Metric-Learning Encoding Models Identify Processing Profiles of Linguistic Features in BERT's Representations

Anonymous ACL submission

Abstract

We introduce Metric-Learning Encoding Models (MLEMs) as a new approach to understand how neural systems represent the theoretical features of the objects they process. As a proof-of-concept, we apply MLEMs to neural representations extracted from BERT, and track a wide variety of linguistic features (e.g., tense, subject person, clause type, clause embedding). We find that: (1) linguistic features are *ordered*: they separate representations of sentences to different degrees in different layers; (2) neural representations are organized *hierarchically*: in some layers, we find clusters of representations nested within larger clusters, following successively important linguistic features; (3) linguistic features are *disentangled* in middle layers: distinct, selective units are activated by distinct linguistic features. Methodologically, MLEMs are superior (4) to multivariate decoding methods, being more robust to type-I errors, and (5) to univariate encoding methods, in being able to predict both local and distributed representations. Together, this demonstrates the utility of Metric-Learning Encoding Methods for studying how linguistic features are neurally encoded in language models and the advantage of MLEMs over traditional methods. MLEMs can be extended to other domains (e.g. vision) and to other neural systems, such as the human brain.

1 Introduction

An open question in neuroscience and Artificial Intelligence is how neural networks, biological or artificial, encode and process natural language. Modern neural language models provide new means to study this question: Unlike in experiments involving humans, language models can be exposed to an extensive range of stimuli while their neural activity is entirely recorded. However, despite this new opportunity, how natural language is neurally encoded, either in models or the human brain, remains largely unknown.

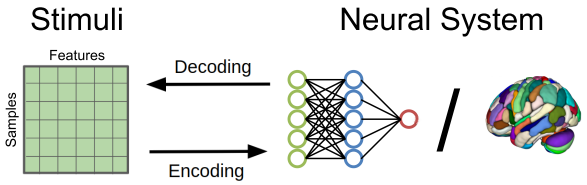

Figure 1: **Encoding vs. Decoding models**: Encoding models predict unit activation from the stimuli, Decoding models predict stimulus features from activations.

An important step towards understanding the neural mechanisms underlying language processing is to understand where and how *linguistic features* are neurally encoded. Linguistic features, such as grammatical number (singular vs. plural), tense (e.g. past, present and future) or verb type (intransitive vs. transitive), are theoretical constructs from linguistics, which aid in the analysis and understanding of natural language, and to discern and categorize various aspects of language, such as its structure and usage. Linguistic features are the building blocks to study more complex linguistic phenomena.

Two general approaches to study the neural encoding of linguistic features can be discerned in the literature: *decoding* and *encoding* methods (Fig. 1; e.g. King et al., 2020b). In the decoding setup, the goal is to predict features of the stimulus from neural activations, typically using standard machine-learning classifiers (aka, 'diagnostic probes') (Alain and Bengio, 2016; Adi et al., 2016; Hupkes and Zuidema, 2017; Belinkov and Glass, 2018; Conneau et al., 2018; Tenney et al., 2019; Arps et al., 2022). In the encoding setup, the arrow is reversed, and the goal is to predict neural activity from a set of features, typically, using regularized regression methods. The encoding approach is most commonly used in neuroscience (e.g. Wehbe et al., 2014; Caucheteux et al., 2021; Caucheteux and King, 2022; Pasquiou et al., 2022; Oota et al., 2022; Pasquiou et al., 2023).

Decoding and encoding methods have different merits and limitations. One main limitation of decoding methods is that the decodability of a given feature does not guarantee its causal role. For instance, a certain feature can be decodable from neural activations not because it has a mechanistic role but only because it correlates with another feature that has such a role. In encoding models, this limitation can be addressed to some extent by introducing the feature of interest and confound features, testing their relative importance in predicting neural activations. However, a common limitation of encoding models is that they are uni- rather than multi-variate, where the goal is typically to predict the neural activity of one *single* unit of the model at a time, from a single electrode in the brain or from a single fMRI voxel. Encoding approaches are thus limited in their ability to study *distributed* representations across many units.

Here, we introduce a simple approach, which preserves the good from both worlds, by extending encoding methods to the multivariate case within a metric-learning framework (Kulis, 2013). We call it Metric-Learning Encoding Models, or MLEMs for short. To study the neural encoding of linguistic features, we created four new datasets, whose stimuli contrast various linguistic features. We then presented stimuli from these datasets to BERT (Devlin et al., 2019), extracted its neural activations and studied them using MLEMs.

The main contributions of our study are: (1) A new framework to study neural encoding in large language models; (2) A new set of probing datasets with their corresponding generating codes; (3) Identification of orders among linguistic features, for all model layers, with respect to their dominance in the neural representations; (4) Identification of hierarchical patterns in the neural representations of linguistic information; (5) Identification of a strong *disentanglement* of linguistic features in layers of BERT, discovered by *contrasting uni- and multivariate* encoding models; (6) Demonstration of the limitation of multivariate decoding methods compared to encoding approaches.

2 Related Literature

Metric-Learning Encoding Models (MLEMs) start from the assumption that to effectively capture multivariate, distributed neural encoding of linguistic information, one should model *distances* among neural representations rather than individual activations (e.g. units or electrodes). Given a set of inputs (e.g. sentences), where each is represented along a set of features (e.g. linguistic features), the goal is to learn a *metric function* (aka, a distance function), which is defined over pairs of inputs and computed based on their features. The optimal metric function is the one that minimizes the differences between the modelled distances among the inputs and the empirical (neural) ones. This optimal metric can be derived using standard metric-learning methods (Kulis, 2013).

MLEMs are therefore closely related to classic work on *second-order isomorphism* between representations of a given system and the represented entities (Shepard and Chipman, 1970). Second-order isomorphism suggests that while the representations and the represented entities belong to different spaces, the similarity between the entities and their representations can be quantified by comparing the pairwise distances within each domain, thus 'second-order' similarity. Second-order isomorphism has also been used to compare representations of two different systems, such as two artificial neural networks (Laakso and Cottrell, 2000) or brains, which is known as Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008; Abnar et al., 2019).

However, in its simple form, second-order isomorphism does not include any *learning*, which is a limitation when operations performed by the representing system strongly deform the input structure. Feature-reweighted representational similarity analysis (FR-RSA; Jozwik et al., 2017; Storrs et al., 2021; Kaniuth and Hebart, 2022), a variant of RSA, addresses this limitation by learning a set of weights over the set of features. However, feature weighing is not enough to ensure that the modelled neural distances have metric properties (e.g. non-negativity and triangle inequality).

MLEMs extend second-order isomorphism methods by: (1) explicitly learning a metric function, which ensures that the learned weights preserve distance properties; (2) allowing the introduction of *interactions* among features into the model, while preserving metric properties (unlike FR-RSA); (3) incorporating conditional permutation importance testing (Chamma et al., 2023), which is the current state-of-the-art in handling possible strong correlations among features.

The closest variant of RSA, which is similar to our approach is Representational Similarity *Learning* (RSL; Oswal et al., 2016). However, our work

differs from this work by learning an encoding rather than a decoding model. Furthermore, our focus here is on neural encoding of linguistic feature in large language models. Finally, we highlight other multivariate encoding approaches, which follow different approaches, such as Reduced-Rank Ridge Regression (Mukherjee and Zhu, 2011), and the Back-to-back regression (King et al., 2020a).

3 Experimental Setup

3.1 Datasets

To study the neural encoding of linguistic features, we created four original datasets. Each dataset contains a list of sentences and their corresponding list of linguistic features. They were generated using grammars manually-crafted to cover central phenomena from linguistics. The first three datasets target specific linguistic phenomena by contrasting a small set of linguistic features: *(1) The Short-Sentence Dataset* mostly targets *word-level* features, such as grammatical number, subject type or tense; *(2) The Relative-Clause Dataset* targets more *structural* phenomena, and in particular, center-embedding vs. right-branching embeddings; *(3) The Long-Range-Agreement Dataset* targets sentence-level features, and in particular, long-range subject-verb dependencies. Finally, the fourth dataset, *(4) The Large Dataset* lumps together all the linguistic phenomena from the first three datasets, and adds others such as quantified sentences, with and without bound pronouns, and full clause embeddings. The datasets are described in Table A.1, with templates and examples, and they are fully available online with their generative code. To secure a clean interpretation of the relative contributions of the different features, we checked for correlations between linguistic features in all four datasets. Fig. A.1 shows that strong correlations only remain in the Large Dataset.

3.2 Language Model and Neural Representations

We studied the neural encoding of linguistic features in BERT (Devlin et al., 2019), a highly studied model, which allows us to compare our results to previous findings about layer-wise neural encoding of language in this model. Furthermore, the original BERT was trained not only on masked-language modeling but also on next-sentence prediction using a special <CLS> token. Here, we consider the embedding on <CLS> as an aggregated sentence-level representation (Jawahar et al., 2019; Rogers et al., 2021). For each sentence, we thus obtain one representation vector in $\mathbb{R}^{768}$ per layer.

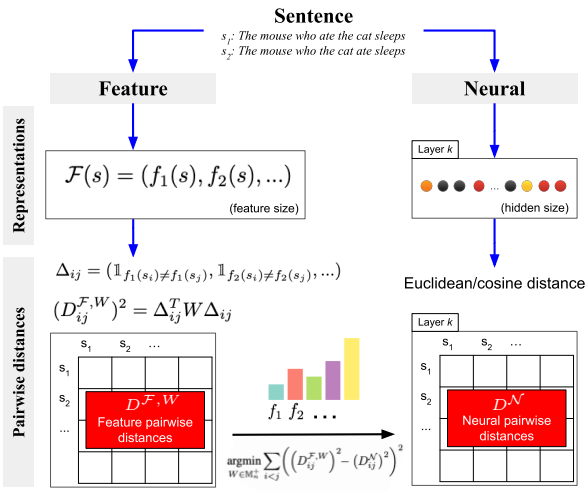

Figure 2: **A Metric-Learning Encoding Model**: MLEMs infer the relative importance of features by finding the best alignment between distances in feature space and in neural space.

3.3 Metric-Learning Encoding Models

We consider a set of N stimuli (sentences), each characterized by a set of (linguistic) features $\mathcal{F}$. MLEMs compute two types of pairwise distances. First, we compute *pairwise neural distances* $D^{\mathcal{N}}$ (right branch in Fig. 2) as the standard distance (e.g. Euclidean or cosine distance) between the neural responses of a set of units (e.g. a layer) for any two sentences. Second, we compute *pairwise feature distances* $D^{\mathcal{F},W}$ (left branch of Fig. 2) as follows. We start from *feature difference vectors*, which indicate on which features two sentences differ: $\Delta(s_i, s_j) = (\mathbb{1}_{f(s_i)\neq f(s_j)})_{f\in\mathcal{F}}$. Then, feature distances are computed using a standard bi-linear form parameterized by a PSD matrix $W \in \mathbb{M}_n^+$:

$$\left(D_{ij}^{\mathcal{F},W}\right)^2 = \Delta(s_i, s_j)^T W \Delta(s_i, s_j)$$

MLEMs, as metric-learning methods, optimize W to bring the pairwise feature distances as close as possible to the neural ones, across all (i, j) pairs of stimuli:

$$W^* = \underset{W\in\mathbb{M}_n^+}{\operatorname{argmin}} \sum_{i<j} \left(\left(D_{ij}^{\mathcal{F},W}\right)^2 - \left(D_{ij}^{\mathcal{N}}\right)^2 \right)^2 + \lambda ||W||_2^2$$

Algorithms such as OASIS (Chechik et al., 2010) solve this optimization problem under the PSD constraint, by projecting the weight matrix on the PSD

manifold every several optimization steps. When W is assumed to be diagonal (no interaction terms), the optimization problem can be reduced to a least-squares problem, and the PSD constraint becomes a non-negativity constraint on the diagonal terms.

Feature Importance To properly derive the contribution of each feature to the neural distances, we estimated *feature importance* by conducting conditional-permutation tests (Chamma et al., 2023), which is the state-of-the-art for feature importance estimation in the case of mild correlations among features.

Model Training and Evaluation For simplicity, we focused here on the diagonal case of W and trained a standard Ridge model with a non-negativity constraint on the parameters. We optimized for λ using nested cross-validation (CV; $\lambda \in 10^{[-4,4]}$; To facilitate λ optimization across all models, target values were min-max scaled into $[0, 1]$). We evaluated the model using the coefficient-of-determination score R^2 and report the average across CV splits.

4 Results

4.1 The Processing Profiles of Linguistic Features across BERT's Layers

The geometry of neural representations, and in particular, distances among representations, is informative about underlying computations in the model. Representations that are nearby in neural space are more prone to confusion, having similar effects on downstream computations. Conversely, representations that are relatively distant in neural space can have disparate effects and be thus important for downstream computations. Identifying which linguistic features cause large neural separations among sentence representations would thus suggest their computational role in each processing stage of the model. Which linguistic features create large neural distances among sentence embeddings in BERT? How do the effects of linguistic features on neural distances vary as a function of layer?

To study these questions, we presented each of the four datasets to BERT, extracted the corresponding embeddings, and trained an MLEM on the embeddings from each layer. We then inspected the resulting *Feature Importances* (FIs; Section 3.3), which quantify the contribution of each linguistic feature in predicting neural distances. Figure 3A shows Feature Importance (continuous lines) in log scale for the Short-Sentence Dataset, for the top four features achieving maximum FI across all layers ($p < 0.01$). See Figure B.2 for more features. This provides, for each feature, a processing profile across all layers of the model.

The Processing Profiles of Simple Features (clause type, subject properties, tense). Figure 3A shows that the clause type feature ("He smiles" vs. "Who smiles?") is the most dominant feature across all layers of the models, followed by subject type ("He smiles" vs. "The boy smiles"), tense ("He smiles" vs. "He smiled") and then subject person ("He smiles" vs. "I smile"), which achieves significant FIs only in deeper layers (statistically insignificant FIs were removed from the plot). Note that for Tense and Subject Person, the profile of FIs significantly changes from early to deeper layers of the model. In particular, the tense feature becomes more important in middle layers of the model, peaking at layer 5 of the model. Similarly, the feature importance of subject person becomes significant only in later layers of the models, from layer 7 on. A control analysis, training MLEMs on a non-trained BERT presented with the same datasets, showed that, without training, FI profiles remain largely constant across all layers (Figure B.4). This discrepancy between the processing profiles of trained and non-trained BERT, in particular for tense and subject person, suggests that neural computations associated with these features are triggered in middle layers of BERT - tense at layer 5 and subject person at layer 7.

The Processing Profile of Attachment site for Relative Clauses and Prepositional Phrases. For the Relative-Clause and Long-Range-Agreement Datasets, Figure 3C shows the processing profiles for the top 4 features with maximal FIs. These datasets contain structural features, such as whether the relative clause is attached to the subject (center embedded) or to the object (right branching) of the verb, and whether the embedded clause itself lacks its subject or its object. The dataset also traces word features, such as grammatical number (singular/plural), and low-level confound features, such as word frequency. The most striking change in FI profile, which is consistent for the two datasets (Figure B.1A), occurs for the attachment-site feature (red color). At layer 5, the processing profile of attachment site crosses three orders of magnitude of FI values. This suggests that neural computations associated

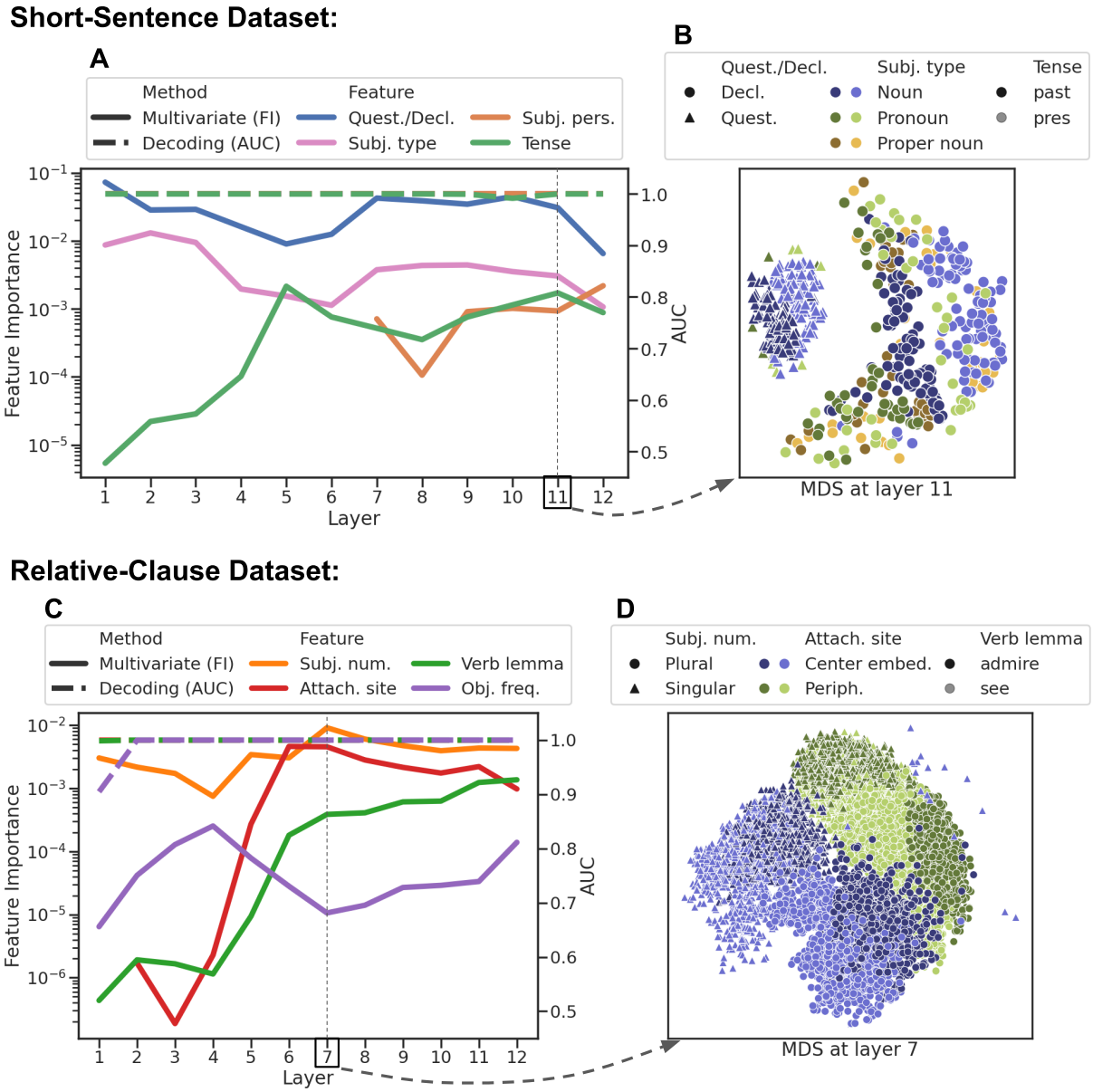

Figure 3: **Processing Profiles of Linguistic Features and Hierarchical Patterns. A&C:** Feature Importances for the Short-Sentence and Relative-Clause datasets for the top 4 features. The area under of curve (AUC) of a decoding baseline is shown on a second y-axis. **B&D:** Multidimensional (MDS) plots for example layers with high FIs (layer 11 and 7). Hierarchical patterns are observed according to the order of FIs: *Quest./Decl.* > *Subj. type* > *Tense* and *Subj. num.* > *Attach. site* > *Verb lemma*.

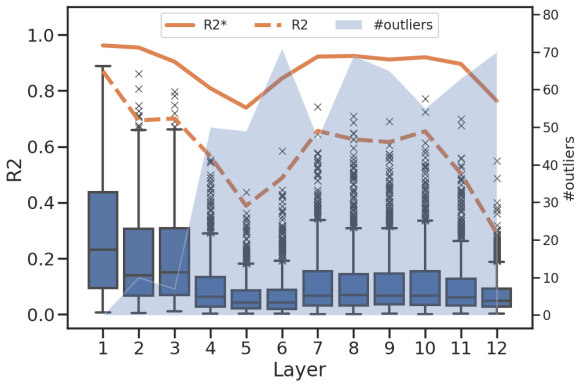

Figure 4: **Performance of Univariate Model (blue) vs. Multivariate MLEMs (orange).** Univariate: barplot of the unit R^2-scores per layer, outlier units (R^2 above 1.5 Interquantile Range above 75%) are shown with crosses, and counted (light-blue background with the second y-axis). Multivariate: The R^2 for the MLEMs trained on the entire layer (continuous line) and the optimal MLEM (dashed; section 4.3) are shown in orange.

with attachment site becomes dominant in these layers.

Processing Profiles are Robust across Model Initializations. Are the processing profiles robust across BERT initializations? To test this, we repeated the analyses for the Short-Sentence dataset with 25 more models of BERT, trained with different seeds. Results show that the processing profiles are consistent across all models (Figure B.3), and that, for all models, the resulting order among linguistic features is the same: clause type, subject type, tense and subject person.

Hierarchical Representations of Linguistic Features in BERT Embeddings. Is the order among linguistic features reflected in the geometry of neural representations? To visualize the results for the FIs, we projected the sentence representations from the original BERT onto the plane while preserving their pairwise neural distances in the original space as much as possible using Multi-Dimensional Scaling (MDS; Kruskal, 1964). Figure 3B illustrates this for one layer of BERT, where FI values are all high. Sentence embeddings (dots in the plot) are marked following the ordered linguistic features, showing that embeddings are first mostly separated by clause type (declarative and interrogative sentences; circles vs. triangles), then by subject type (different colors) and, finally, based on tense (shades). This shows that the order of FIs is reflected in the geometrical organization of sentence representations in neural space.

Figure 3D illustrates this for layer 7 of BERT, for the Relative-Clause Datasets. It shows that sentences are first clustered based on grammatical number (circles vs. triangles), then each cluster is divided into two identifiable clusters for sentence structure (different colors), and, finally, each of the sub-clusters is further divided into subsequent nested clusters for verb lemma (shades). Sentence representations at layer 7 thus form a hierarchical structure, with nested clusters, whose levels follow the order among the most dominant linguistic features. This hierarchical organization is not a necessity: it does not hold for earlier layers of the model and only emerges at layer 5, where the sudden increase of the sentence-structure feature occurs (Fig. 3C). A similar hierarchical organization is also identified for the Long-Range-Agreement Dataset with respect to the same three features (Fig. B.1). These hierarchical patterns become visible when sentences in the MDS are marked following the ordered features, and they would therefore have been hard to detect without the results from the MLEM.

4.2 Local and Distributed Neural Encoding of Linguistic Information

A central question about information encoding in neural networks is whether information is locally encoded - in a single unit, or a small set of units (Quiroga et al., 2005; Bowers, 2009) – or whether it is encoded distributedly, 'spread out' across several units in the network (Rumelhart et al., 1986; Smolensky, 1990; Gelder, 1992). Which linguistic features are locally encoded and which are distributedly encoded in BERT embeddings?

To study this question, we suggest a new approach, which contrasts univariate and multivariate encoding models. If a unit of the model locally encodes a specific feature, then a univariate encoding model containing this feature would perform well in predicting unseen neural activity (high R^2). However, if a feature is distributedly encoded, across many units, then univariate models would only capture part of the variance and would therefore have low performance compared to a multivariate encoding model. Comparing univariate and multivariate models could therefore reveal the type of encoding (local or distributed) in a network.

We thus trained univariate encoding models, which minimally differ from the multivariate MLEMs, by training an MLEM on the activations of each unit of each layer of the model. This en-

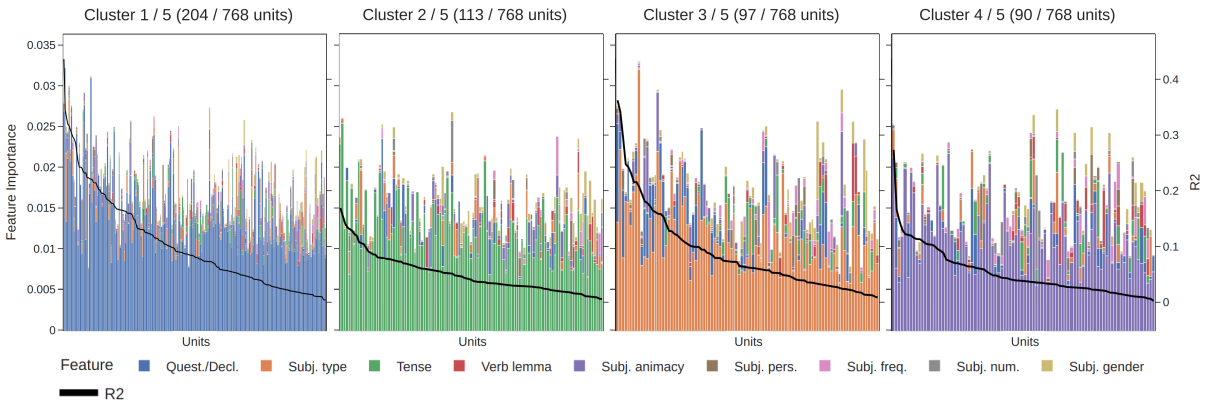

Figure 5: **Disentanglement of Linguistic Features into Separate Clusters of Units in Layer 5 of BERT.** The first four clusters from a K-means clustering of the univariate FI profiles of all units in layer 5. Each stacked bar shows the feature-importance profile (colors) of a single unit. Every cluster shows a specialization in one *single* feature, which is clearly observable from the dominant color (feature) in each cluster. Units in each cluster are sorted by decreasing R^2 (black line).

sures a close comparison between the univariate and multivariate methods. This resulted in a total of 768 units $\times$ 12 layers univariate models, each providing a profile of feature importance and a performance score (R^2).

We applied these models to the Short-Sentence Dataset. Figure 4 compares the performances of the univariate models and the MLEMs, showing that: (1) overall, MLEMs outperform the univariate models, across all layers. Thus, neural separation can be better predicted at the population compared to the unit level, which suggests that at least some linguistic information is distributedly encoded; (2) from layer 4 on, the median of the univariate distributions decreases (horizontal lines) while the distribution also becomes narrower (blue boxes). This tendency suggests an increase in distributed code in middle layers of the model; (3) However, from layer 4 on, the number of outlier units starts to increase (shaded background), and the performance of these units becomes comparable to that of the multivariate models. This increase in outlier units in middle layers of the model suggests that at least some linguistic information is locally encoded by such units. We further study this in what follows.

Strong Disentanglement in the Neural Encoding of Linguistic Features in Layer 5 of BERT. Comparing the performances of uni- and multivariate models suggests, so far, that some linguistic information is distributedly encoded in middle layers of the model, whereas other types of information are more locally encoded, by specific 'outlier' units. However, comparing model performances cannot tell which type of linguistic information is encoded in these units. We therefore next inspected the feature importances of the univariate models. For this, we clustered all units based on their FI profiles, using a standard K-means algorithm, and the silhouette method to determine the right number of clusters. Before clustering, to make FI profiles of different units comparable, we normalized the FI profile of each unit by its R^2. Figure 5 shows, for each cluster, the individual FI profile of each of its units (in a stacked bar), sorted by the performance of the univariate model.

Remarkably, every cluster (except for a fifth catch-all cluster; Figure E.1) has a single dominant linguistic feature whose feature importance is an order of magnitude above that of any other feature. Many of the units in each cluster, and in particular 'outlier' units, are selective to the dominant feature. These results show a strong specialization and disentanglement of linguistic information in layer 5 of BERT. BERT, unlike other models such as variational auto-encoders (Higgins et al., 2016), is not trained with an explicit disentanglement term. Disentanglement therefore spontaneously emerges in the model during training. Finally, we note that earlier layers of BERT also show disentanglement, but to a smaller number of features. Later layers show no disentanglement with respect to our features and all have a catch-all cluster (see Figure E.1).

4.3 Advantages of MLEMs over Traditional Decoding and Encoding Approaches

Our analyses show that MLEMs are better suited for the study of multivariate encoding of linguis-

tic features than Diagnostic Probes (e.g. Jawahar et al., 2019; Puccetti et al., 2021; Miaschi et al., 2020; Rogers et al., 2021). The limitation of diagnostic probes is visible in Figure 3A&C. Each of the four top features can be perfectly decoded from the sentence representations of any layer of BERT. The Area Under the Curve (AUC; 3-fold nested cross-validation) of a linear Ridge classifier is at ceiling, in all cases. In fact, it is possible that the entire stimulus can be recomposed, and hence the features. MLEMs, unlike diagnostic probes, jointly includes several features, as well as possible 'low-level' confound features (e.g. sentence length), and disentangle their respective contributions (FIs) in explaining neural activity. MLEMs are thus less prone to false-positive errors by including possible confound variables in the model, and specifically detect which features are encoded, and where.

MLEMs also have advantages over univariate encoding methods in that they better predict differences in neural activity. To show this, for each layer, we also trained an *optimal* MLEM in the following way. We first sorted the units of each layer based on the R^2 performance of corresponding univariate models. We then, incrementally, trained an MLEM on only the top-d units, for $d \in [1, 768]$. Finally, we define the optimal number of encoding units, d^*, as the number of units required to achieve maximal MLEM performance, which we denote by R^{2*} (Figure D.1). Figure 4 shows the resulting R^{2*} for all layers of the models, showing that the MLEMs outperform univariate models, across all layers, including all outliers in each layer. Taken together, this shows the superiority of MLEMs over traditional decoding and encoding approaches.

5 Discussion

We introduce Metric-Learning Encoding Models (MLEMs): a new framework for the study of neural encoding of linguistic features. Like decoding approaches, it is *unit*-multivariate, that is, all units enter the same model. As such, it can capture distributed encoding of features. Like encoding approaches, it is *feature*-multivariate, that is all features enter the same model. As such, it can properly disentangle the specific contributions of different features. MLEMs extend other multivariate approaches, such as Representational Similarity Analysis (Kriegeskorte et al., 2008), by precisely modelling a distance (i.e. imposing the matrix of parameters to be positive-semi definite).

We tested MLEMs on BERT, using 4 original probing datasets, which provided: quantitative measures for the dominance of various linguistic features across layers, the identification of previously undescribed hierarchical patterns of neural activity, encoding profiles of linguistic features and strong disentanglement of representations in several layers. This demonstrates the capability of MLEMs to study neural encoding in large language models.

MLEMs are applicable to the activations of any language processing system, be they other artificial models, or the human brain. As such, the current results provide a reference point for predictions of how other systems process language.

MLEMs are also applicable to domains outside of language, such as vision: they are designed to close the gap between theoretically motivated features in any domain (linguistic features, geometric features, etc.) and any system of representations.

6 Conclusions

Linguistic features can be represented in vastly different ways in language models. Metric-Learning Encoding Models show that some features dominate the neural representations, creating large neural distances among sentence representations. These features can be encoded locally in single units, or in multiple, redundant units, or they can be encoded in a more complex, distributed manner. The method is applicable to any measurable neural system (artificial or human), and to any domain (language, vision) with a feature-based theory.

Ethical statement

This paper presents work whose goal is to bridge closer together the fields of Machine Learning and Linguistics; being theoretical in nature, we do not feel like any societal risks need to be specifically highlighted.

Limitations

For simplicity, we assumed here that there are no interactions among linguistic features in predicting neural distances among sentence representations. However, such interactions are common in many problems, including in language. The framework of MLEMs allows a straightforward way to introduce interactions (Section 3.3), while, in contrast to other approaches (such as RSA), it preserves the metric property of the learned distances. The introduction of interaction terms into the study of

linguistic features with MLEMs is therefore an interesting direction for future work.

The question of the type of encoding - local or distributed - has been a major topic of research in neuroscience (Rumelhart et al., 1986; Rolls, 2017; Bowers, 2017). This work shows how this question can be addressed by contrasting univariate and multivariate MLEMs. However, it only provides a proof of concept for a single model. Calibrating this more and extending this with additional models and datasets could in the future reveal how different systems may find different solutions to the same computational problem.

References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Jelle Zuidema. 2019. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. In *Proceedings of the ACL-Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 191–203.

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *ArXiv*, abs/1610.01644.

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for constituency structure in neural language models. *ArXiv*, abs/2204.06201.

Yonatan Belinkov and James R. Glass. 2018. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Jeffrey S Bowers. 2009. On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological review*, 116(1):220.

Jeffrey S Bowers. 2017. Grandmother cells and localist representations: a review of current thinking.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2021. Long-range and hierarchical language predictions in brains and algorithms. *ArXiv*, abs/2111.14232.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5.

Ahmad Chamma, Denis A Engemann, and Bertrand Thirion. 2023. Statistically valid variable importance assessment through conditional permutations. *arXiv preprint arXiv:2309.07593*.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3).

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Tim van Gelder. 1992. Defining 'distributed representation'. *Connection science*, 4(3-4):175–191.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Dieuwke Hupkes and Willem H. Zuidema. 2017. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *J. Artif. Intell. Res.*, 61:907–926.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Kamila M Jozwik, Nikolaus Kriegeskorte, Katherine R Storrs, and Marieke Mur. 2017. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in psychology*, 8:1726.

Philipp Kaniuth and Martin N Hebart. 2022. Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *NeuroImage*, 257:119294.

Jean-Rémi King, Francois Charton, Maxime Oquab, and David Lopez-Paz. 2020a. Measuring causal influence with back-to-back regression: the linear case.

Jean-Rémi King, Laura Gwilliams, Chris Holdgraf, Jona Sassenhagen, Alexandre Barachant, Denis Engemann, Eric Larson, and Alexandre Gramfort. 2020b. Encoding and decoding framework to uncover the algorithms of cognition. *The Cognitive Neurosciences (Sixth Edition)*, 58.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.

Joseph B. Kruskal. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27.

Brian Kulis. 2013. Metric learning: A survey. *Found. Trends Mach. Learn.*, 5:287–364.

Aarre Laakso and Garrison Cottrell. 2000. Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, 13(1):47–76.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. *arXiv preprint arXiv:2010.01869*.

Ashin Mukherjee and Ji Zhu. 2011. Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. 2022. Joint processing of linguistic properties in brains and language models. *ArXiv*, abs/2212.08094.

Urvashi Oswal, Christopher R. Cox, Matthew A. Lambon Ralph, Timothy T. Rogers, and Robert D. Nowak. 2016. Representational similarity learning with application to brain networks. In *International Conference on Machine Learning*.

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. 2022. Neural language models are not born equal to fit brain data, but training helps. In *International Conference on Machine Learning*.

Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, and Christophe Pallier. 2023. Information-restricted neural language models reveal different brain regions' sensitivity to semantics, syntax and context. *Neurobiology of Language*.

Giovanni Puccetti, Alessio Miaschi, and Felice Dell'Orletta. 2021. How do bert embeddings organize linguistic knowledge? In *Proceedings of deep learning inside out (DeeLIO): the 2nd workshop on knowledge extraction and integration for deep learning architectures*, pages 48–57.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Edmund T Rolls. 2017. Cortical coding. *Language, Cognition and Neuroscience*, 32(3):316–329.

David E Rumelhart, James L McClelland, and CORPORATE PDP Research Group. 1986. *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, Ian Tenney, and Ellie Pavlick. 2022. The multiberts: Bert reproductions for robustness analysis.

Roger N Shepard and Susan Chipman. 1970. Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17.

Paul Smolensky. 1990. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216.

Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. 2021. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of cognitive neuroscience*, 33(10):2044–2064.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *ArXiv*, abs/1905.06316.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

A Datasets

In this appendix, we provide a description of the datasets. They are available online, together with the labels and the grammars used to generate and extend them.

Dataset	Template	#features	#sentences
Short Sentences Simple 2/3 words sentences	Subj. verb. I play. Who/Which [subj.] verb? Which princess sang?	9	522
Relative Clause 2 by 2 design Peripheral/Center Embedding (CE) Subj./obj. relative clause (s/oRC)	[CE, sRC]: Subj. who verb obj. verb obj. The woman who sees the princess admires the actress. [Periph., oRC]: Subj. verb obj. who subj. verb. The woman sees the princess who the actress admires. [CE, oRC] ... [Periph., sRC] ...	14	7680
Long-Range Agreement (*) 2 by 2 design Periph./CE Number (in)congruence (NC/NI)	[Periph., NC]: Subj. verb obj. who subj. verb. The woman sees the princess near the actress. [CE, NI] (*): Subj. who verb obj. verb obj. The woman near the princesses sees the actress. [Periph., NI] ... [CE, NC] ...	14	7680
Large	All of the above Embedding under propositional attitude John knows that/whether/who... Quantification and binding theory Everyone sees himself/him/me	25	3120

Table A.1: Brief description of the 4 datasets. The Large dataset contains more linguistic phenomena and features; it is nonetheless smaller because it uses less lexical variability.

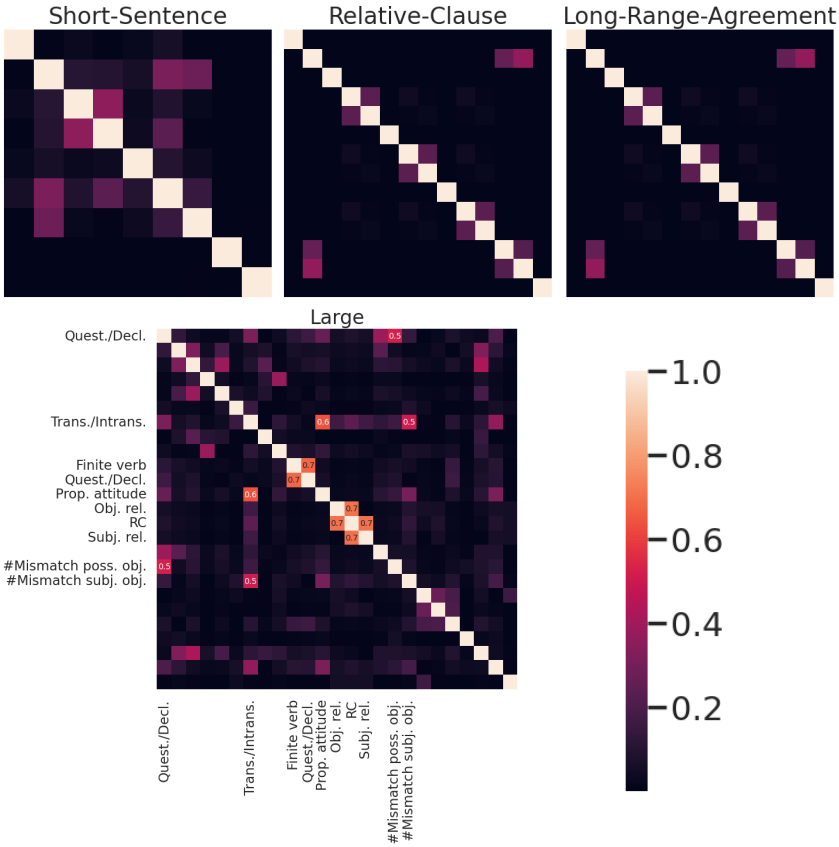

Figure A.1: Correlations between all the features of the 4 datasets in the matrix of pairwise feature distances. Only correlations above 0.5 are annotated. The Large dataset has more features and could benefit from a re-encoding, as some correlations are not negligible.

B MLEM Feature Importance

In this appendix, we provide more information about feature importances across datasets, and even models (showing several seeds of BERT as well as untrained versions).

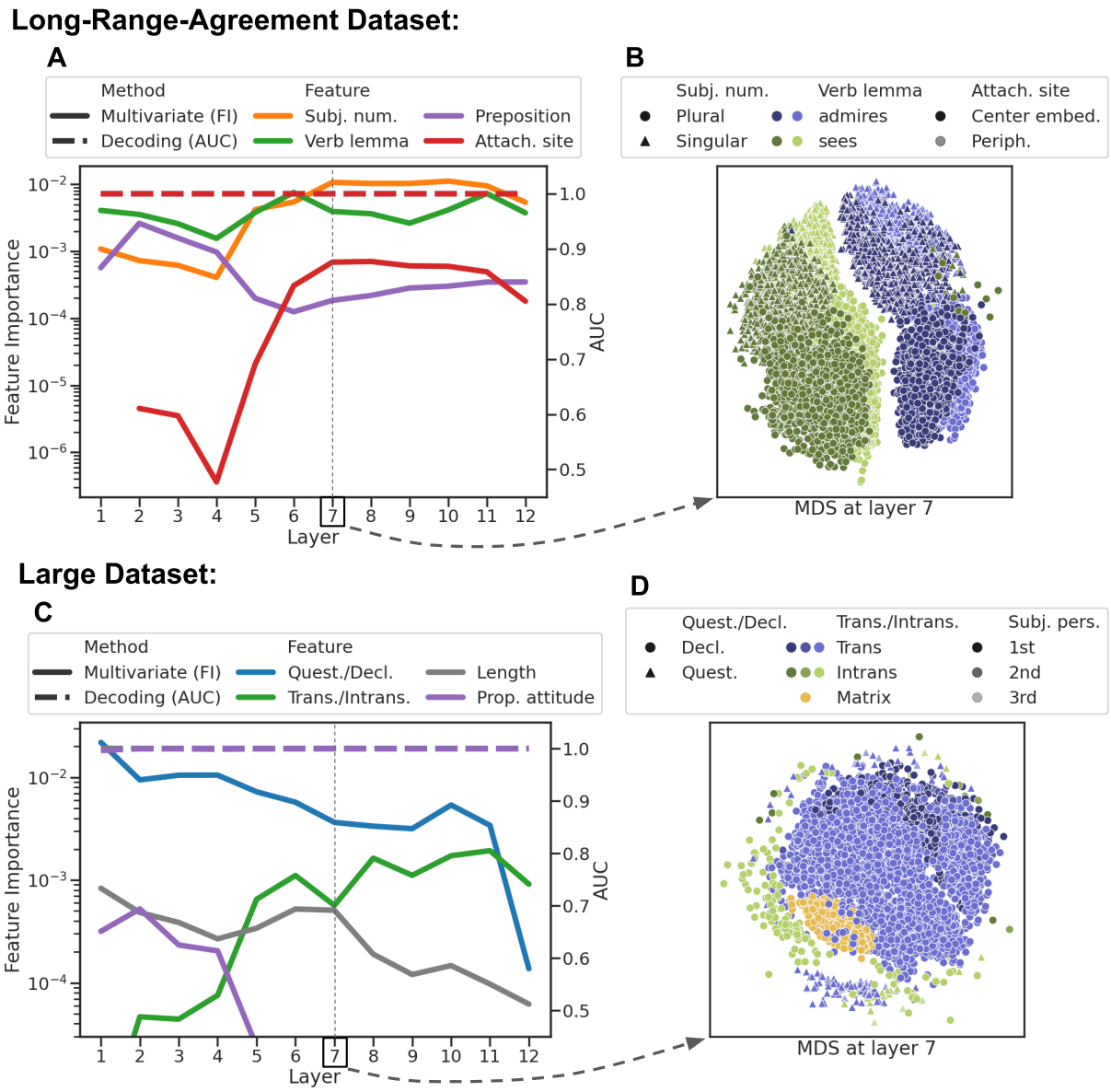

Figure B.1: This is the continuation of Fig. 3 for the remaining datasets: MLEM FIs along with the corresponding decoding AUC for the top 4 features for the **Long-Range Agreement** and **Large** datasets. The MDS plot at layer 8 and 7, respectively, show hierarchical clustering according to *Subj. num.* > *Verb lemma* > *Attach. site* and *Quest./Decl.* > *Trans./Intrans.* > *Subj. pers.*.

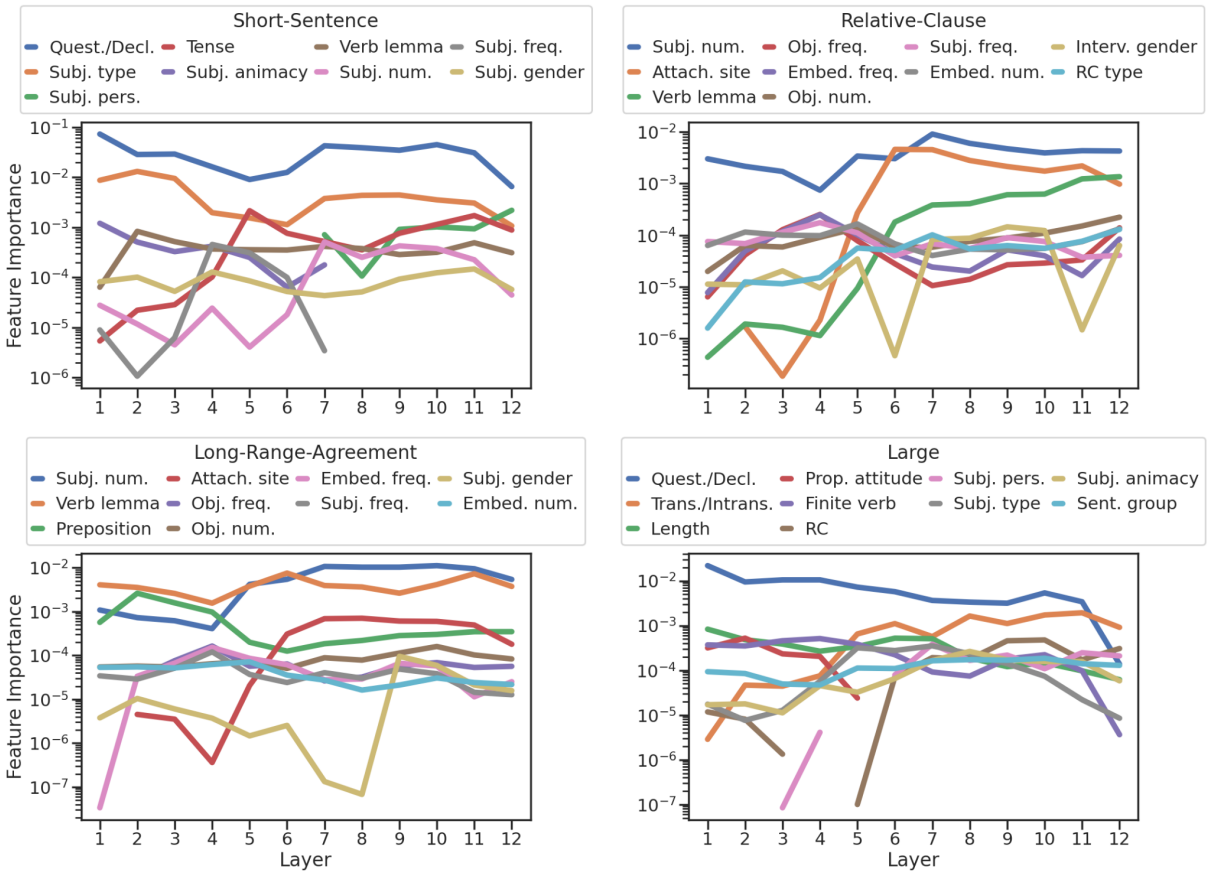

Figure B.2: MLEM FIs with more features displayed (top 10) for all datasets.

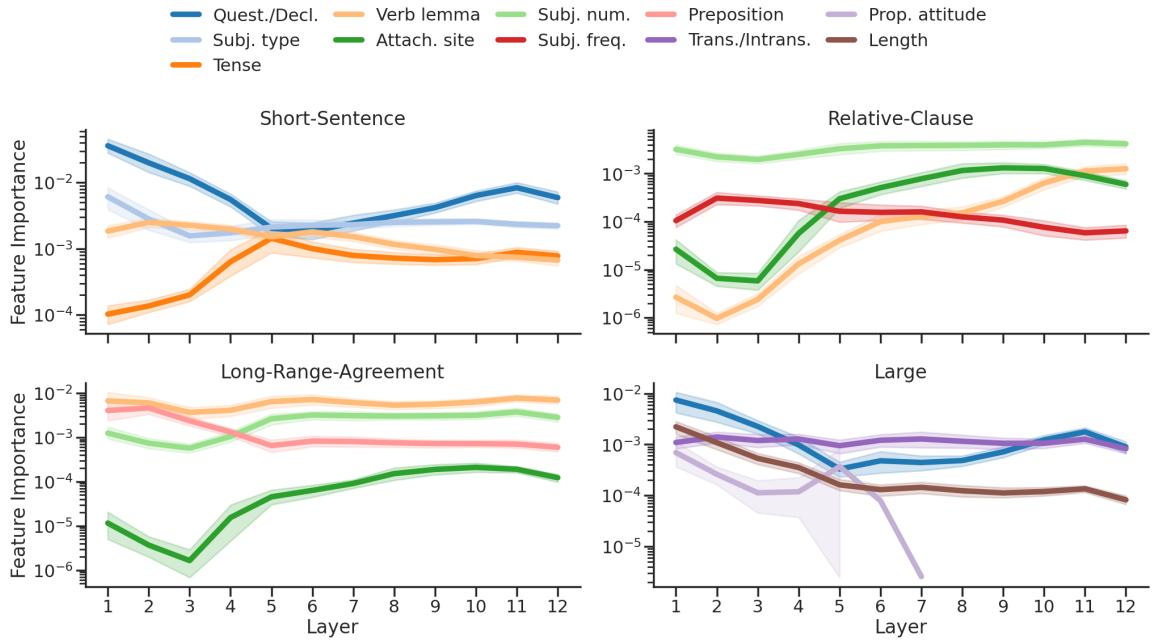

Figure B.3: MLEM FIs for the 25 available seeds of MultiBERTS (Sellam et al., 2022) on the 4 datasets. The average over the seeds is in dark and the 95% confidance interval in light. For readability, only the top 4 features for each dataset are shown.

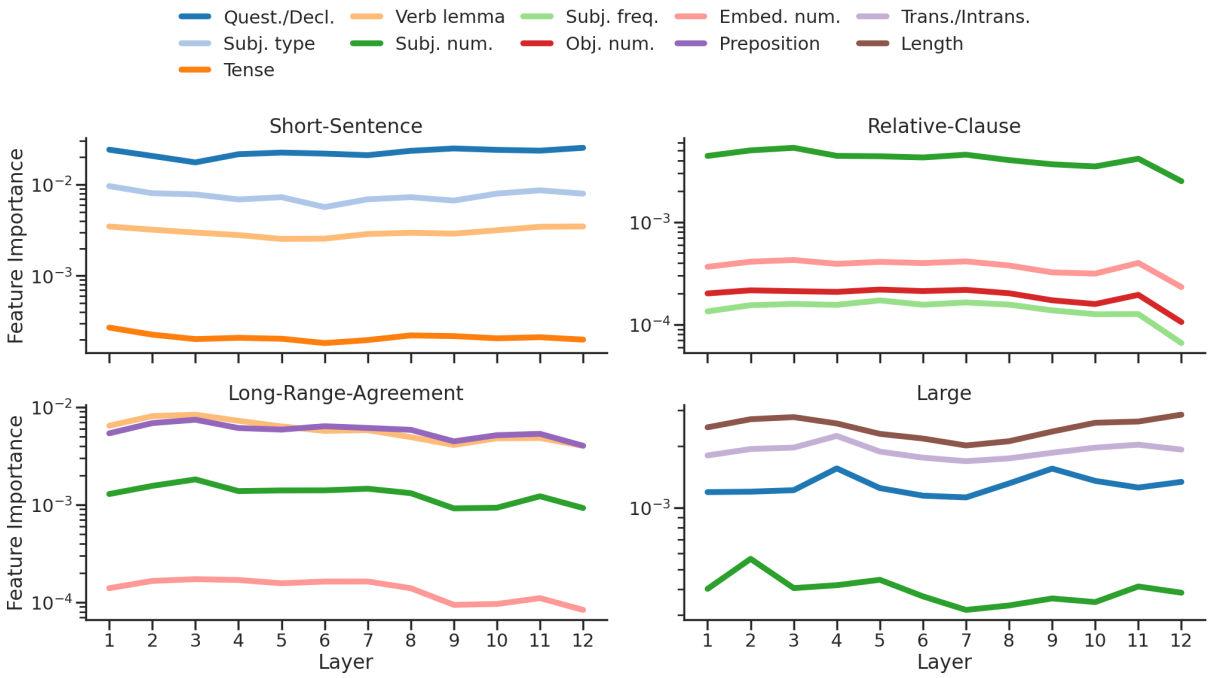

Figure B.4: MLEMs FIs on untrained BERT (multiberts-seed-0-step-0k) for the top 4 features on each dataset.

C Visualization of the features in the representations

In this appendix we offer various visualizations of the way the different features spread in representation space, revealing both order and hierarchical effects.

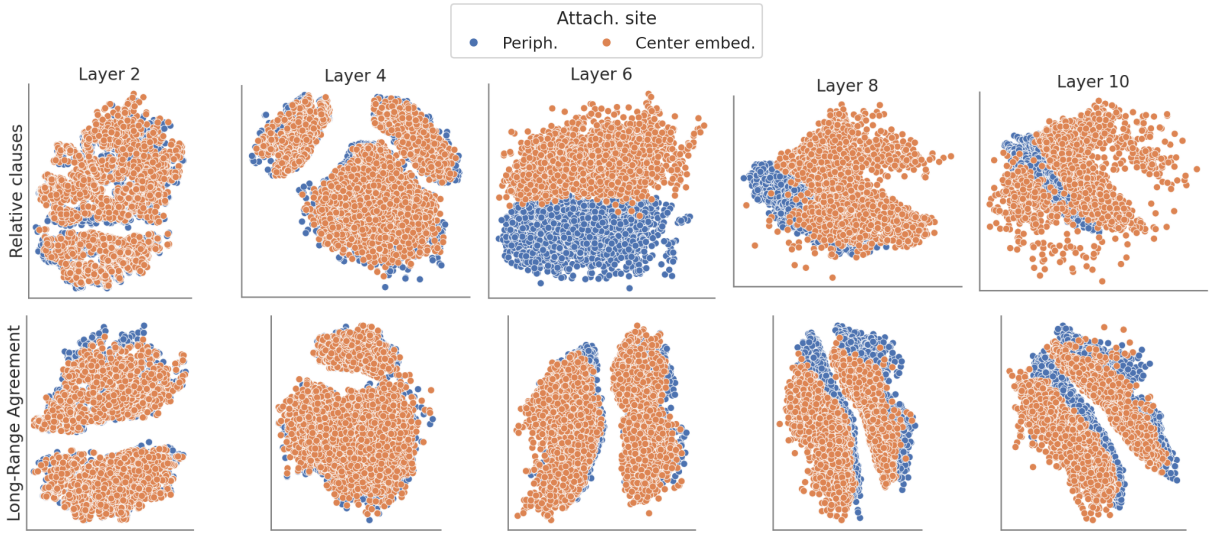

Figure C.1: MDS for the Relative Clause and Long-Range Agreement datasets. It shows the increasing importance of *Attachment site* across layers.

D Univariate vs. multivariate encoding

In this appendix, we show how one can combine the univariate and multivariate analysis to reveal that there is an optimal number of units, between 10 and 100 units across the 768 unit layers, at which maximal fit is achieved. We also show that at each layer the units are grouped in clusters specialized in a feature.

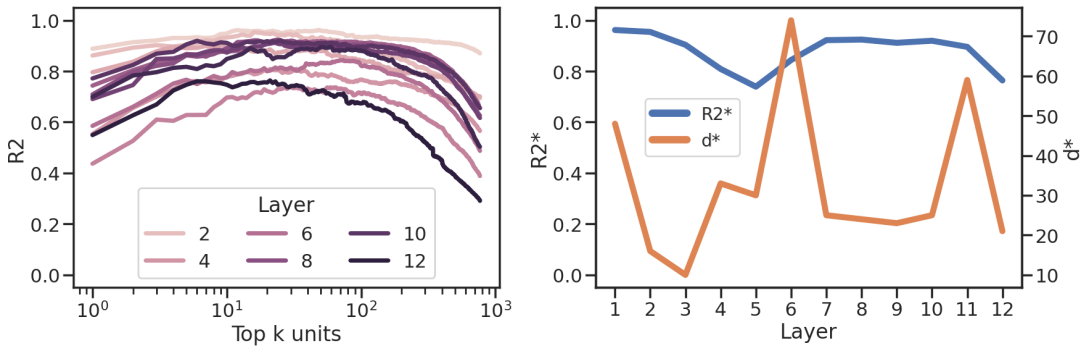

Figure D.1: **Left:** Multivariate R^2 for each layer when considering more and more units, sorted by their univariate R^2. **Right:** Maximum R^2 score achieved for each layer on the previous plot, namely R^{2*}, along with the dimensionality d^*, i.e. the number of units needed to achieve R^{2*}.

E Clusterization of units and feature specialization

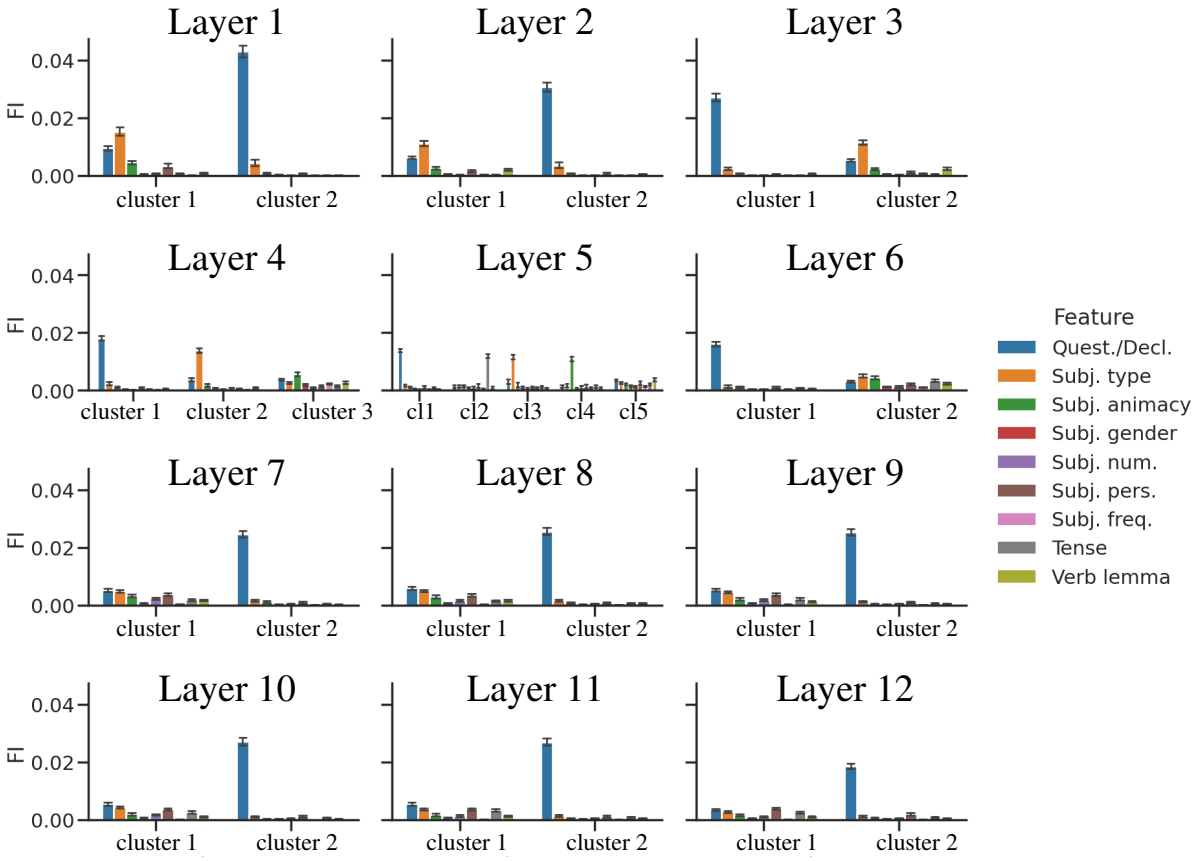

Figure E.1: **Average FIs by cluster obtained by K-means and the silhouette method for each layer.** We see that layers 1-3 have two clusters, one specialized in Quest./Decl. and the other in Subj. type. At layer 4 there are three clusters with a new one for Subj. animacy. Layer 5 has five clusters, a new one for Tense and a catch-all one. Layers 6-12 have two clusters each, one for Quest./Decl. and a catch-all one.