Mapping Transformer Layer Hierarchy to Brain Functional Organization Through Integrated Gradients

Andrea Corsico
Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca
a.corsico@campus.unimib.it

ABSTRACT
Encoding models that map linguistic features to brain activity have traditionally been evaluated on controlled stimuli, leaving open questions about their behavior under naturalistic conditions. In this work, we investigate how the layer-wise representations of GPT-2 align with the functional organization of the human language network using fMRI data from subjects watching the TV series *Friends*. Beyond prediction accuracy, we employ Integrated Gradients (IG) to analyze which embedding dimensions drive the encoding model's predictions for different brain regions.
Contextualized representations substantially outperform static embeddings, with intermediate layers (7–8) achieving peak performance. The IG analysis reveals that mean and max pooling capture complementary information—temporal regions preferentially rely on mean pooling while frontal regions favor max pooling—validating concatenated pooling strategies. Crucially, the Feature Importance Maps reveal an organizational structure aligned with anatomical groupings (frontal vs. temporal) rather than the hypothesized syntactic-semantic dissociation.
These findings demonstrate that Integrated Gradients can illuminate not only which models predict brain activity best, but how their internal representations map onto cortical organization—revealing that anatomical rather than functional principles may govern this alignment under naturalistic conditions.

Key words: NLP, XAI, fMRI Encoding models, GPT-2, Brain-model alignment

1 INTRODUCTION

1.1 Motivation and Background

Encoding models are a standard approach in computational neuroscience for linking features derived from external stimuli to patterns of neural activity. In this framework, one learns a mapping from a stimulus representation—often obtained from NLP or computer vision models—to the corresponding brain responses. This approach naturally connects neuroscience with established areas of deep learning, where pretrained models provide rich and structured embedding spaces suitable for modeling complex cognitive processes [5].

In the intersection between neuroscience and NLP literature, many encoding studies have traditionally relied on controlled stimuli such as isolated words or fixed-length sentences. Within this setting, a central line of inquiry focuses on how internal representations at different network depths align with neural representations in the brain, and how different model architectures (static vs. contextual, recurrent vs. transformer-based) capture distinct levels of linguistic structure. Far fewer studies, however, examine naturalistic stimuli such as movies, narratives, or spontaneous conversation, despite their increasing relevance and ecological validity. Naturalistic paradigms are more challenging: they involve continuous, time-aligned input, highly variable linguistic structure, and the need to aggregate sequences of different lengths into fixed-size vectors that match the hemodynamic resolution of fMRI. At the same time, they offer a more realistic view of how language is processed in everyday contexts and allow testing NLP models under conditions that more closely resemble natural comprehension [9][6].

A further practical challenge in naturalistic settings is that the number of words within each fMRI time window varies substantially. This requires transforming variable-length sequences of embeddings into coherent, fixed-size representations. Pooling operations such as mean or max pooling provide simple and widely used solutions, but they may highlight different aspects of the linguistic input. For this reason, in the present work we also consider how Integrated Gradients can help interpret the contribution of different pooling strategies and the role they play in shaping the model's predictive features.

1.2 Research Question

Building on the motivations presented above, the present project addresses the following research questions:

(i) **Layer-wise comparison in a naturalistic setting.** We aim to extend layer-wise analyses commonly performed on controlled linguistic stimuli by evaluating how the internal representations of a transformer model (GPT-2) align with neural activity under a fully naturalistic paradigm. This includes testing, for each layer, the degree to which its representations predict responses across distinct language-selective regions defined in prior work by Fedorenko and colleagues.

(ii) **Static versus contextual representations.** As a baseline, we compare the performance of contextualized representations to a static embedding model (FastText). This contrast allows us to assess whether layer-wise improvements in prediction arise from contextual integration, from depth-dependent abstraction, or simply from the model's representational capacity.

(iii) **Role of pooling under naturalistic constraints.** Because naturalistic stimuli produce variable amounts of linguistic input within each fMRI time window, we must aggregate word-level embeddings into fixed-size vectors. We compare mean and max pooling and ask whether these operations emphasize different aspects of the stimulus. Integrated Gradients (IG) are used as an interpretability tool to examine how each pooling strategy contributes to the predictions of the encoding model.

(iv) **Layer-dependent evolution of linguistic components.** Beyond prediction accuracy, we use IG to explore whether the relative importance of syntactic versus semantic components of the embedding space changes across the depth of the network, and whether these shifts correspond to the functional characteristics of the brain regions being predicted. This analysis aims to shed light on how linguistic information is reorganized across layers and how this reorganization relates to the selectivity of specific cortical areas.

2 METHODS

2.1 Dataset and Language Brain Regions

For this project we used the linguistic and fMRI subset released as part of the Algonauts 2025 Challenge, derived from the CNeuromod dataset [4]. The stimulus material consists of the first six seasons of the TV series *Friends*: seasons 1–5 were used as training data, while season 6 served as held-out test data. For each episode, time-aligned transcripts were provided in the form of text segments binned into fixed 1.49 s windows, corresponding to the repetition time (TR) of the fMRI acquisition.

In total, the dataset comprises 137,681 TRs, equivalent to approximately 57 hours of language transcripts. However, due to the naturalistic nature of the stimulus, not all TR windows contain spoken dialogue. Figure 1(a) shows the distribution of word counts per TR: approximately one quarter of the TRs contain no words (corresponding to silent periods, non-verbal scenes, or pauses in dialogue). These empty TRs were excluded from all subsequent analyses, yielding a final dataset of 103,313 TRs containing at least one word. Figure 1(b) shows the distribution of GPT-2 tokens per non-empty TR, reflecting the actual input length to the encoding models after tokenization.

This variable-length input is a defining characteristic of naturalistic paradigms and motivates the use of pooling strategies to aggregate token embeddings into fixed-size representations matched to the fMRI acquisition.

The fMRI volumes were supplied in MNI space and parcellated using the Schaefer 1000-parcel atlas (2 mm resolution). Since the full atlas includes parcels from multiple large-scale networks, we restricted the analysis to language-selective cortex following the functional regions identified by Fedorenko and colleagues. To do so, we downloaded the publicly available Fedorenko language masks, transformed them into the same MNI space as the Schaefer atlas, and computed voxelwise overlap between the two. Parcels from the Schaefer-1000 atlas showing at least 5% overlap with a Fedorenko language mask were assigned to the corresponding functional region. This procedure yielded five left-hemisphere macro-ROIs covering the core frontal and temporal components of the human language network [3].

A summary of the five macro-ROIs is reported in Table 1, including their abbreviation, extended anatomical label, number of Schaefer parcels selected from the original atlas, and a brief description of their linguistic function. See Appendix A) for graphical representation of the involved brain reasons.

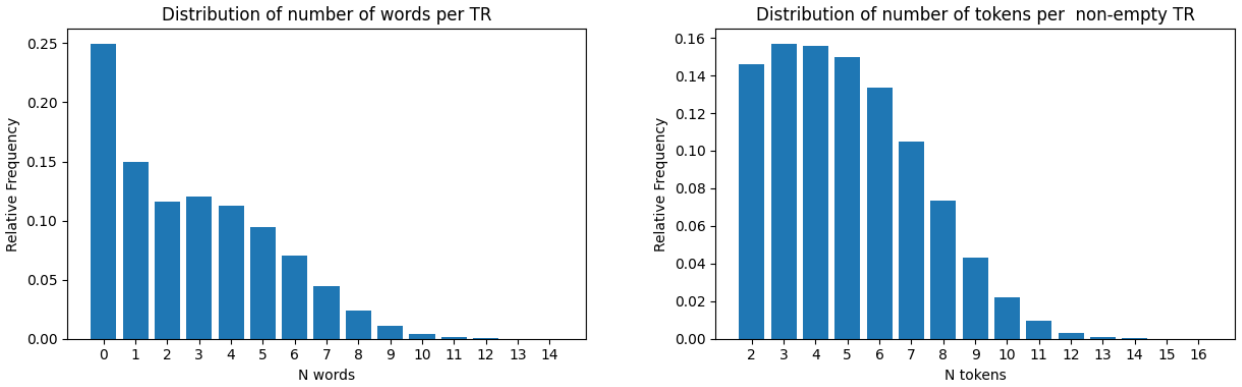

Figure 1. Distribution of linguistic content per TR window. (a) Number of words per TR across the full dataset (137,681 TRs). Approximately 25% of TRs contain no words and were excluded from subsequent analyses. (b) Number of GPT-2 tokens per non-empty TR (103,313 TRs).

Table 1. Summary of chosen Fedorenko Language Network ROIs

ROI	Extended name & localization	Main function
LH_IFG	Left Inferior Frontal Gyrus (pars opercularis)	Syntactic composition; combining phrase structure.
LH_IFGorb	Left Inferior Frontal Gyrus (pars orbitalis)	High-level semantic processing; integration of linguistic meaning.
LH_MFG	Left Middle Frontal Gyrus	Linguistic control processes; verbal working memory.
LH_AntTemp	Left Anterior Temporal Lobe	Semantic composition; combining word meanings.
LH_PostTemp	Left Posterior Temporal Lobe	Syntax–semantics integration.

2.2 Feature Extraction

The linguistic features (embeddings) used as input for the encoding models were extracted from a pre-trained Transformer *decoder-only* model GPT-2 small [8]. This model was chosen because systematic studies have shown that it is the best family of models for this task [9]; the small variant was selected in order to limit the number of comparisons. The primary goal of this feature extraction process was to obtain **contextualized representations** across the network's hierarchy to assess how the depth-dependent organization of the Transformer aligns with the functional organization of the human language network.

2.2.1 Contextualized Feature Generation

The extraction was performed layer-wise, spanning from **Layer 1 through Layer 12**, where the base token embedding dimension is 768. The raw linguistic input was first processed using the **tokenizer specific to the employed Transformer model** (GPT-2 small tokenizer), which converts the text into sub-word units.

A methodological decision was made to **remove all explicit punctuation markers** from the raw text prior to tokenization. While punctuation defines sentence and clause boundaries, this practice is aligned with prior research on neural language modeling which has shown that removing non-lexical elements often **maximizes the effectiveness and predictive power of encoding models** [7]. By prioritizing lexical and functional units, we optimized the $\mathbf{X} \rightarrow \mathbf{Y}$ mapping (correlation r).

Crucially, the syntactic structure is robustly encoded by elements retained in the feature vector: **functional words** (e.g., articles, prepositions, conjunctions), **verb conjugations**, and **morphological features**. Therefore, the Integrated Gradients analysis, particularly in syntax-related ROIs like LH_IFG, investigates the importance of the **contextualized implicit syntactic features** rather than relying on explicit structural markers.

2.2.2 *Temporal Alignment and Pooling*

The naturalistic stimulus material (the TV series *Friends*) presented sequences of variable linguistic input, which had to be converted into fixed-size vectors temporally matched to the fMRI acquisition. This was achieved through the following steps:

(i) **Temporal Matching:** Token embeddings were aligned with the fMRI Repetition Time (**TR**) windows of **1.49 seconds**.
(ii) **HRF Delay:** A temporal offset corresponding to an HRF (Hemodynamic Response Function) delay of **3 TRs** was applied to account for the lag between neural activity and the measured BOLD signal (Figure 2).
(iii) **Aggregation (Pooling):** To transform the variable-length sequences of embeddings within each TR window into a fixed-size vector, two distinct pooling strategies were utilized in parallel:

- **Mean Pooling:** Calculates the arithmetic mean of all token embeddings within the TR window, capturing the average linguistic information of the period.
- **Max Pooling:** Takes the maximum value across the feature dimension for all token embeddings within the TR window, designed to capture the most salient linguistic event or activation peak during the period.

The final feature vector, used as input for the encoding MLP, was generated by **concatenating** the Mean Pooled vector (768 dimensions) and the Max Pooled vector (768 dimensions). This resulted in a comprehensive feature vector of **1536 dimensions** for each time point. This concatenation is central to Research Question (iii), allowing the use of Integrated Gradients to interpret the relative contribution of each pooling strategy to the model's predictions.

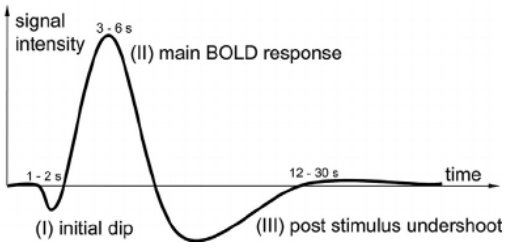

Figure 2. The Hemodynamic Response Function (HRF) describes the delayed and sluggish BOLD signal response to neural activity. The peak occurs approximately 4–6 seconds after stimulus onset, motivating the 3 TR delay used in our temporal alignment.

2.2.3 *Static Baseline*

As a baseline for comparison against the contextualized representations, a static embedding model, **FastText**, was employed. FastText provides 300-dimensional word embeddings, which were processed using the identical temporal alignment and aggregation procedures (Mean/Max pooling and concatenation), resulting in a 600-dimensional feature vector. This ensures a methodologically consistent evaluation of the benefits provided by dynamic contextual integration over static representations [2].

2.3 Preliminary Validation of Embedding Structure

Before applying Integrated Gradients to investigate the syntactic-semantic organization of the embedding space, we conducted a preliminary validation to assess whether the pooled representations preserve any distinction between syntactically-dominant and semantically-dominant linguistic inputs.

For each TR window containing at least four words, we performed Part-of-Speech (POS) tagging using spaCy and computed two scores: a *syntactic score*, defined as the proportion of function words (determiners, prepositions, auxiliaries, pronouns, conjunctions, particles), and a *semantic score*, defined as the proportion of content words (nouns, proper nouns, verbs, adjectives, adverbs, interjections, numerals).

To maximize the contrast between the two linguistic profiles, we applied an extreme filtering procedure: TR windows in the 90th percentile for syntactic score and below the 50th percentile for semantic score were labeled as "syntactic", while TR windows in the 90th percentile for semantic score and below the 50th percentile for syntactic score were labeled as "semantic". From each pool, 1000 samples were randomly selected, yielding a balanced dataset of 2000 TR windows.

For each of the 12 Transformer layers, we extracted the corresponding pooled embeddings (1536 dimensions) and applied

dimensionality reduction via PCA (100 components) followed by t-SNE (perplexity = 30; default value) to visualize the two-dimensional structure of the embedding space (Figure 3).

The t-SNE visualizations revealed limited separability between syntactically-dominant and semantically-dominant TR windows across all layers, as confirmed by low silhouette scores (mean ≈ 0.05). However, visual inspection suggested a partial spatial segregation in the early layers (1–4), with syntactic and semantic samples showing a tendency toward distinct regions of the embedding space. This pattern progressively attenuated in deeper layers (8–12), where the two groups appeared more uniformly intermixed.

These preliminary findings suggest that while the pooled embeddings retain a weak trace of the syntactic-semantic distinction—particularly in early layers—the temporal aggregation process substantially dilutes this separation. This observation anticipates the main IG analysis results, which reveal an organizational structure more aligned with anatomical groupings (frontal vs. temporal regions) than with the hypothesized functional dissociation between syntactic and semantic processing.

Figure 3. t-SNE visualizations of pooled embeddings for each Transformer layer, comparing 1000 syntactically-dominant (blue) and 1000 semantically-dominant (red) TR windows.

2.4 Encoding Models

Our methodology involved training a total of 12 Layers $\times$ 5 ROIs (60 unique combinations per subject) to map linguistic feature vectors ($\mathbf{X}$, 1536 dimensions) to the measured fMRI signal within the target region of interest. The models employed were simple **One-Hidden-Layer Multi-Layer Perceptrons (MLPs)**. This architecture was chosen to maintain interpretability and reduce the risk of overfitting inherent in more complex deep learning models.

2.4.1 *Data Preprocessing and Alignment*

Prior to training, the fMRI and linguistic feature data underwent several critical preprocessing steps:

(i) **Temporal Alignment (HRF Delay):** The linguistic features were temporally shifted to account for the slow nature of the BOLD signal (the Hemodynamic Response Function, HRF). A fixed delay of **3 TRs** was applied, corresponding to approximately **4.5 seconds**, to align the stimulus onset with the peak hemodynamic response in the brain regions.

(ii) **Silence/Gap Removal:** All time points corresponding to periods of silence were removed from both the feature matrix ($\mathbf{X}$) and the fMRI response matrix ($\mathbf{Y}$). This step ensures that the model is only trained on high-information linguistic moments, enhancing the signal-to-noise ratio.

(iii) **Standardization:** After the removal of irrelevant samples, the feature data were standardized. A **StandardScaler** was fitted exclusively on the training data, and this fitted scaler was then applied consistently to both the training and testing sets to prevent data leakage.

2.4.2 *Hyperparameter Optimization*

Hyperparameters for the MLP models were optimized using the framework **Optuna**, which employs a Bayesian optimization approach to efficiently search the configuration space [1]. The goal of the optimization was to maximize the mean Pearson correlation coefficient across the test set. The hyperparameter search space included:

- **Learning Rate (lr):** Suggested on a log-scale (10^{-5} to 10^{-3}).
- **Weight Decay (λ):** Suggested on a log-scale (10^{-4} to 10^{-2}).
- **Hidden Dimension (hidden_dim):** Suggested as a categorical choice ([32, 64, 128, 256, 512, 1024]).
- **Batch Size:** Suggested as a categorical choice ([512, 1024, 2048]).
- **Dropout Rate:** Suggested on a uniform scale (0.4 to 0.6).

2.4.3 *Architectural Consistency Strategy*

A crucial methodological decision was implemented to isolate the impact of the linguistic features (Layer 1 through 12) from the architectural complexity of the models:

(i) **Initial Optimization:** Full hyperparameter optimization using Optuna was performed **only for the features extracted from Layer 1** of the Transformer model, for each of the five target ROIs (e.g., LH_IFGorb, LH_AntTemp, etc.). Models were trained using the Adam optimizer.

(ii) **Architecture Freezing:** The optimal set of architectural hyperparameters (specifically, hidden_dim and dropout_rate) found for Layer 1 features were then **fixed and re-used** for the models trained on the features of Layers 2 through 12 for the corresponding ROI.

This strategy ensures that any observed change in performance or in the feature importance patterns (e.g., analyzed by Integrated Gradients) across the 12 layers is primarily attributable to the **evolution of the linguistic representation itself** (semantic vs. syntactic information at different depths), rather than being confounded by differences in model capacity or regularization across the layers.

2.5 Integrated Gradients Attribution

To analyze the functional relationship between the internal representations of the Transformer model and the measured brain activity, we employed an Explainable AI (XAI) approach using the **Integrated Gradients (IG)** method [10].

2.5.1 *Conceptual Shift: From Token Attribution to Feature Disentanglement*

Traditionally, in NLP, IG is utilized as a retrospective tool to assign an importance score to each input *token*, thereby explaining which word most significantly contributed to a model's output (e.g., a classification or prediction).

In this project, we deliberately **invert this conventional XAI paradigm** to focus the lens of interpretation not on the input data, but on the neural data:

(i) Our **interpretive reference is the known neural function** of the target brain region (ROI)—which is the model's output—rather than the model's predictive accuracy alone.

(ii) The objective shifts from studying token inputs to studying the **abstract, aggregated feature dimensions** (the 1536 feature vector).

This novel application of IG seeks to understand *how* different abstract components within the linguistic embedding are "decoded" by specific functional brain areas. The IG score quantifies the necessary contribution of each 1536 feature dimension for the encoding model to correctly predict the ROI's activation.

This feature-centric attribution analysis directly addresses three core aspects of our Research Questions:

- **Functional Disentanglement:** To ascertain whether features hypothesized to encode specific linguistic components (e.g., syntactic dimensions) are systematically more important for predicting the activity of regions known for that function (e.g., LH_IFG) compared to regions associated with other functions (e.g., semantic processing in LH_AntTemp).
- **Pooling Strategy Preference:** To determine if the ROI's predictive signal is predominantly driven by features derived from **Mean Pooling** or **Max Pooling**, thereby establishing whether the brain region is more sensitive to the average information content or the peak saliency within the TR window.
- **Hierarchical Evolution:** To trace how the relative importance of these features evolves and specializes across the 12 layers of the Transformer architecture.

2.5.2 Integrated Gradients Calculation Methodology

Integrated Gradients are computed by numerically approximating the path integral of the prediction function's gradient along a straight-line path from a designated *baseline* (a neutral input) to the actual input instance.

(i) **Target Function (Wrapper):** Due to the multi-voxel nature of the fMRI signal, the model's target function was defined as the **mean output across all voxels** that comprise the target ROI.

$$F_{ROI}(\mathbf{X}) = \text{Mean}(\text{MLP}(\mathbf{X})) \tag{1}$$

where $\mathbf{X}$ is the 1536-dimensional feature vector.

(ii) **Baseline Selection:** The baseline was defined as the **mean vector of the training data set** $\mathbf{X}_{\text{train}}$. Given the preceding standardization of the data, this vector approximates a zero vector, serving as a non-informative, neutral starting point for the attribution path.

(iii) **Parameters and Convergence:** The numerical approximation of the integral was performed using $N_{\text{steps}} = 80$.

(iv) **Memory Optimization (Batching):** To circumvent the excessive memory allocation required by calculating gradients for the entire test set simultaneously, the IG process was implemented with a **mini-batching** strategy (e.g., BATCH_SIZE $= 2048$). Each batch was processed independently, and the resulting attributions were concatenated before final aggregation, ensuring mathematical equivalence while maintaining computational feasibility.

(v) **Final Feature Importance Map (FIM):** The final FIM for each of the 60 Layer $\times$ ROI combinations was derived by: (a) computing the mean IG across all test samples (FIM$_S$) for each subject, and (b) aggregating the mean of FIM$_S$ across all subjects.

3 RESULTS AND DISCUSSION

3.1 Layer-Wise Encoding Performance

Figure 4 shows the encoding performance (mean voxel correlation) across all 12 layers of GPT-2 and the FastText baseline for each ROI. Shaded areas represent $\pm$ SEM across subjects.

3.1.1 Overall Performance

All ROIs showed substantial improvement when using GPT-2 representations compared to the static FastText baseline, confirming the advantage of contextualized embeddings for predicting brain activity in language regions. The improvement was most pronounced for temporal regions, which approximately doubled their correlation values from FastText to the best-performing GPT-2 layers.

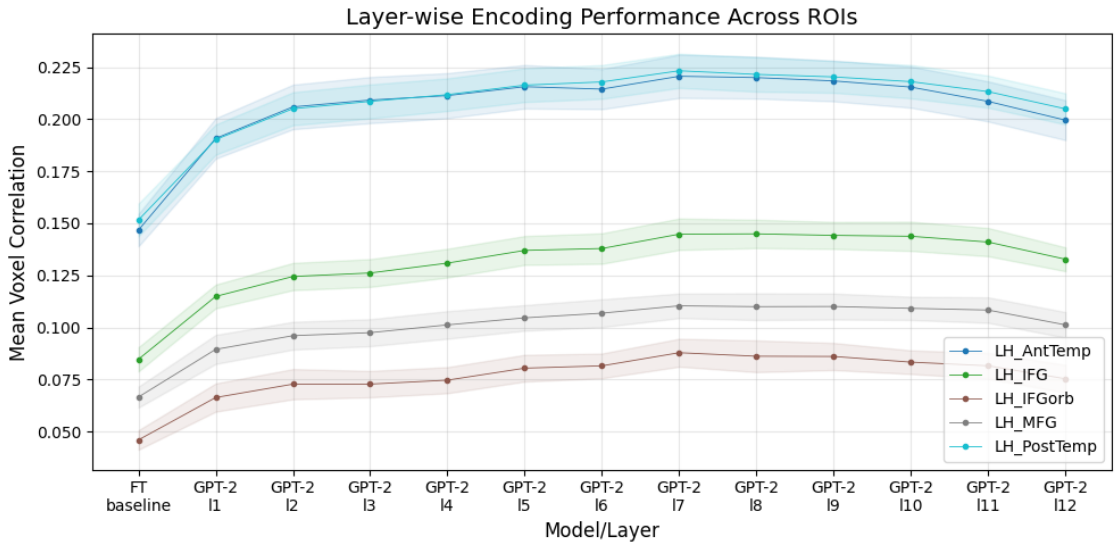

Figure 4. Layer-wise encoding performance across ROIs. Lines show mean voxel correlation for each layer of GPT-2 and the FastText baseline. Shaded areas represent ± SEM across subjects. Temporal regions (LH_AntTemp, LH_PostTemp) achieve the highest correlations, with peak performance at layers 7–8.

3.1.2 Regional Differences

A clear hierarchy emerged across ROIs: temporal regions (LH_AntTemp, LH_PostTemp) achieved the highest correlations ($r \approx 0.21$–0.23 at peak), followed by LH_IFG ($r \approx 0.15$), LH_MFG ($r \approx 0.11$), and LH_IFGorb ($r \approx 0.09$). This ranking was consistent across all layers and suggests that temporal regions are more predictable from linguistic embeddings than frontal regions, possibly reflecting differences in the nature of information processed by these areas.

3.1.3 Layer-Wise Trajectory

All ROIs exhibited a similar layer-wise pattern: a steep increase from layer 1 to layer 2, continued improvement through the middle layers, a peak around layers 7–8, followed by a gradual decline in the final layers (10–12). This inverted-U pattern is consistent with previous findings suggesting that intermediate layers capture the most brain-relevant linguistic representations, while early layers encode lower-level features and late layers may overfit to the language modeling objective.

3.2 Pooling Strategy Analysis

To investigate whether mean and max pooling capture different aspects of the linguistic input, we analyzed the Feature Importance Maps (FIM) derived from Integrated Gradients across all layer-ROI combinations.

3.2.1 Relative Contribution of Pooling Methods

For each layer-ROI combination, we computed the relative contribution of mean pooling by summing the absolute IG values for the mean pooling dimensions (1–768) and dividing by the total importance across all 1536 dimensions.

Figure 5 shows a clear dissociation between brain regions: temporal regions (LH_AntTemp, LH_PostTemp) showed a consistent preference for mean pooling ($> 50\%$), reaching approximately 60% in layer 12 for LH_AntTemp. Frontal regions (LH_IFGorb, LH_IFG, LH_MFG) exhibited the opposite pattern, favoring max pooling in early-to-mid layers before converging toward balance in deeper layers.

Notably, this dissociation follows an anatomical rather than functional organization. LH_IFGorb and LH_AntTemp are both associated with semantic processing, yet they exhibit opposite pooling preferences. This suggests that pooling preference reflects regional signal characteristics rather than a syntactic-semantic distinction.

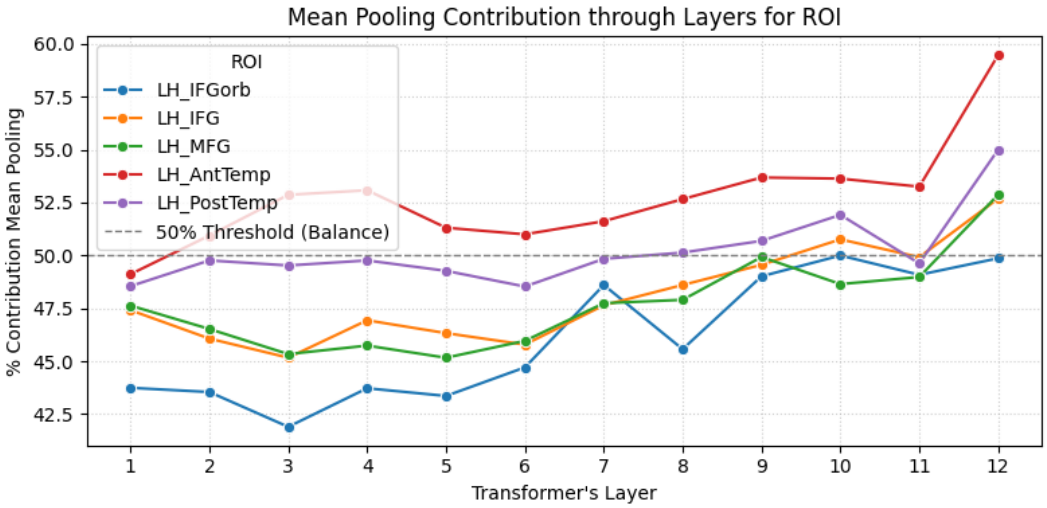

Figure 5. Relative contribution of mean pooling across layers for each ROI. Values above 50% indicate preference for mean pooling, below 50% indicate preference for max pooling. Temporal regions show consistent mean pooling preference, while frontal regions favor max pooling in early layers.

3.2.2 Statistical Validation

Permutation tests (10,000 permutations, KS statistic) revealed significant distributional differences ($p < 0.05$) primarily in early-to-mid layers (1–6), particularly for LH_IFGorb and LH_PostTemp. The effect attenuated in deeper layers (8–12), which typically yield the best encoding performance (see Appendix B, Figure 9).

However, absence of distributional differences does not imply redundancy. Spearman correlations between mean and max FIM vectors remained low across all combinations (range: -0.02 to 0.22, mean ≈ 0.10), as did Jaccard indices for top-100 feature overlap (range: 0.07 to 0.26, mean ≈ 0.13; see Appendix B, Figures 10 and 11).

3.2.3 Summary

These results demonstrate that mean and max pooling capture fundamentally different aspects of the linguistic signal: different brain regions preferentially rely on different pooling strategies, and the features driving predictions show minimal overlap between methods. The concatenation of both pooling types therefore appears to be a well-motivated methodological choice, allowing the encoding model to leverage complementary information that would be lost using either strategy alone.

3.3 Feature Attribution Across Brain Regions

Having established that mean and max pooling capture complementary information, we next examined how the importance patterns of embedding dimensions relate to the functional organization of the language network. Specifically, we asked whether the FIM vectors reveal a syntactic-semantic dissociation across ROIs, or whether a different organizational principle emerges.

3.3.1 Similarity of Feature Patterns Across ROIs

To quantify the similarity of feature importance patterns between brain regions, we computed pairwise Spearman correlations between L1-normalized FIM vectors for each layer. Figure 6 (left) shows the mean correlation matrix averaged across all 12 layers.

A clear clustering emerged: temporal regions (LH_AntTemp, LH_PostTemp) exhibited high mutual correlation ($r \approx 0.78$), indicating that they rely on similar embedding dimensions. Frontal regions (LH_IFGorb, LH_IFG, LH_MFG) formed a separate cluster with moderate internal correlations ($r \approx 0.22$–0.37). Cross-cluster correlations remained low ($r \approx 0.17$–0.19), indicating that frontal and temporal regions emphasize largely distinct dimensions of the embedding space.

To further probe whether this organization reflects functional or anatomical principles, we examined three key ROI pairings (Figure 6, right):

• **IFG ↔ AntTemp** (blue): This pairing represents the hypothesized syntactic-semantic dissociation, with IFG associated with syntactic processing and AntTemp with semantic processing. As expected under a functional dissociation hypothesis, their correlation is low ($r \approx 0.19$–0.25).

• **IFGorb ↔ AntTemp** (green): Both regions are associated with semantic processing, representing a functional association. If the organization were driven by function, we would expect high correlation. Instead, we observe similarly low values ($r \approx 0.13$–0.25), indistinguishable from the dissociated pair.

• **AntTemp ↔ PostTemp** (orange): These regions share anatomical location (temporal cortex) but have distinct functional profiles—AntTemp is primarily semantic, while PostTemp is involved in syntactic-semantic integration. Despite this functional difference, they exhibit the highest correlation ($r \approx 0.78$).

This pattern is inconsistent with a purely functional organization: regions with matching functions (IFGorb and AntTemp) do not show elevated correlation, while regions with shared anatomy but different functions (AntTemp and PostTemp) correlate strongly. Notably, even the AntTemp-PostTemp correlation shows a slight dip in intermediate layers (6–8), precisely where encoding performance peaks—suggesting that the most predictive representations may capture more region-specific information.

These results indicate that anatomical location, rather than putative linguistic function, is the dominant organizing principle governing how different brain regions weight the embedding dimensions.

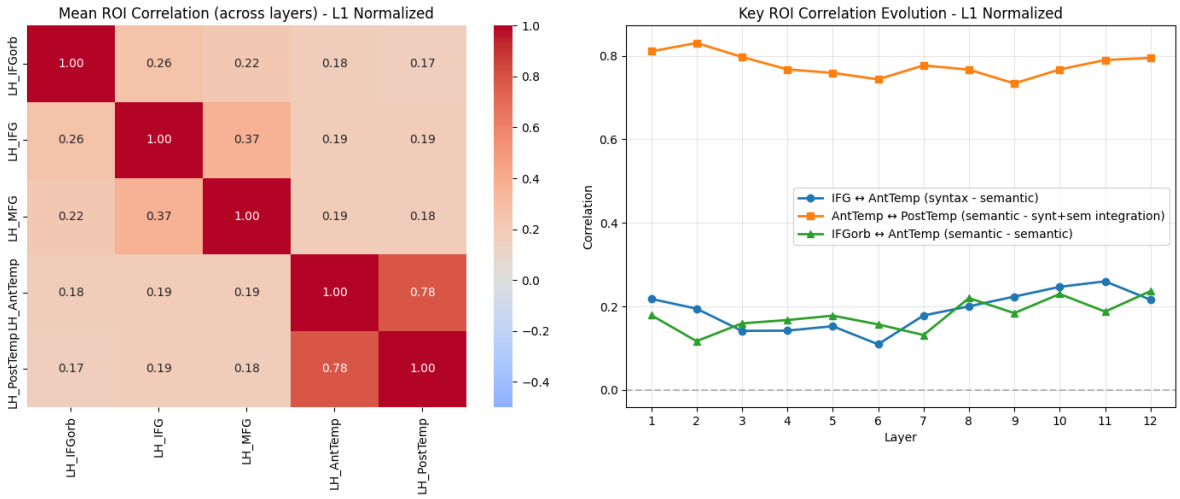

Figure 6. Similarity of feature importance patterns across ROIs. **Left:** Mean correlation matrix (L1-normalized FIM vectors) averaged across all 12 layers, showing a clear frontal-temporal clustering. **Right:** Layer-wise evolution of correlations for three key pairings testing functional vs. anatomical organization. The functional dissociation (IFG-AntTemp, blue) and functional association (IFGorb-AntTemp, green) show similarly low correlations, while the anatomical association (AntTemp-PostTemp, orange) shows high correlation despite functional differences.

3.3.2 Temporal Stability of Feature Patterns

We next examined how stable the FIM patterns are across the Transformer hierarchy within each ROI. For each region, we computed pairwise Spearman correlations between FIM vectors of different layers (see Appendix C, Figure 12).

A striking difference emerged: temporal regions (LH_AntTemp, LH_PostTemp) showed broader diagonal bands, with correlations of 0.50–0.60 persisting even at 3–4 layers distance. In contrast, frontal regions (LH_IFGorb, LH_IFG, LH_MFG) exhibited rapid decorrelation—layer 12 showed near-zero correlation with early layers ($r < 0.10$).

This suggests that temporal regions are sensitive to more abstract, layer-invariant features, while frontal regions capture layer-specific transformations that evolve substantially through the network hierarchy.

3.3.3 Dimension Selectivity Analysis

Finally, we investigated whether individual embedding dimensions show selectivity for specific ROIs. We focused on LH_IFG and LH_AntTemp as these two regions represent the core of the hypothesized syntactic-semantic dissociation: IFG is primarily

associated with syntactic composition, while AntTemp is associated with semantic composition. If embedding dimensions encode distinct linguistic functions, we would expect to find dimensions that are selectively important for one region but not the other.

For each of the 768 dimensions, we computed a selectivity index comparing importance for LH_IFG versus LH_AntTemp, combining contributions from both pooling methods (Figure 7). The selectivity index was defined as:

$$\text{Selectivity}_d = \frac{|IG_{\text{AntTemp},d}| - |IG_{\text{IFG},d}|}{|IG_{\text{AntTemp},d}| + |IG_{\text{IFG},d}|} \tag{2}$$

where positive values indicate AntTemp-selective dimensions (putatively semantic) and negative values indicate IFG-selective dimensions (putatively syntactic).

The distribution of selectivity scores was approximately symmetric around zero (Figure 7, bottom), with 10th percentile at -0.23 (IFG-selective) and 90th percentile at $+0.24$ (AntTemp-selective). While the tails indicate that some dimensions show preferential importance for one region, the majority cluster near zero, indicating moderate rather than strong selectivity.

The heatmap (Figure 7, top) shows that selectivity patterns remain stable across layers: dimensions that are IFG-selective in early layers tend to remain so in deeper layers, and vice versa. This stability suggests that the regional specialization of embedding dimensions is an intrinsic property of the representation, not a layer-dependent phenomenon.

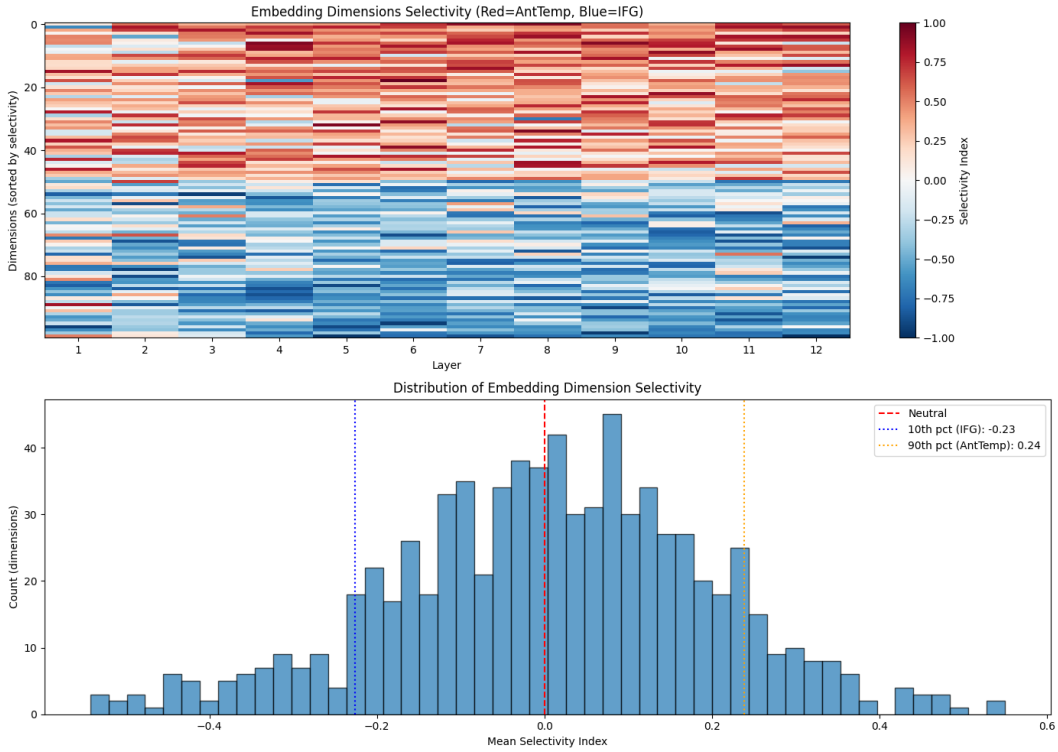

Figure 7. Dimension selectivity analysis comparing LH_IFG (syntactic composition) and LH_AntTemp (semantic composition). **Top:** Selectivity index for the 100 most selective dimensions (50 IFG-selective, 50 AntTemp-selective) across layers; horizontal consistency indicates stable selectivity. **Bottom:** Distribution of mean selectivity across all 768 dimensions; the majority cluster near zero, indicating moderate rather than strong functional specialization.

3.3.4 Summary

The IG analysis reveals an organizational structure aligned with anatomical groupings rather than the hypothesized syntactic-semantic dissociation. Frontal and temporal regions form distinct clusters that rely on different embedding dimensions, but this division does not map cleanly onto syntactic versus semantic processing—both clusters include regions associated with semantic functions (e.g., LH_IFGorb and LH_AntTemp). Temporal regions show more stable feature patterns across the Transformer hierar-

chy, while frontal regions exhibit greater layer-specific specialization. Individual dimensions show moderate selectivity for specific ROIs, but strong dissociation between putatively syntactic and semantic dimensions was not observed.

4 CONCLUSIONS

This work investigated the alignment between Transformer layer representations and the functional organization of the human language network using Integrated Gradients as an interpretability tool under a fully naturalistic paradigm.

4.1 Main Findings

Our results provide clear answers to the four research questions posed at the outset. First, contextualized representations from GPT-2 substantially outperformed static FastText embeddings across all language-selective ROIs, with intermediate layers (7–8) achieving peak encoding performance. This inverted-U pattern confirms that middle layers capture the most brain-relevant linguistic information, consistent with prior work on controlled stimuli.

Second, the analysis of pooling strategies revealed that mean and max pooling capture fundamentally complementary information. Temporal regions preferentially rely on mean pooling, while frontal regions favor max pooling in early layers. Crucially, this dissociation follows anatomical rather than functional lines—regions with similar putative functions (e.g., LH_IFGorb and LH_AntTemp, both linked to semantic processing) exhibited opposite pooling preferences.

Third, and perhaps most importantly, the hypothesized syntactic-semantic dissociation did not emerge from the IG analysis. Instead, the Feature Importance Maps revealed a robust anatomical organization: frontal and temporal regions form distinct clusters that rely on largely non-overlapping embedding dimensions, but this division does not map onto syntactic versus semantic processing. Temporal regions showed stable feature patterns across layers, while frontal regions exhibited rapid layer-specific transformations.

4.2 Implications

These findings have two key implications. Methodologically, they validate the concatenation of mean and max pooling as a principled strategy for naturalistic encoding models, rather than an arbitrary design choice. Theoretically, they suggest that the classical syntactic-semantic dichotomy may be too coarse to capture how Transformer representations align with brain organization—or that the temporal aggregation inherent in naturalistic paradigms dilutes finer-grained functional distinctions.

4.3 Limitations and Future Directions

Several limitations should be acknowledged. The use of a single model (GPT-2 small) and a single naturalistic stimulus (*Friends*) limits generalizability. The POS-based proxy for syntactic versus semantic content is imperfect, and the absence of ground-truth labels for embedding dimensions precludes definitive claims about what individual features encode. Future work could extend this approach to encoder models (e.g., RoBERTa), employ controlled stimuli with parametric manipulation of syntactic complexity, or leverage probing classifiers to independently characterize embedding dimensions before relating them to brain activity.

Despite these limitations, this work demonstrates that Integrated Gradients can serve as a powerful lens for understanding brain-model alignment—not merely by asking which model predicts best, but by revealing how the internal structure of representations maps onto the functional architecture of the brain.

REFERENCES

[1] Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama, 2019, Optuna: A next-generation hyperparameter optimization framework: International Conference on Knowledge Discovery & Data Mining, 2623–2631.

[2] Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov, 2017, Enriching word vectors with subword information: Transactions of the Association for Computational Linguistics, **5**, 135–146.

[3] Fedorenko, E., P.-J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, and N. Kanwisher, 2010, New method for fmri investigations of language: defining rois functionally in individual subjects: Journal of neurophysiology, **104**, 1177–1194.

[4] Gifford, A. T., D. Bersch, M. St-Laurent, B. Pinsard, J. Boyle, L. Bellec, A. Oliva, G. Roig, and R. M. Cichy, 2025, The algonauts project 2025 challenge: How the human brain makes sense of multimodal movies.

[5] Kietzmann, T. C., P. McClure, and N. Kriegeskorte, 2019, Deep neural networks in computational neuroscience.

[6] Lamarre, M., C. Chen, and F. Deniz, 2022, Attention weights accurately predict language representations in the brain: bioRxiv, 2022–12.

[7] Lamprou, Z., F. Pollick, and Y. Moshfeghi, 2022, Role of punctuation in semantic mapping between brain and transformer models: International Conference on Machine Learning, Optimization, and Data Science, Springer, 458–472.

[8] Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, 2019, Language models are unsupervised multitask learners: OpenAI blog, **1**, 9.

[9] Schrimpf, M., I. A. Blank, G. Tuckute, C. Kauf, E. A. Hosseini, N. Kanwisher, J. B. Tenenbaum, and E. Fedorenko, 2021, The neural architecture of language: Integrative modeling converges on predictive processing: Proceedings of the National Academy of Sciences, **118**, e2105646118.

[10] Sundararajan, M., A. Taly, and Q. Yan, 2017, Axiomatic attribution for deep networks: International conference on machine learning, PMLR, 3319–3328.

APPENDIX

A Language-Selective ROIs

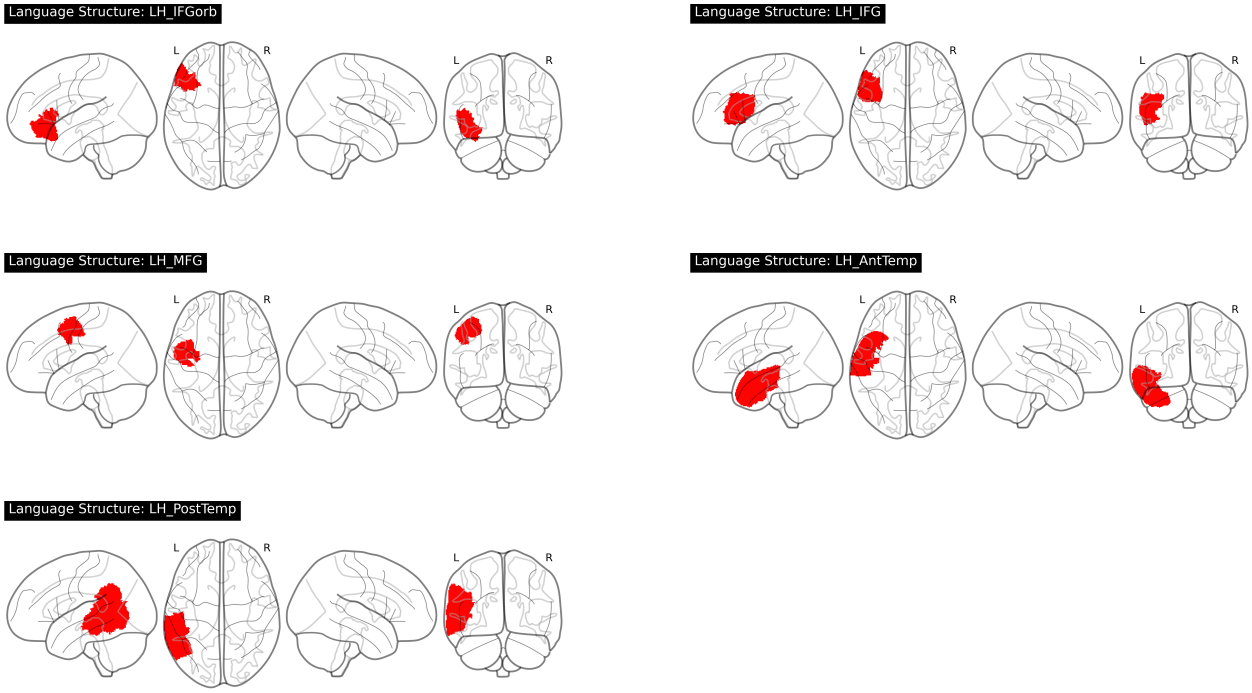

Figure 8. Visualization of the five language-selective ROIs in the left hemisphere, derived from Fedorenko language masks overlaid on the Schaefer-1000 parcellation. Top row: LH_IFGorb (left), LH_IFG (right). Middle row: LH_MFG (left), LH_AntTemp (right). Bottom: LH_PostTemp.

B Pooling Complementarity Analysis

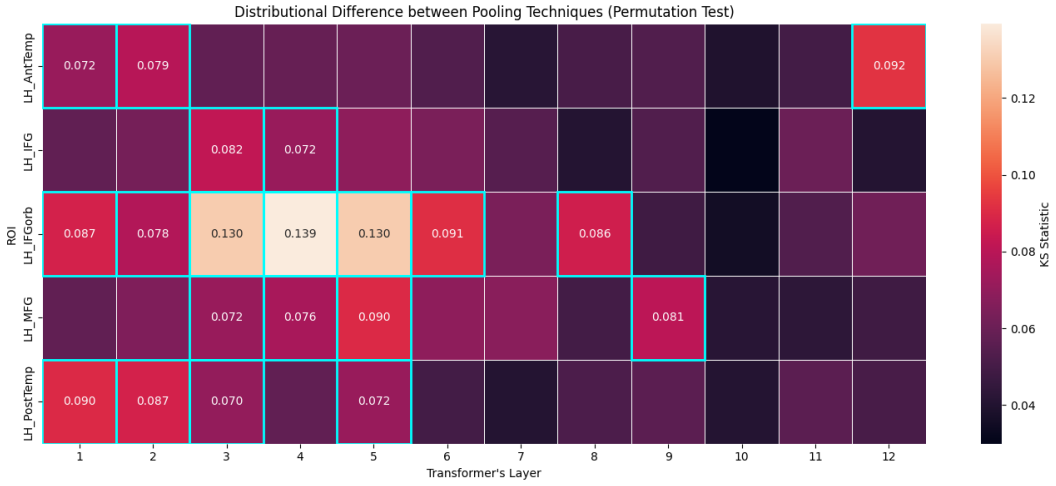

Figure 9. Distributional difference between pooling techniques (permutation test). Cells show KS statistic values; cyan borders indicate significant differences ($p < 0.05$).

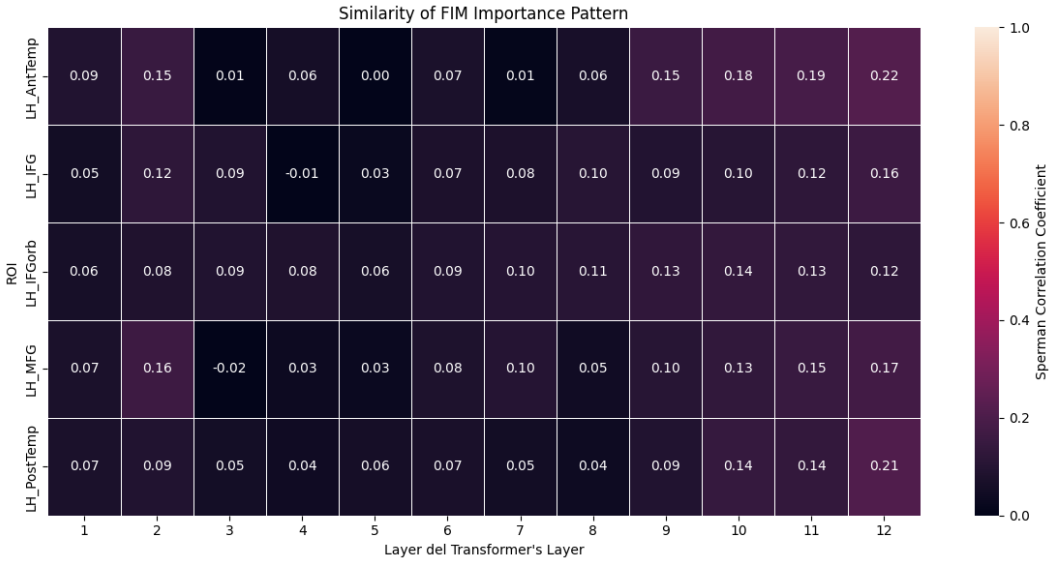

Figure 10. Spearman correlation between mean and max pooling FIM vectors for each layer-ROI combination.

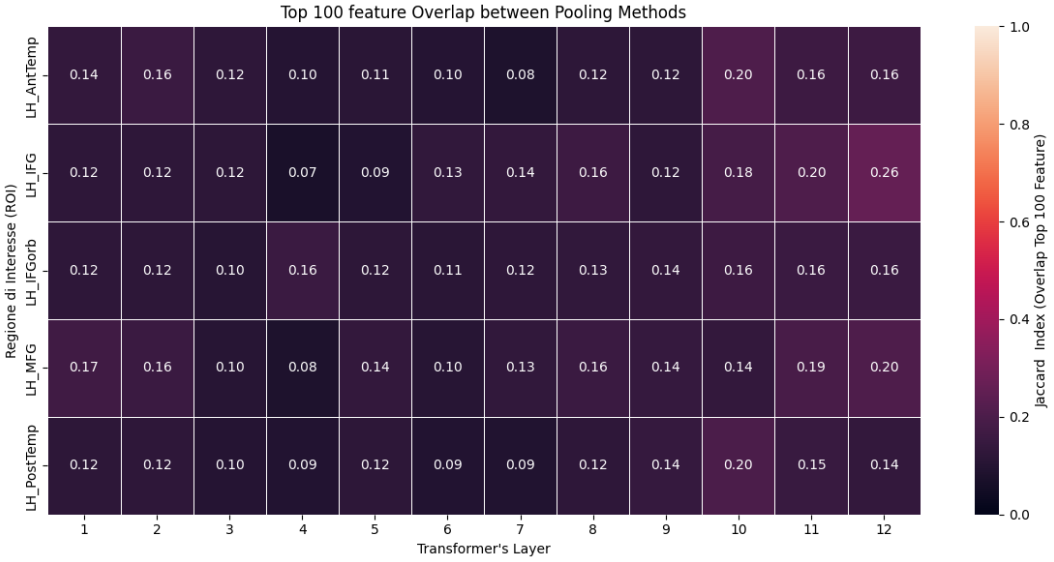

Figure 11. Jaccard index measuring the overlap of top-100 most important features between mean and max pooling.

C Temporal Stability Analysis

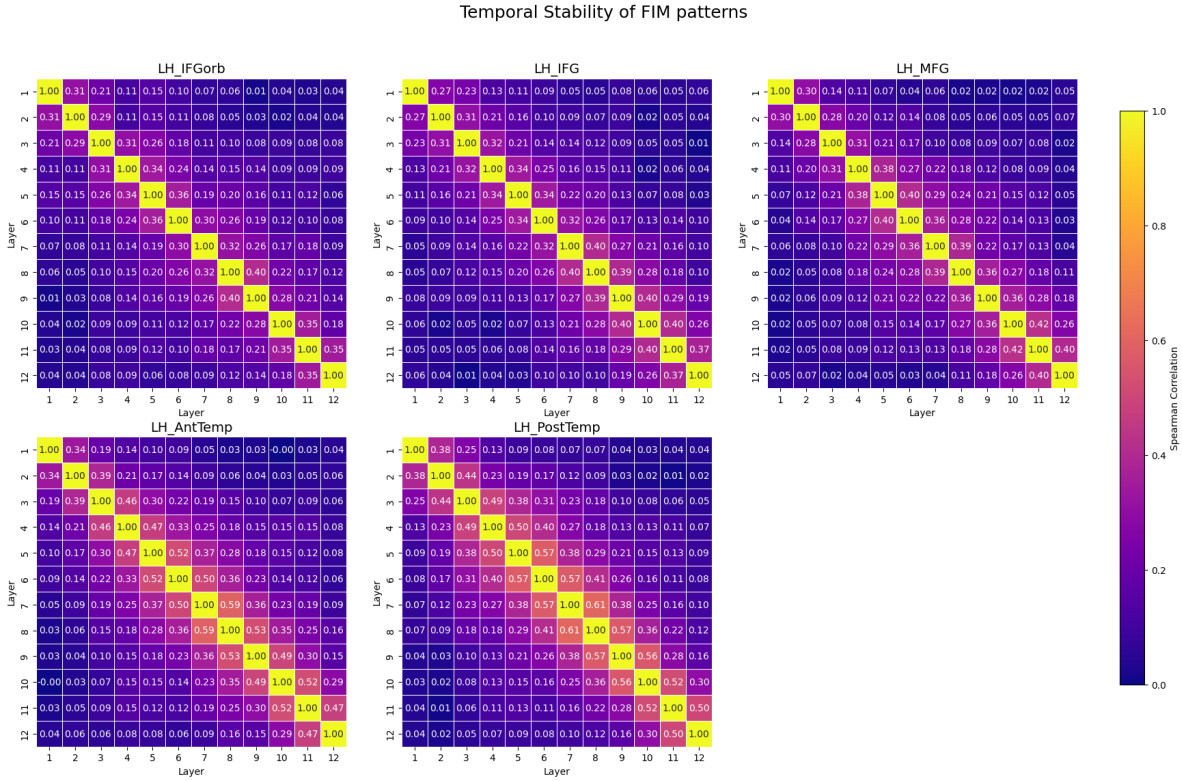

Figure 12. Temporal stability of FIM patterns across the Transformer hierarchy for each ROI.