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

Lucia Domenichelli[1,2,*], Luca Dini[1,2], Dominique Brunato[1] and Felice Dell'Orletta[1]

[1]ItaliaNLP Lab, Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), Pisa, Italy
[2]University of Pisa, Pisa, Italy

**Abstract**

This paper falls within ongoing research aimed at enhancing the human interpretability of neural language models by incorporating physiological data. Specifically, we leverage eye-tracking data collected during reading to explore how such information can guide model behavior. We train a multilingual encoder model to predict eye-tracking features from the Multilingual Eye-tracking Corpus (MECO) and analyze the resulting shifts in model attention patterns, focusing on how attention redistributes across linguistically informed categories such as part of speech, word position, word length, and distance from the syntactic head after fine-tuning. Moreover, we test how this attention shift impacts the representation of the interested words in the embedding space. The study covers both Italian and English, enabling a cross-linguistic perspective on attention and representation shifts in multilingual encoders grounded in human reading behavior.

**Keywords**

Eye-tracking, Neural Attention, Multilingual models, Embedding space, Interpretability

## 1. Introduction and Motivation

Neural language models (NLMs) now match or even surpass human benchmarks on many NLP tasks, yet the logic behind their predictions remains largely hidden behind billions of parameters. To make these systems more transparent and data-efficient, researchers are increasingly borrowing ideas from cognitive science, grounding both training and evaluation in how people actually learn and process language (e.g. [1, 2, 3]). Among the most informative cognitive signals of human language processing is eye-tracking (ET). Decades of psycholinguistic work show that fixation times, regressions, and skips mirror both early lexical access and later integrative processes underlying text comprehension [4, 5]. Leveraging these signals has already boosted model accuracy on a variety of downstream tasks ranging from core linguistic tasks [6] to more applied tasks like sentiment analysis [7], language proficiency assessment [8], machine reading comprehension [9], while also giving us a new lens on model interpretability. Studies by Sood et al. [10] and Eberle et al. [11] found that transformer attention does not always line up with human gaze, whereas Bensemann et al. [12] and Wang et al. [13] revealed stronger links in specific layers, hinting at a layered correspondence between reading behavior and neural representations. Extending this direction, Dini et al. [14] investigate how injecting reading-related information into NLMs through different fine-tuning strategies on ET data affects their attention patterns, as well as their performance on downstream tasks and representation space. Their findings show that this intermediate process increases the correlation between model attention and human attention and it leads to a compression of the embedding space, without generally degrading performance on downstream tasks.

Building on this foundational framework, this paper aims to further highlight the effects of **incorporating information about human reading behavior in a NLM** from a **linguistically informed perspective**. Specifically, we examine how fine-tuning on eye-tracking signals leads to **shifts in model attention**, and how these shifts affect the **structure of word representations**. To explore this, we extract a set of linguistic features, capturing progressively more complex language phenomena, from the input text and analyze how attention is redistributed across word classes defined by these features. In parallel, we assess how these attention shifts influence the embedding space, both at a global level and within the local representational geometry of specific word classes.

The code for our experiments is publicly available on GitHub.

## 2. Related work

Our study intersects two complementary lines of research within NLMs interpretability. The first investigates ET data as a diagnostic signal to evaluate the alignment between model behavior and human cognitive processing, particularly through the lens of attention mechanisms. The second focuses on analysing model's attention mechanisms (Section 2.2) and representational space (Section 2.3) in relation to linguistic structure.

*CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy*
*Corresponding author.
lucia.domenichelli@ilc.cnr.it (L. Domenichelli)
© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

## 2.1. Eye-tracking and NLMs

In recent years, eye-tracking has emerged as a prominent physiological signal in NLP research due to its affordability and ease of collection compared to methods like fMRI or MEG. Public resources such as the GECO corpus [15], the MECO corpus [16], and the WE-RDD dataset [17] now let researchers probe gaze behaviour at scale across languages and reading paradigms.

Work with these corpora has split in two directions. The former injects gaze-derived features, into neural architectures, typically lifting accuracy on downstream tasks. The latter, which motivates our study, treats ET as a diagnostic for a model's internal workings.

The first systematic comparison came from Sood et al. [18], who matched attention maps from CNNs, LSTMs and Transformers against human fixations. Their findings reveal that while transformers performed the best, they showed the weakest alignment with gaze. Eberle et al. [11] confirmed that even after task-specific fine-tuning, large Transformers stayed distant from human reading patterns. Conversely, Bensemann et al. [12] reported that raw dwell times correlate strongly with the earliest BERT layers, a relation that persists as model size grows. Morger et al. [19] extended the inquiry cross-lingually and found robust correlations, especially for monolingual encoders, between human word-importance rankings and model saliency. Most recently, Wang et al. [20] showed that deeper layers of NLMs once again echo fixation metrics, hinting at a layered, non-monotonic link between model depth and cognitive fidelity.

## 2.2. Model Attention Dynamics

The role of attention mechanisms in NLMs has been a subject of extensive research and debate. While attention weights are often interpreted as providing insight into model reasoning, a growing body of research has questioned their reliability as faithful explanations of model decisions. Some studies suggest that attention can highlight important input elements, yet others argue that attention distributions can be manipulated without significantly affecting predictions, casting doubt on their explanatory power [21, 22]. In response to these concerns, alternative attribution methods have been proposed—such as attention rollout [23] and gradient-based techniques [24]—which aim to better capture the pathways through which information influences predictions. As part of this debate, a parallel line of work has explored whether attention aligns with known linguistic structures, such as syntactic dependencies or PoS categories, offering a complementary perspective on its interpretability. The foundational study by Clark et al. [25] showed that certain attention heads in BERT consistently focus on syntactic phenomena, such as attending to an entity's determiners or subjects attending to their verbs. However, fine-tuning on syntactic or semantic tasks had minimal effect on altering self-attention patterns. Vig and Belinkov [26] conducted a comprehensive analysis of attention head interpretability in GPT-2 using both visualization and quantitative measures. Their results indicate a layer-specific linguistic sensitivity, with different types of linguistic information—such as PoS and syntactic dependencies—being more salient in particular layers. They also found stronger alignment with syntactic dependencies in the model's middle layers. Htut et al. [27] directly evaluated the extent to which attention aligns with gold-standard dependency parses. By computing the correspondence between attention distributions and syntactic head-dependent pairs, they showed that BERT's attention does not systematically reflect syntactic dependency structures, particularly in deeper layers.
Taken together, these studies suggest that while attention mechanisms can exhibit linguistically meaningful behavior in isolated cases—especially in specific layers or individual heads—they do not consistently encode syntactic or morpho-syntactic structure.

## 2.3. Geometry of the embedding space

Transformer models learn a high-dimensional *embedding space* in which every token is represented by a dense vector that encodes both meaning and syntax. A consistent finding is that these vectors occupy only a narrow cone of the space, an *anisotropic* layout sometimes called the representation degradation effect [28, 29, 30]. In NLP, such behaviour is often viewed as harmful because it can hide fine-grained linguistic cues [31, 32, 33]. Yet theory and broader machine-learning evidence show that anisotropy can arise naturally under stochastic gradient descent and may even aid generalization, especially when models project data onto low-dimensional manifolds [34, 35, 36, 37]. In this respect, studying the impact of various fine-tuning objectives and downstream tasks provides important insights into how they shape the geometry of the embedding space [34, 35, 36]. While still relatively limited, a growing body of work has begun to examine the relationship between embedding space properties and linguistic phenomena. For example, Hernandez and Andreas [38] show that linguistic features tend to be encoded in lower-dimensional subspaces in the early layers of both ELMo and BERT and that relational features (like dependency relations between pairs of words) are encoded less compactly than categorical features like part of speech. More recently, Cheng et al. [39] analyzed representation compression in pre-trained language models from both geometric and information-theoretic perspectives. Their findings reveal a strong correlation between these two views and show that the intrinsic geometric dimension of linguistic data is predic-

tive of its coding length under the language model.

To the best of our knowledge, no systematic study has examined how eye-tracking fine-tuning affects attention patterns and the resulting embedding representations across different linguistic phenomena. Moreover, cross-linguistic analyses of these changes following cognitively motivated fine-tuning remain scarce.

# 3. Dataset

For our analysis, we leverage two distinct datasets: the Multilingual Eye-tracking Corpus (MECO) to finetune the model on human gaze modeling and treebanks from the Universal Dependencies (UD) project to extract linguistically motivated features and compute model attention shifts and representation structure induced by fine-tuning on ET data.

## 3.1. Eye-tracking data: The MECO Corpus

MECO [16] is a multilingual collection featuring reading behavior from both native (L1) and second-language speakers across 13 languages. We focus on the L1 subsets for English and Italian, chosen for their typological diversity and data completeness, allowing for a controlled yet cross-linguistic perspective on gaze modeling.

Each participant in MECO read 12 encyclopedic-style texts, covering general knowledge topics. To ensure consistency and limit computational costs, we selected the largest subsets of users who had read the majority of sentences. For Italian, we included 9 participants who read all sentences. For English, since no participant completed the full set, we selected 25 participants who all read the same set of sentences, missing only two in common.

We used five ET features intended to represent early, late and contextual signals of human reading processes: **First Fixation Duration**: the duration of the first fixation landing on the word; **Gaze Duration**: the summed duration of fixations on the word in the first pass, i.e., before the gaze leaves it for the first time; **Total Reading Time**: the cumulative amount of time spent reading a word, capturing both fixations and potential interruptions (e.g., regressions or pauses); **First-run Number of Fixations**: the number of fixations on a word during the first pass; **Total Number of Fixations**: the number of discrete fixations on areas of interest overall.

## 3.2. Universal Dependencies Treebanks

To analyze how model attention weights and embedding space shift following fine-tuning on eye-tracking data, we relied on linguistically annotated corpora from UD treebanks [40]. Specifically, for Italian, we employed the subsection corresponding to the training set of the Italian Stanford Dependency Treebank (ISDT), which contains $\approx 13,000$ sentences drawn from a variety of textual genres. For English, we used the training set of the English Web Treebank (EWT) [41], including $\approx 12,000$ sentences , also multi-genre. UD corpora were chosen due to their gold-standard syntactic and part-of-speech annotations, which provide a reliable foundation for our fine-grained linguistic analyses. Additionally, the cross-linguistically consistent annotation schema offered by UD enables meaningful comparisons across typologically distinct languages.

# 4. Our Approach

We propose a **linguistically informed framework** to investigate the impact of injecting human reading behaviour into a pre-trained NLM, focusing on its effects on attention and word representations. The approach consists of two main stages: first, we fine-tune the model on predicting several ET features; then, we compare the pre-trained and fine-tuned models along three axes: i) Correlation between model attention and human attention; ii) Attention distribution over input tokens; iii) Sentence representations in the embedding space.

To enable a more fine-grained analysis of how ET fine-tuning affects word representations, we condition our evaluations on the following linguistic features extracted from the UD treebanks: **word length** in characters, **part of speech** category, **position** in the sentence, and **distance** from the syntactic head.

For our experiments we used XLM-RoBERTa-base, a 12 layer multilingual encoder-based model. In what follows, we outline the methodological choices and implementation details of our experimental setting.

## 4.1. ET injection into the Model

To inject reading-related information into the model, we leverage the set of eye-tracking features from MECO described in Section 3.1. Unlike most prior work—which typically aggregates eye-tracking data across participants, with few exceptions [42]—we treat each reader individually, conducting experiments separately for each subject. This design choice is motivated by the intrinsic variability observed in reading behavior, even among skilled readers [43, 44, 45], and enables a more accurate modeling of reader-specific dynamics.

After a hyperparameter tuning phase using 5-fold cross-validation, we **fine-tune the model to predict five word-level eye-tracking features**, training a separate model for each individual reader.

Since the MECO dataset provides annotations at the word level, while the model's tokenizer splits some words into subword units, we follow standard practice [46] and

assign eye-tracking features only to the first sub-token of each word, ignoring the rest during training[1].

To examine whether the fine-tuned model develops a more human-like attention pattern, we compute the **correlation between model attention and human attention** before and after fine-tuning. For model attention, we consider the attention weights received by each word when computing the representation of the beginning-of-sentence token (<s>), which is the only token used during the eye-tracking prediction phase and serves as a global summary of the sentence. To account for subword tokenization, we follow the same approach used during fine-tuning and associate attention scores to the first sub-token of each word. As a proxy for human attention, we choose the *Total Reading Time* feature (see Section 3.1). For each reader, we thus compute the correlation between their eye-tracking data and the attention patterns of both the pre-trained and the fine-tuned model across all layers, allowing us to assess whether the latter aligns more closely with human reading behavior.

### 4.2. Assessing the Role of ET fine-tuning on Word Representations

To assess how fine-tuning on ET affects the model's internal dynamics for attention and embedding space, we leverage the linguistic features from the treebanks described in Section 4.

Specifically, to compute the **attention shifts**, for each value of these features, we analyse the amount of attention the corresponding words receive before and after fine-tuning. This allows us to characterize shifts in attention distribution across different linguistic phenomena and across all layers of the models. Firstly, we normalize the attention scores for each sentence (excluding BOS and EOS tokens) so that their sum is 1. Attention shifts are quantified as the percentage change in the average attention received by tokens with a given feature value, before fine-tuning. A positive shift indicates increased attention to these tokens after fine-tuning, while a negative shift reflects a decrease. This allows us to identify which linguistic categories gain or lose prominence after incorporating eye-tracking supervision.

To analyze the **shifts in the embedding space**, we rely on two complementary metrics. *(i) IsoScore* [47] offers a scale-invariant measure of isotropy: lower scores indicate that the embedding variance is concentrated along fewer directions, pointing to a more anisotropic space. *(ii) Linear Intrinsic Dimensionality (Linear-ID)* [48] estimates the dimensionality of the smallest linear subspace that captures the embeddings, providing a proxy for their geometric complexity.

[1]The fine-tuning is run for 50 epochs, using a learning rate of $5e-05$, a weight decay of 0.01, and a warm-up ratio of 0.05.

The two metrics were computed on the first sub-token of each word in the UD treebanks. In line with the other analyses, we compare the embedding spaces of the pre-trained and fine-tuned models to assess whether ET fine-tuning leads to more compact or more isotropic representations, as reflected by changes in these metrics.

All reported scores are first computed for each user individually and then averaged across all users.

## 5. Results

### 5.1. Correlation between model and human attention

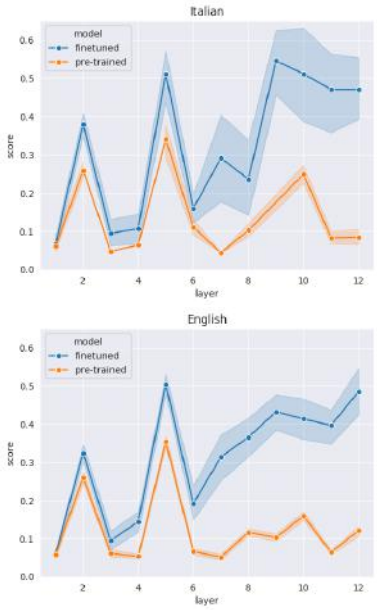


**Figure 1:** Correlation between model attention and human attention (p-value < 0.05).

As a first evaluation step, we computed the correlation between human attention and model attention, both before and after fine-tuning on eye-tracking data. As we are interested in the strength rather than the direction of the association, we considered the absolute values of the correlation coefficients. For the fine-tuned models, we computed the correlation between the model's attention weights and the Total Reading Time of the specific user on which each model was fine-tuned. For the pre-trained model, which is not finetuned to any individual reader, we calculated the correlation between its attention weights and the Total Reading Time of each user independently, and subsequently averaged the resulting coefficients. Figure 1 reports the comparison of Spearman correlation coefficients, averaged across all users.

In line with results reported in [14, 49], **fine-tuning on ET data consistently leads to stronger correlation coefficients between model and human attention, particularly in the deeper layers** of the model. This effect is evident in both Italian and English. The overall patterns are remarkably similar across the two languages, although the correlation scores for Italian are slightly higher on average.

## 5.2. Analysis of the Attention Shifts

This section reports the analysis of the attention shifts induced by fine-tuning on ET data. We grouped tokens into classes for the values of the linguistic features detailed in Section 4. To enhance readability and interpretability, for each linguistic feature we visualised only the most representative values. Rather than applying a strict frequency threshold, we heuristically excluded rare or degenerate cases (e.g., for token length, extremely long tokens such as URLs), retaining typical and frequent values that better reflect standard linguistic patterns. Each figure also includes an "AVG" column summarizing the average shift across all layers, offering a high-level view of the attention reallocation patterns.

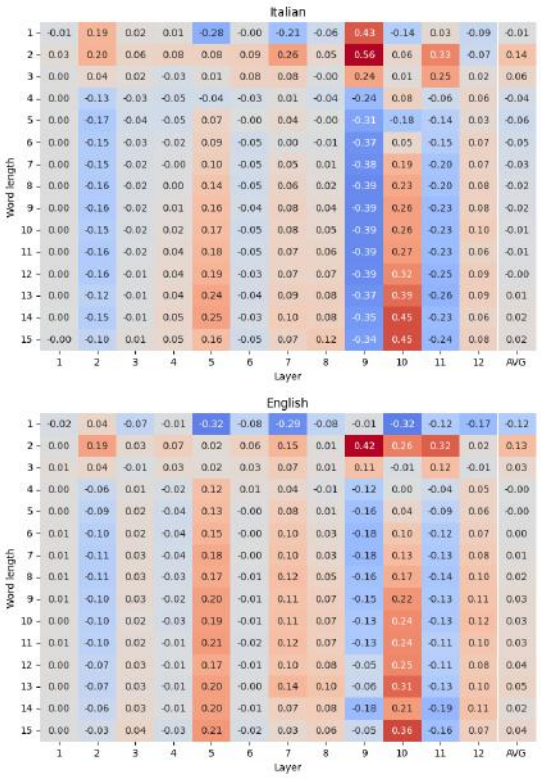


**Figure 2:** Attention shift for word length.

Figure 2 reports the results of the attention shift analysis for word length, showing three distinct patterns. First, **single-character words consistently lose attention after fine-tuning**, with particularly sharp drops observed in layers 5, 7, and 10. An exception appears in Italian, where these short words receive a notable increase in attention in layer 9. Second, **short words** (2-3 characters) **exhibit a general increase** in attention across most layers, especially pronounced in layer 9 and 11, suggesting that the fine-tuned model places greater importance on these short words. Finally, **longer words** (4+ characters) **show a more complex pattern, with attention picks and decreases alternating across layers**. Interestingly, layers 5 and 10 display a gradual increase in attention starting from 6-tokens long, suggesting that it may encode length-sensitive distinctions post-fine-tuning.

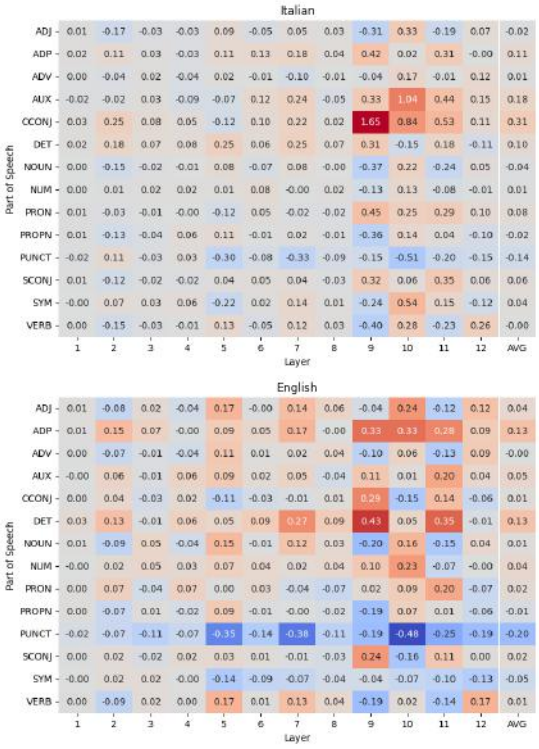


**Figure 3:** Attention shift for UD Parts of Speech.

Figure 3 shows the attention shift analysis across Parts of Speech. Overall, we observe a reduction in attention to punctuation marks (PUNCT) across layers, reinforcing the word length analysis and suggesting that the **model learns to down-weight non-lexical tokens after fine-tuning on eye-tracking data**. In contrast—and somewhat unexpectedly, given existing psycholinguistic evidence on human reading behavior—, **functional words** like adpositions (ADP), determiners (DET), and auxiliary verbs (AUX) **receive increased attention**, likely reflecting their importance in building the syntactic structure

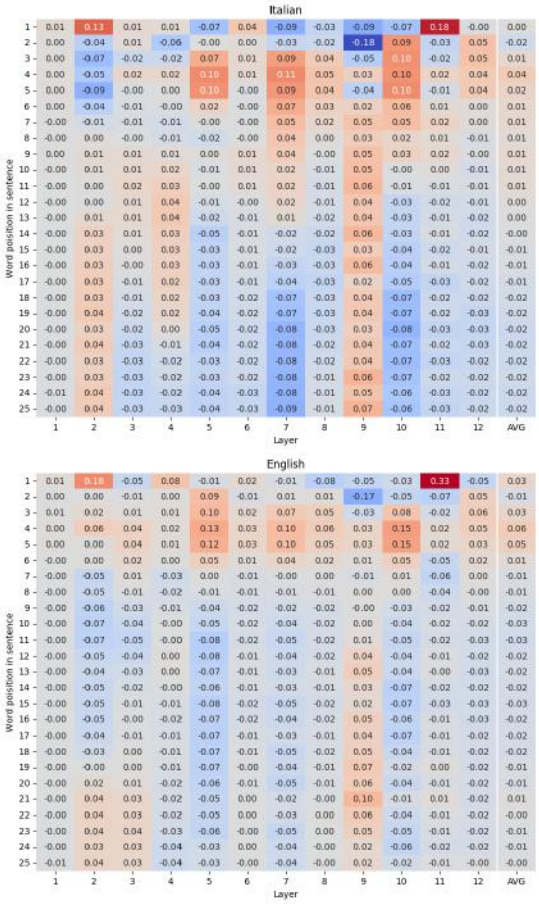


**Figure 4:** Attention shift for word position in sentence.

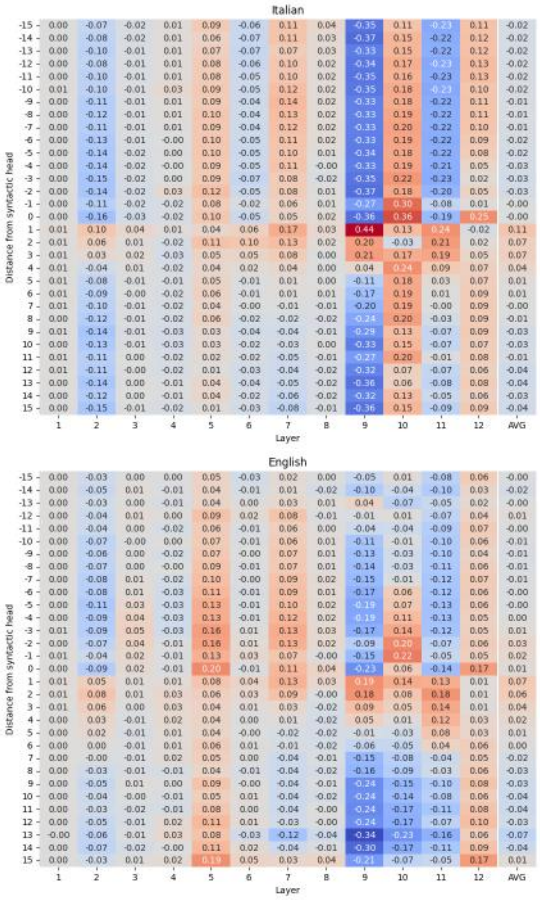


**Figure 5:** Attention shifts for distance from syntactic head.

and sentence interpretation. Additionally, a language-specific effect is visible in Italian, where coordinating conjunctions (CCONJ) gain notable attention across several layers. While similar shifts occur sporadically in the English model, they are less consistent and often offset by decreases in other layers.

As regards the attention shifts based on the word's position within the sentence (Figure 4), we noted that for both languages **tokens appearing earlier in the sentence generally receive slightly more attention after fine-tuning**, whereas those occurring later receive less. An exception is observed for the first two tokens, which deviate from this trend. Layer-specific behaviors also emerge: for instance, layers 2 and 9 tend to increase attention toward later tokens, while most other layers show the opposite effect, emphasizing earlier positions. Notably, layer 2 and layer 11 both show sharp increases in attention to the first token, suggesting a potential reweighting of sentence-initial information after exposure to human reading patterns. Interestingly, quantitative data from the used UD treebanks show that early sentence positions largely correspond to syntactically central elements—particularly the root, which anchors the clause and governs the structure of major complements. The observed shift in attention may therefore reflect the model's increased sensitivity to syntactic organization cues at sentence onset, especially in specific layers. This behavior is also well-documented in psycholinguistic studies and indicative of incremental parsing, where early elements guide syntactic and semantic expectations during sentence comprehension.

Figure 5 shows the attention shifts for the head-dependent distance parameter. A positive value indicates that the head follows the dependent, while a negative one that the head precedes it. The special value 0 is assigned to the root of the sentence. On average, it emerged that **tokens that are syntactically closer to their head tend to receive more attention after fine-tuning, particularly when the head follows the dependent**. This suggests that **fine-tuning on ET data encourages**

**the model to prioritize syntactic dependencies that align with typical reading dynamics**, where upcoming heads may draw anticipatory processing effort.

### 5.3. Shifts in the Embedding Space

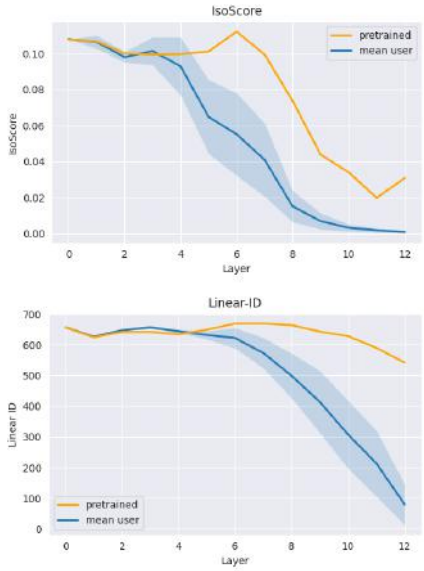


**Figure 6:** IsoScore (Top) and Linear Intrinsic Dimensionality (Bottom) of word embeddings from all model layers, before and after fine-tuning, averaged across users.

For space reasons, this section is limited to results for Italian; results for English show comparable trends and are provided in Appendix B. In the pre-trained model, IsoScore stays flat at $\approx 0.10$ through layer 6 and drops only in the final layers. After ET fine-tuning, the decline starts at layer 4, leaving layers 1–3 unchanged but rendering the upper layers markedly more anisotropic (Fig. 6, top). Linear-ID mirrors this pattern: the pre-trained model sustains $\approx$ 650 effective dimensions across all layers, whereas the fine-tuned one contracts from layer 4 onward and collapses to $< 100$ dimensions by layer 12 (Fig. 6, bottom). For these phenomena, as well as the ones to follow, the reduction of IsoScore and Linear-ID after fine-tuning is statistically significant ($p<0.05$ based on the Wilcoxon signed-rank test).

These results align with findings reported in [14] on how ET fine-tuning influences the embedding space shift. The linguistically informed analysis provides additional insights. Considering words grouped by *part of speech* and *head-dependent distance* (analyses for the remaining features are given in Appendix B), some main trends emerge. For POS (Figures 7 and 8), **the pre-trained model assigns content words (NOUN, VERB, PROPN) the highest-dimensional, most isotropic subspaces, with functional words and punctuation confined to lower dimensions**. Since content words exhibit high semantic diversity, the model tends to distribute their embeddings across many nearly orthogonal directions, resulting in broader and more isotropic subspaces. Function words, being few but very frequent and semantically uniform, collapse into a tight, anisotropic region, yielding lower IsoScore and Linear-ID. **Fine-tuning compresses all POS categories in the upper stack**, erasing the hierarchy above layer $\sim$ 6 while retaining it below; content words still display slightly greater variability. The observed contraction of the embedding space and loss of isotropy mirror the new optimization objective imposed during fine-tuning: to solve the ET task, the model no longer requires highly granular lexical representations, even for content words, so the latent geometry collapses accordingly. Turning to syntactic structure as captured by dependency distance (Figures 9 and 10), we observe **a notable asymmetry already in the pre-trained model based on the position of the dependent**: right dependents (and specifically within $d \in [-3, -1]$ of the head) display higher Linear-ID and isotropy, while left ones are confined to lower-dimensional and less uniform subspaces. This phenomenon appears highly interesting and, to the best of our knowledge, has not been previously reported, warranting a more in-depth investigation. **After fine-tuning, the model applies a uniform compression across all distance bins in the upper layers** (from layer 8 onward), while preserving the strong distinction between left and right dependents in earlier layers.

## 6. Conclusion

In this paper, we proposed a linguistically informed approach to study the impact of incorporating human reading behavior into a NLM. Our main findings reveal systematic and interpretable changes across both attention patterns and the representation space. Fine-tuning on eye-tracking shifts attention toward syntactic cues and sentence-initial elements, while reducing focus on non-informative tokens like punctuation, especially in middle and upper layers. Some of these trends partially mirror human reading dynamics and warrant further investigation. At the representational level, we observe substantial compression and increased anisotropy, especially for functional words and tokens close to their syntactic heads. We believe these preliminary findings confirm the value of analyzing attention and representation spaces through a linguistic lens, and open several avenues for future research, including how compression effects and cognitively grounded attention patterns may support the development of smaller, more efficient models through human-inspired inductive biases.

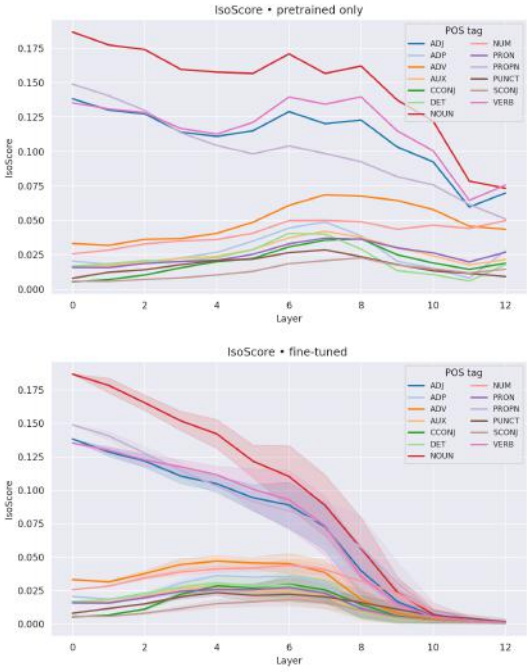


**Figure 7:** Isotropy before (top) and after (bottom) fine-tuning, shown for the 13 most frequent POSs

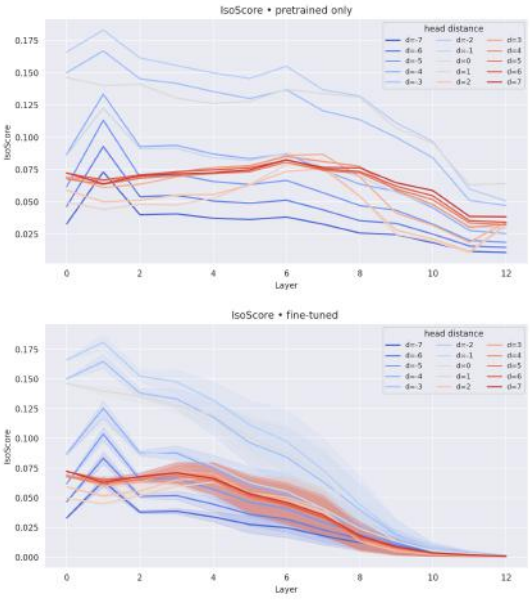


**Figure 9:** Isotropy before (top) and after (bottom) fine-tuning, shown for syntax head distance (up to 7 tokens distance).

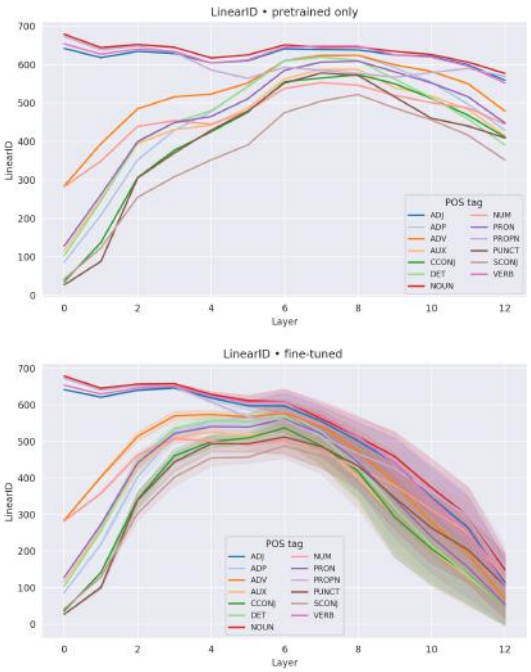


**Figure 8:** Linear-ID before (top) and after (bottom) fine-tuning, shown for the 13 most frequent POS classes.

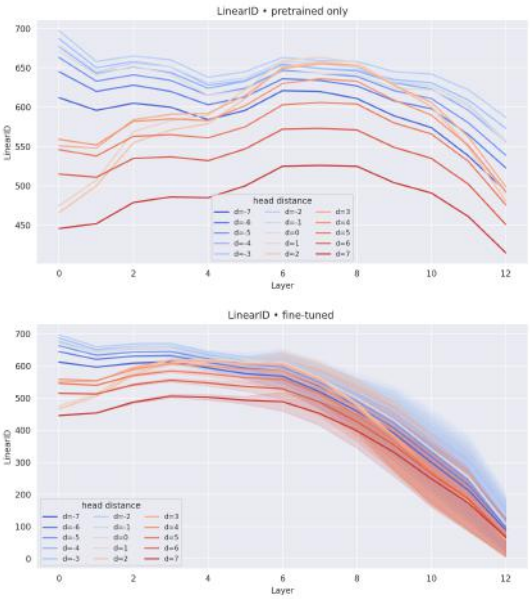


**Figure 10:** Linear-ID before (top) and after (bottom) fine-tuning, shown for syntax head distance (up to 7 tokens distance).

## Acknowledgments

This work has been supported by:

- FAIR - Future AI Research (PE00000013) projects under the NRRP MUR program funded by the NextGenerationEU.;
- The project "XAI-CARE" funded by the European Union - Next Generation EU - NRRP M6C2 "Investment 2.1 Enhancement and strengthening of biomedical research in the NHS" (PNRR-MAD-2022-12376692_VADALA' – CUP F83C22002470001)
- The project "Human in Neural Language Models" (IsC93_HiNLM), funded by CINECA3 under the ISCRA initiative;
- Language Of Dreams: the relationship between sleep mentation, neurophysiology, and neurological disorders - PRIN 2022 2022BNE97C_SH4_PRIN2022.

## References

[1] N. Hollenstein, M. Barrett, M. Troendle, F. Bigiolli, N. Langer, C. Zhang, Advancing NLP with cognitive language processing signals, CoRR abs/1904.02682 (2019).

[2] L. Evanson, Y. Lakretz, J.-R. King, Language acquisition: do children and language models follow similar learning stages?, in: Annual Meeting of the Association for Computational Linguistics, 2023. URL: https://api.semanticscholar.org/CorpusID:259089351.

[3] A. Yedetore, T. Linzen, R. Frank, R. T. McCoy, How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9370–9393. URL: https://aclanthology.org/2023.acl-long.521/. doi:10.18653/v1/2023.acl-long.521.

[4] M. A. Just, P. A. Carpenter, A theory of reading: from eye fixations to comprehension., Psychological review 87 (1980) 329.

[5] K. Rayner, Eye movements in reading and information processing: 20 years of research., Psychological bulletin 124 (1998) 372.

[6] M. Barrett, J. Bingel, F. Keller, A. Søgaard, Weakly supervised part-of-speech tagging using eye-tracking data, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016, pp. 579–584.

[7] A. Mishra, D. Kanojia, S. Nagar, K. Dey, P. Bhattacharyya, Leveraging cognitive features for sentiment analysis, in: S. Riezler, Y. Goldberg (Eds.), Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 156–166. URL: https://aclanthology.org/K16-1016/. doi:10.18653/v1/K16-1016.

[8] Y. Berzak, B. Katz, R. Levy, Assessing language proficiency from eye movements in reading, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1986–1996. URL: https://aclanthology.org/N18-1180/. doi:10.18653/v1/N18-1180.

[9] J. Malmaud, R. Levy, Y. Berzak, Bridging information-seeking human gaze and machine reading comprehension, in: R. Fernández, T. Linzen (Eds.), Proceedings of the 24th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2020, pp. 142–152. URL: https://aclanthology.org/2020.conll-1.11/. doi:10.18653/v1/2020.conll-1.11.

[10] E. Sood, S. Tannert, P. Mueller, A. Bulling, Improving natural language processing tasks with human gaze-guided neural attention, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 6327–6341. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/460191c72f67e90150a093b4585e7eb4-Paper.pdf.

[11] O. Eberle, S. Brandl, J. Pilot, A. Søgaard, Do transformer models show similar attention patterns to task-specific human gaze?, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4295–4309. URL: https://aclanthology.org/2022.acl-long.296. doi:10.18653/v1/2022.acl-long.296.

[12] J. Bensemann, A. Y. Peng, D. B. Prado, Y. Chen, N. Ö. Tan, P. M. Corballis, P. Riddle, M. Witbrock, Eye gaze and self-attention: How humans and transformers attend words in sentences, Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (2022). URL: https://api.semanticscholar.org/CorpusID:248780077.

[13] B. Wang, B. Liang, L. Zhou, R. Xu, Gaze-infused bert: Do human gaze signals help pre-trained language models?, Neural Comput. Appl. 36 (2024) 12461–12482. URL: https://doi.org/10.1007/s00521-024-09725-8. doi:10.1007/s00521-024-09725-8.

[14] L. Dini, L. Domenichelli, D. Brunato, F. Dell'Orletta, From human reading to NLM understanding: Evaluating the role of eye-tracking data in encoder-based models, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 17796–17813. URL: https://aclanthology.org/2025.acl-long.870/. doi:10.18653/v1/2025.acl-long.870.

[15] U. Cop, N. Dirix, D. Drieghe, W. Duyck, Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading, Behavior Research Methods 49 (2017) 602–615. URL: https://api.semanticscholar.org/CorpusID:11567309.

[16] N. Siegelman, S. Schroeder, C. Acartürk, H.-D. Ahn, S. Alexeeva, S. Amenta, R. Bertram, R. Bonandrini, M. Brysbaert, D. Chernova, et al., Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco), Behavior research methods (2022) 1–21.

[17] O. Raymond, Y. Moldagali, N. Al Madi, A dataset of underrepresented languages in eye tracking research, in: Proceedings of the 2023 Symposium on Eye Tracking Research and Applications, ETRA '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3588015.3590128. doi:10.1145/3588015.3590128.

[18] E. Sood, S. Tannert, D. Frassinelli, A. Bulling, N. T. Vu, Interpreting attention models with human visual attention in machine reading comprehension, in: R. Fernández, T. Linzen (Eds.), Proceedings of the 24th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online, 2020, pp. 12–25. URL: https://aclanthology.org/2020.conll-1.2/. doi:10.18653/v1/2020.conll-1.2.

[19] F. Morger, S. Brandl, L. Beinborn, N. Hollenstein, A cross-lingual comparison of human and model relative word importance, in: S. Dobnik, J. Grove, A. Sayeed (Eds.), Proceedings of the 2022 CLASP Conference on (Dis)embodiment, Association for Computational Linguistics, Gothenburg, Sweden, 2022, pp. 11–23. URL: https://aclanthology.org/2022.clasp-1.2.

[20] X. Wang, X. Li, X. Li, C. Biemann, Probing large language models from a human behavioral perspective, in: T. Dong, E. Hinrichs, Z. Han, K. Liu, Y. Song, Y. Cao, C. F. Hempelmann, R. Sifa (Eds.), Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024, ELRA and ICCL, Torino, Italia, 2024, pp. 1–7. URL: https://aclanthology.org/2024.neusymbridge-1.1/.

[21] S. Jain, B. C. Wallace, Attention is not Explanation, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556. URL: https://aclanthology.org/N19-1357/. doi:10.18653/v1/N19-1357.

[22] S. Serrano, N. A. Smith, Is attention interpretable?, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2931–2951. URL: https://aclanthology.org/P19-1282/. doi:10.18653/v1/P19-1282.

[23] S. Abnar, W. Zuidema, Quantifying attention flow in transformers, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4190–4197. URL: https://aclanthology.org/2020.acl-main.385/. doi:10.18653/v1/2020.acl-main.385.

[24] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 782–791.

[25] K. Clark, U. Khandelwal, O. Levy, C. D. Manning, What does BERT look at? an analysis of BERT's attention, in: T. Linzen, G. Chrupała, Y. Belinkov, D. Hupkes (Eds.), Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Florence, Italy, 2019, pp. 276–286. URL: https://aclanthology.org/W19-4828/. doi:10.18653/v1/W19-4828.

[26] J. Vig, Y. Belinkov, Analyzing the structure of attention in a transformer language model, in: BlackboxNLP@ACL, 2019. URL: https://api.semanticscholar.org/CorpusID:184486755.

[27] P. M. Htut, J. Phang, S. Bordia, S. R. Bowman, Do attention heads in bert track syntactic dependencies? (2019). URL: https://arxiv.org/abs/1911.12246. arXiv:1911.12246.

[28] K. Ethayarajh, How contextual are contextual-

ized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 55–65. URL: https://aclanthology.org/D19-1006/. doi:10.18653/v1/D19-1006.

[29] N. Godey, É. Clergerie, B. Sagot, Anisotropy is inherent to self-attention in transformers, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 35–48. URL: https://aclanthology.org/2024.eacl-long.3/.

[30] J. Gao, D. He, X. Tan, T. Qin, L. Wang, T. Liu, Representation degeneration problem in training natural language generation models, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=SkEYojRqtm.

[31] X. Cai, J. Huang, Y. Bian, K. Church, Isotropy in the contextual embedding space: Clusters and manifolds, in: International conference on learning representations, 2021.

[32] Z. Zhang, C. Gao, C. Xu, R. Miao, Q. Yang, J. Shao, Revisiting representation degeneration problem in language modeling, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 518–527.

[33] T. Mickus, D. Paperno, M. Constant, K. van Deemter, What do you mean, BERT? assessing bert as a distributional semantics model, in: A. Ettinger, G. Jarosz, J. Pater (Eds.), Proceedings of the Society for Computation in Linguistics 2020, Association for Computational Linguistics, New York, New York, 2020, pp. 279–290. URL: https://aclanthology.org/2020.scil-1.35/.

[34] R. Diehl Martinez, Z. Goriely, A. Caines, P. Buttery, L. Beinborn, Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 5999–6011. URL: https://aclanthology.org/2024.emnlp-main.344/. doi:10.18653/v1/2024.emnlp-main.344.

[35] W. Rudman, C. Eickhoff, Stable anisotropic regularization, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=dbQH9AOVd5.

[36] A. Machina, R. Mercer, Anisotropy is not inherent to transformers, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 4892–4907. URL: https://aclanthology.org/2024.naacl-long.274/. doi:10.18653/v1/2024.naacl-long.274.

[37] A. Ansuini, A. Laio, J. H. Macke, D. Zoccolan, Intrinsic dimension of data representations in deep neural networks, Advances in Neural Information Processing Systems 32 (2019).

[38] E. Hernandez, J. Andreas, The low-dimensional linear geometry of contextualized word representations, in: Conference on Computational Natural Language Learning, 2021. URL: https://api.semanticscholar.org/CorpusID:234742544.

[39] E. Cheng, C. Kervadec, M. Baroni, Bridging information-theoretic and geometric compression in language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2023, p. 12397–12420.

[40] M.-C. de Marneffe, C. D. Manning, J. Nivre, D. Zeman, Universal dependencies, Computational Linguistics 47 (2021) 255–308. URL: https://doi.org/10.1162/coli_a_00402. doi:10.1162/coli_a_00402.

[41] N. Silveira, T. Dozat, M.-C. de Marneffe, S. Bowman, M. Connor, J. Bauer, C. D. Manning, A gold standard dependency corpus for English, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), 2014.

[42] S. Brandl, N. Hollenstein, Every word counts: A multilingual analysis of individual human alignment with model attention, in: Y. He, H. Ji, S. Li, Y. Liu, C.-H. Chang (Eds.), Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online only, 2022, pp. 72–77. URL: https://aclanthology.org/2022.aacl-short.10/. doi:10.18653/v1/2022.aacl-short.10.

[43] A. J. Parker, T. J. Slattery, Spelling ability influences early letter encoding during reading: Evidence from return-sweep eye movements, Quarterly Journal of Experimental Psychology 74 (2021) 135–149. URL: https://doi.org/10.1177/1747021820949150. doi:10.1177/1747021820949150, pMID: 32705948.

[44] J. Ashby, K. Rayner, C. Clifton, Eye movements of highly skilled and average readers: Differential effects of frequency and predictability,

The Quarterly Journal of Experimental Psychology Section A 58 (2005) 1065–1086. doi:10.1080/02724980443000476.

[45] T. J. Slattery, M. Yates, Word skipping: Effects of word length, predictability, spelling and reading skill, Quarterly Journal of Experimental Psychology 71 (2018) 250–259. doi:10.1080/17470218.2017.1310264.

[46] N. Hollenstein, F. Pirovano, C. Zhang, L. Jäger, L. Beinborn, Multilingual language models predict human reading behavior, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 106–123. URL: https://aclanthology.org/2021.naacl-main.10/. doi:10.18653/v1/2021.naacl-main.10.

[47] W. Rudman, N. Gillman, T. Rayne, C. Eickhoff, IsoScore: Measuring the uniformity of embedding space utilization, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3325–3339. URL: https://aclanthology.org/2022.findings-acl.262/. doi:10.18653/v1/2022.findings-acl.262.

[48] J. H. Lee, T. Jiralerspong, L. Yu, Y. Bengio, E. Cheng, Geometric signatures of compositionality across a language model's lifetime (2025). URL: https://arxiv.org/abs/2410.01444. arXiv:2410.01444.

[49] L. Dini, L. Moroni, D. Brunato, F. Dell'Orletta, In the eyes of a language model: A comprehensive examination through eye-tracking data, Neurocomputing (2025). In press.

# A. Shift in the embeddings space - Extra features

This Appendix section contains the analysis of Section 5.3 conducted on the remaining linguistic features: word length, Figures A.1 and A.2, and word index in sentence, Features A.3 and A.4. As in Section 5.3, a clear hierarchy emerges among the new feature classes. For *word length*, tokens 6–10 characters long retain the highest IsoScore and Linear-ID before collapsing, like all other bins, under fine-tuning.

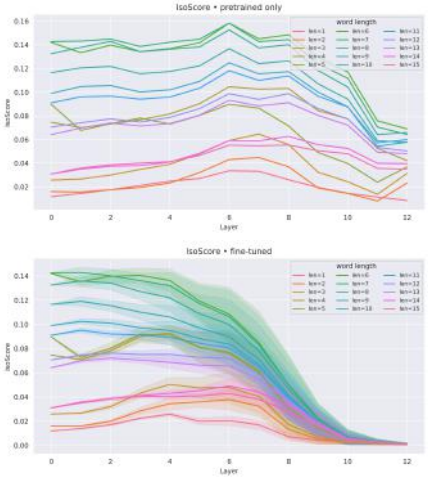


**Figure A.1:** Isotropy before (left) and after (right) fine-tuning, shown for word length (up to 15 tokens).

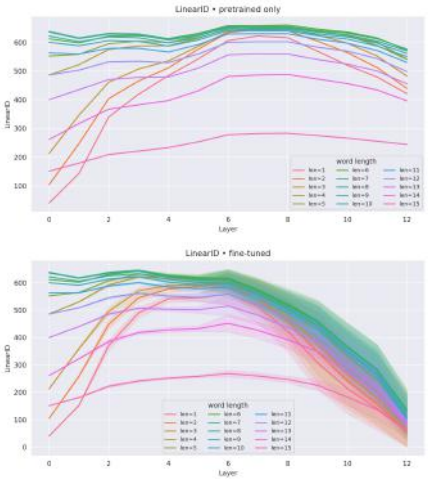


**Figure A.2:** Linear-ID before (left) and after (right) fine-tuning, shown for word length (up to 15 tokens).

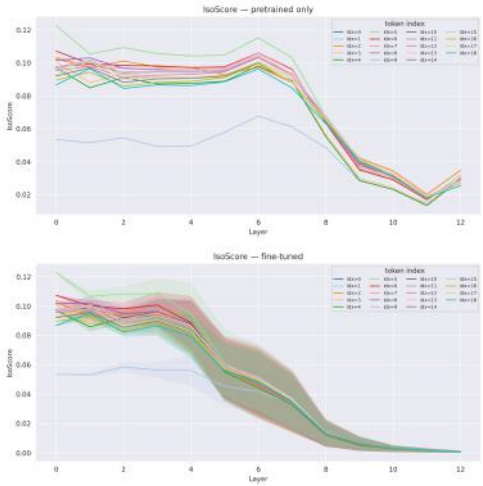


**Figure A.3:** Isotropy before (left) and after (right) fine-tuning, shown for word index (up to index 18).

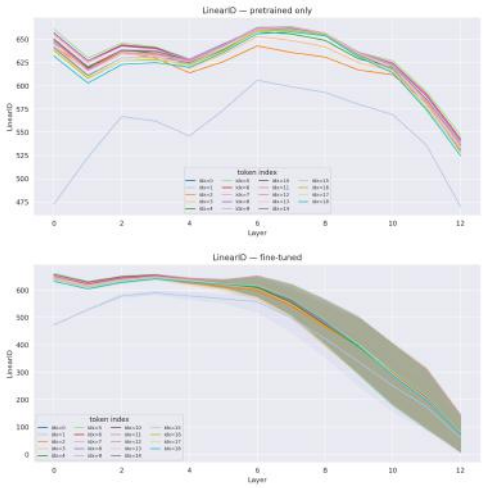


**Figure A.4:** Linear-ID before (left) and after (right) fine-tuning, shown for word index (up to index 18).

For the *word-index* feature, position 1 is the most distinctive. Lexical composition of these classes will be addressed in future work.

# B. Shift in the embedding space - English dataset

We report the scores on the English word embeddings. The results are comparable to those on the italian dataset. Further exploration of parallels and differences will be the focus of future work.

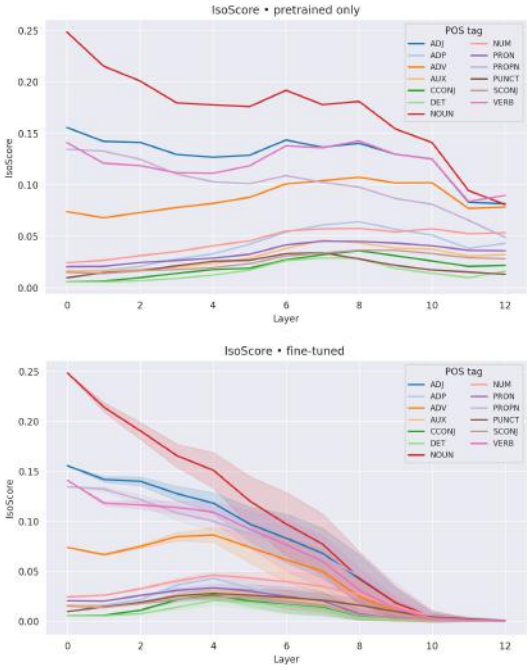


**Figure B.1:** Isotropy before (top) and after (bottom) fine-tuning, shown for the 13 most frequent POS classes.

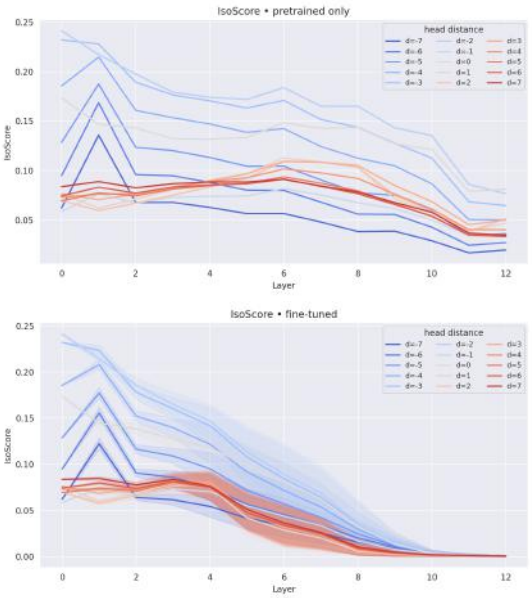


**Figure B.3:** Isotropy before (top) and after (bottom) fine-tuning, grouped by syntactic head distance (up to 7 words of distance).

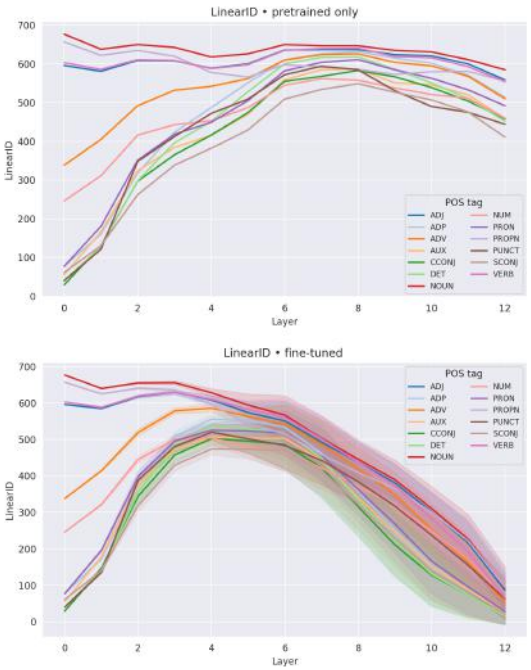


**Figure B.2:** Linear-ID before (top) and after (bottom) fine-tuning, shown for the 13 most frequent POS classes.

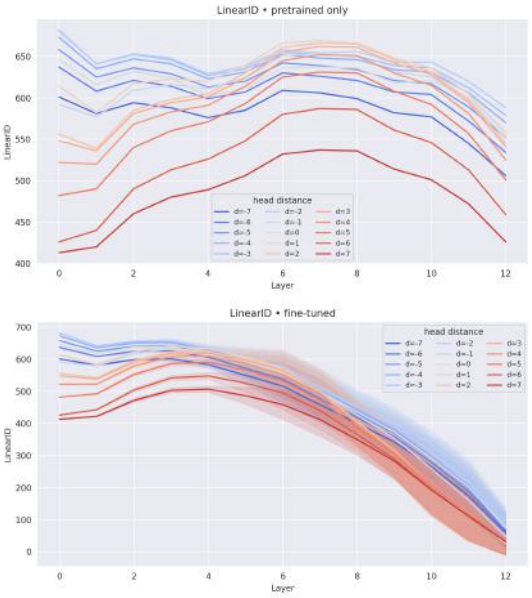


**Figure B.4:** Linear-ID before (top) and after (bottom) fine-tuning, grouped by syntactic head distance (up to 7 words of distance).

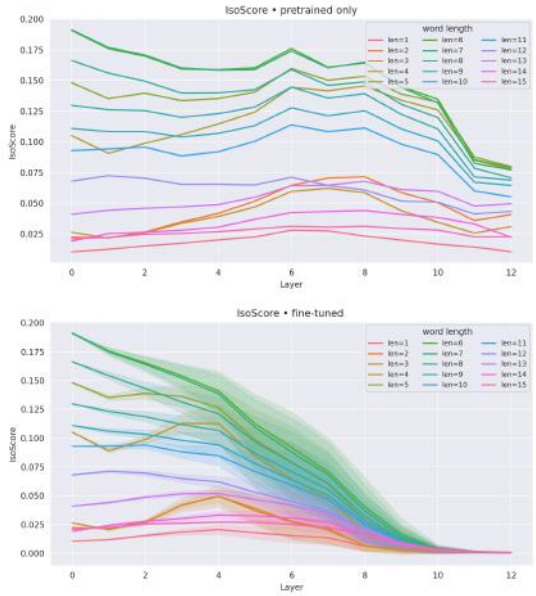


**Figure B.5:** Isotropy before (top) and after (bottom) fine-tuning, shown for word length (up to 15 tokens).

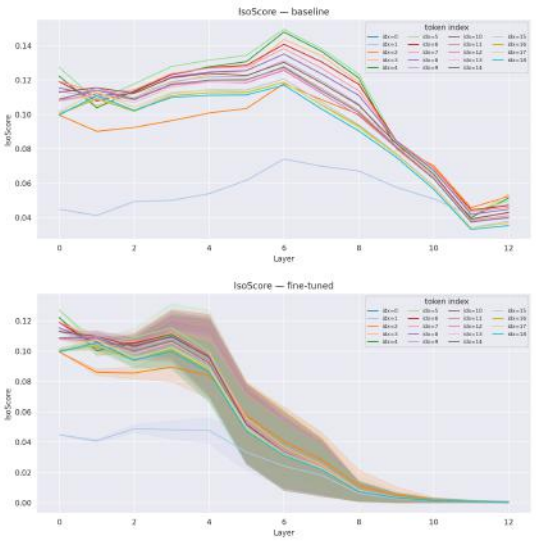


**Figure B.7:** Isotropy before (top) and after (bottom) fine-tuning, shown for word index (up to index 18).

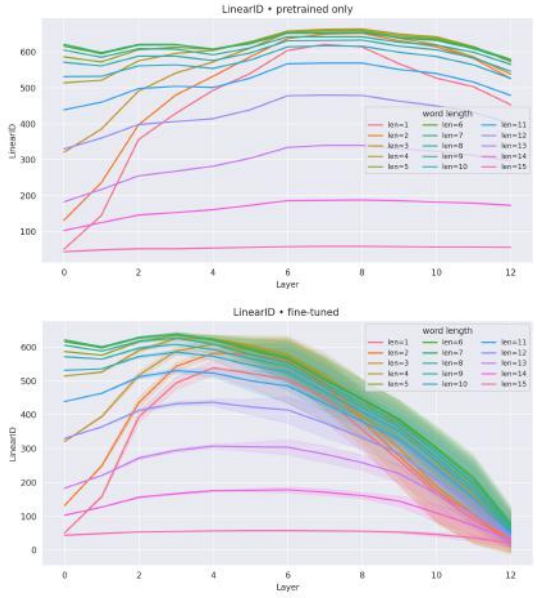


**Figure B.6:** Linear-ID before (top) and after (bottom) fine-tuning, shown for word length (up to 15 tokens).

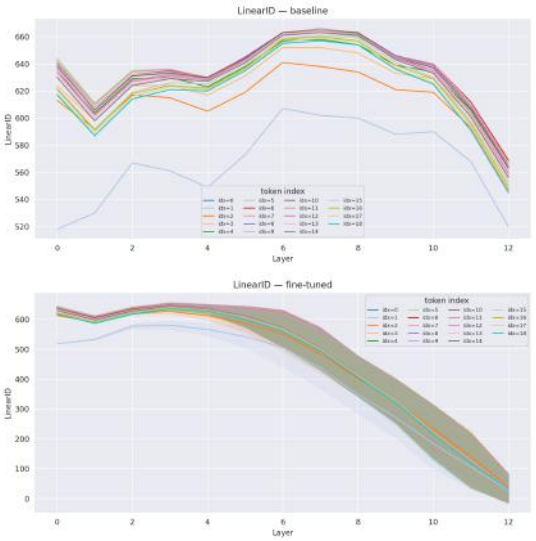


**Figure B.8:** Linear-ID before (top) and after (bottom) fine-tuning, shown for word index (up to index 18).