and fine-tuning in large language models." *Advances in Neural Information Processing Systems* 37 (2024): 18122-18165.
- Tulchinskii, Eduard, et al. "Intrinsic dimension estimation for robust detection of ai-generated texts." *Advances in Neural Information Processing Systems* 36 (2023): 39257-39276.
- Valeriani, Lucrezia, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga (2023). “The geometry of hidden representations of large transformer models”. In: Advances in Neural Information Processing Systems 36, pp. 51234–51252
- Elhage, et al., "A Mathematical Framework for Transformer Circuits", Transformer Circuits Thread, 2021.
- Gardinazzi, Yuri, et al. "Persistent topological features in large language models." *arXiv preprint arXiv:2410.11042* (2024).
- Li, Qing, et al. "HD-NDEs: Neural Differential Equations for Hallucination Detection in LLMs." *arXiv e-prints* (2025): arXiv-2506
- Dini, Luca, et al. "From Human Reading to NLM Understanding: Evaluating the Role of Eye-Tracking Data in Encoder-Based Models." *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 2025.
- Dini, Luca, et al. "In the eyes of a language model: A comprehensive examination through eye-tracking data." *Neurocomputing* (2025): 130617.