frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 31 January 2025
DOI 10.3389/frai.2025.1477246

OPEN ACCESS

EDITED BY
Leonie Weissweiler,
The University of Texas at Austin, United States

REVIEWED BY
Wesley Scivetti,
Georgetown University, United States
Qing Yao,
The University of Texas at Austin, United States

*CORRESPONDENCE
Patrick Krauss
✉ patrick.krauss@fau.de

RECEIVED 07 August 2024
ACCEPTED 08 January 2025
PUBLISHED 31 January 2025

CITATION
Ramezani P, Schilling A and Krauss P (2025) Analysis of argument structure constructions in the large language model BERT.
Front. Artif. Intell. 8:1477246.
doi: 10.3389/frai.2025.1477246

Analysis of argument structure constructions in the large language model BERT

Pegah Ramezani[1,2], Achim Schilling[2,3] and Patrick Krauss[2,3]*

[1]Department of English and American Studies, University of Erlangen-Nuremberg, Erlangen, Germany, [2]Pattern Recognition Lab, Cognitive Computational Neuroscience Group, University of Erlangen-Nuremberg, Erlangen, Germany, [3]Neuroscience Lab, University Hospital Erlangen, Erlangen, Germany

Understanding how language and linguistic constructions are processed in the brain is a fundamental question in cognitive computational neuroscience. In this study, we investigate the processing and representation of Argument Structure Constructions (ASCs) in the BERT language model, extending previous analyses conducted with Long Short-Term Memory (LSTM) networks. We utilized a custom GPT-4 generated dataset comprising 2000 sentences, evenly distributed among four ASC types: transitive, ditransitive, caused-motion, and resultative constructions. BERT was assessed using the various token embeddings across its 12 layers. Our analyses involved visualizing the embeddings with Multidimensional Scaling (MDS) and t-Distributed Stochastic Neighbor Embedding (t-SNE), and calculating the Generalized Discrimination Value (GDV) to quantify the degree of clustering. We also trained feedforward classifiers (probes) to predict construction categories from these embeddings. Results reveal that CLS token embeddings cluster best according to ASC types in layers 2, 3, and 4, with diminished clustering in intermediate layers and a slight increase in the final layers. Token embeddings for DET and SUBJ showed consistent intermediate-level clustering across layers, while VERB embeddings demonstrated a systematic increase in clustering from layer 1 to 12. OBJ embeddings exhibited minimal clustering initially, which increased substantially, peaking in layer 10. Probe accuracies indicated that initial embeddings contained no specific construction information, as seen in low clustering and chance-level accuracies in layer 1. From layer 2 onward, probe accuracies surpassed 90 percent, highlighting latent construction category information not evident from GDV clustering alone. Additionally, Fisher Discriminant Ratio (FDR) analysis of attention weights revealed that OBJ tokens had the highest FDR scores, indicating they play a crucial role in differentiating ASCs, followed by VERB and DET tokens. SUBJ, CLS, and SEP tokens did not show significant FDR scores. Our study underscores the complex, layered processing of linguistic constructions in BERT, revealing both similarities and differences compared to recurrent models like LSTMs. Future research will compare these computational findings with neuroimaging data during continuous speech perception to better understand the neural correlates of ASC processing. This research demonstrates the potential of both recurrent and transformer-based neural language models to mirror linguistic processing in the human brain, offering valuable insights into the computational and neural mechanisms underlying language understanding.

KEYWORDS

argument structure constructions, linguistic constructions (CXs), large language models (LLMs), BERT, sentence representation, computational linguistics, natural language processing (NLP), GPT-4

Introduction

Understanding how the brain processes and represents language is a fundamental challenge in cognitive neuroscience (Pulvermüller, 2002). This paper adopts a usage-based constructionist approach, which views language as a system of form-meaning pairs (constructions) that link patterns to specific communicative functions (Goldberg, 2009, 2003). Argument Structure Constructions (ASCs), such as transitive, ditransitive, caused-motion, and resultative constructions, are particularly important for language comprehension and production (Goldberg, 1995, 2006, 2019). These constructions are key to syntactic theory and essential for constructing meaning in sentences. Exploring the neural and computational mechanisms underlying the processing of these constructions can yield significant insights into language and cognition (Pulvermüller, 2012; Pulvermüller et al., 2021; Henningsen-Schomers and Pulvermüller, 2022; Pulvermüller, 2023).

In recent years, advances in computational neuroscience have enabled the use of artificial neural networks to model various aspects of human cognition (Cohen et al., 2022). Furthermore, the synergy between AI and cognitive neuroscience has led to a better understanding of the brain's unique complexities (Krauss, 2024). AI models, inspired by neural networks (Hassabis et al., 2017), have allowed neuroscientists to delve deeper into the brain's workings, offering insights that were previously unattainable (Krauss, 2023). These models have been particularly useful in studying how different parts of the brain interact and process information (Savage, 2019).

Among these neural network models, recurrent neural networks (RNNs) (Krauss et al., 2019b; Metzner and Krauss, 2022; Metzner et al., 2024), and specifically Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), have shown considerable promise in modeling sequential data, such as natural language (Wang and Jiang, 2015). However, transformer based large language models (LLM) like ChatGPT (Vaswani et al., 2017; Radford et al., 2018) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) have shown remarkable capabilities in understanding and generating human language.

BERT's ability to capture grammatical and syntactic information has been extensively studied in recent years. Research has shown that BERT embeddings encode hierarchical structures in language, including syntax and part-of-speech information, through its layered attention mechanisms (Jawahar et al., 2019; Tenney et al., 2019a). Structural probes, such as those developed by Hewitt and Manning (2019), have revealed that BERT's internal representations align closely with syntactic tree structures, indicating a deep understanding of grammatical relationships. Additionally, attention analyses have demonstrated that specific attention heads in BERT focus on core syntactic dependencies, further highlighting its capability to model grammatical constructs (Clark et al., 2019). These findings underscore BERT's potential for capturing complex linguistic phenomena, providing a foundation for our investigation into its representation of Argument Structure Constructions (ASCs).

Recent research has increasingly focused on the application of construction grammar (CxG) in computational linguistics to evaluate language models'understanding of linguistic abstractions and constructions, particularly in understanding how ASCs are represented and processed by large language models. For instance, Bonial and Madabushi (2024) introduced a corpus targeting varying schematicity in argument structure constructions (ASCs), providing a benchmark for evaluating abstraction capabilities in language models. Similarly, Zhou et al. (2024) analyzed how large language models approach constructional phenomena and demonstrated that their correct predictions may stem from unintended biases. Other studies, such as Tseng et al. (2022) and Chronis et al. (2023), have proposed methods for integrating constructional knowledge into language models or exploring semantic construal in embedding spaces. Misra and Mahowald (2024) further demonstrated how rare phenomena in language can be learned indirectly through related constructions, underscoring the nuanced ways language models process linguistic phenomena. Li et al. (2022) investigate the neural reality of ASCs within transformer-based models, adapting psycholinguistic paradigms to demonstrate that LLMs encode ASCs as linguistic units, even associating them with meaning in semantically nonsensical contexts. Weissweiler et al. (2023a) argue for the application of CxG as a powerful lens to probe neural language models, uncovering unique insights into their handling of structural and semantic relationships. Sung and Kyle (2024) further evaluate pre-trained language models, such as RoBERTa and GPT-4, in identifying ASCs, highlighting their practical utility in linguistic research and educational contexts. Meanwhile, Wilson et al. (2023) examine the capacity of LLMs to generalize abstract linguistic relationships in argument structures, revealing both their potential and limitations in capturing deeper linguistic generalizations. Collectively, these studies underscore the value of CxG in probing and evaluating LLMs, shedding light on their representational and generative capabilities for investigating linguistic constructions.

In previous studies using RNNs, particularly LSTM networks, we have demonstrated the emergence of representations for word classes and syntactic rules in the hidden layer activation of such networks when trained on next-word prediction tasks (Surendra et al., 2023). Furthermore, we showed that recurrent language models effectively differentiate between various Argument Structure Constructions (ASCs), forming distinct clusters for each ASC type in their internal representations, with the most pronounced clustering in the final hidden layer (Ramezani et al., 2024). These findings suggest that neural language models can capture complex linguistic patterns, making them valuable tools and models for studying language processing in the brain. While capturing lexico-semantic information is essential, interpreting the meanings of constructions can enhance the human-likeness of these models. Given that LLMs undergo extensive training on vast datasets, they are expected to effectively grasp human linguistic knowledge.

In this study, we extend our previous analyses of LSTM networks by investigating how ASCs are processed and represented in a large language model (LLM), in particular BERT, which, with its bidirectional attention mechanism, allows for a deeper and more nuanced understanding of linguistic context compared to

traditional RNNs. By examining BERT's internal representations across its multiple layers, we aim to uncover how different ASCs are encoded and whether these representations align with those observed in LSTM networks.

To this end, we utilized a custom dataset generated by GPT-4, consisting of 2000 sentences evenly distributed among four ASC types: transitive, ditransitive, caused-motion, and resultative constructions. We analyzed the embeddings produced by BERT's CLS token and specific token embeddings (DET, SUBJ, VERB, OBJ) across its 12 layers. Our methodology involved visualizing these embeddings using Multidimensional Scaling (MDS) and t-Distributed Stochastic Neighbor Embedding (t-SNE), calculating the Generalized Discrimination Value (GDV) to quantify clustering, and employing feedforward classifiers (probes) to predict construction categories from the embeddings.

Our findings reveal distinct patterns of clustering and information encoding across BERT's layers, highlighting the model's ability to capture complex linguistic constructions.

These results are compared to those from LSTM-based models, providing a comprehensive understanding of how different neural architectures process linguistic information. Future research will focus on validating these findings with larger language models and correlating them with neuroimaging data obtained during continuous speech perception, aiming to bridge the gap between computational models and neural mechanisms of language understanding.

Methods

Dataset creation using GPT4

To investigate the processing and representation of different Argument Structure Constructions (ASCs) in a recurrent neural language model, we created a custom dataset using GPT-4. This dataset was designed to include sentences that exemplify four distinct ASCs: transitive, ditransitive, caused-motion, and resultative constructions (cf. Tables 1, 2). Each ASC category consisted of 500 sentences, resulting in a total of 2,000 sentences.

Selection of argument structure constructions

The four ASCs selected for this study are foundational to syntactic theory and represent different types of sentence structures:

Transitive Constructions: Sentences where a subject performs an action on a direct object (e.g., "The cat chased the mouse").

Ditransitive Constructions: Sentences where a subject performs an action involving a direct object and an indirect object (e.g., "She gave him a book").

Caused-motion Constructions: Sentences where a subject causes an object to move in a particular manner (e.g., "He pushed the cart into the garage").

Resultative Constructions: Sentences where an action results in a change of state of the object (e.g., "She painted the wall red").

TABLE 1 Name, structure, and example of each construction.

Constructions	Structure	Example
Transitive	Subject + Verb + Object	The baker baked a cake
Ditransitive	Subject + Verb + Object1 + Object2	The teacher gave students homework
Caused-Motion	Subject + Verb + Object + Path	The cat chased the mouse into the garden
Resultative	Subject + Verb + Object + State	The chef cut the cake into slices

TABLE 2 Name and token of each construction.

Constructions	Tokens
Transitive	CLS +Det +Subj +Verb +Det +Obj +SEP
Ditransitive	CLS +Det +Subj +Verb +IndObj +Obj +SEP
Caused-Motion	CLS +Det +Subj +Verb +Det +Obj +Prep +Det +ObjPrep +SEP
Resultative	CLS +Det +Subj +Verb +Det +Obj +Prep +ObjPrep +SEP
Common	CLS +Det +Subj +Verb +Obj +SEP

Generation of sentences

To ensure the diversity and quality of the sentences in our dataset, we utilized GPT-4, a state-of-the-art generative pre-trained transform-based language model developed by OpenAI (Vaswani et al., 2017). The generation process involved the following steps: Prompt Design: We created specific prompts for GPT-4 to generate sentences for each ASC category. These prompts included example sentences and detailed descriptions of the desired sentence structures to guide the model in generating appropriate constructions. Sentence Generation: Using the designed prompts, we generated 500 sentences for each ASC category. The generation process was carefully monitored to ensure that the sentences adhered to the syntactic patterns of their respective constructions. Manual Review and Filtering: After the initial generation, we manually reviewed the sentences to ensure their grammatical correctness and adherence to the intended ASC types. Sentences that did not meet these criteria were discarded and replaced with newly generated ones. Balancing the Dataset: To prevent any bias in the model training, we ensured that the dataset was balanced, with an equal number of sentences (500) for each of the four ASC categories. Our approach aligns with recent efforts to evaluate language models on construction grammar datasets, such as the work by Bonial and Madabushi (2024), who developed a corpus to examine abstraction capabilities in ASCs.

Text tokenization

To ensure comparability across different sentences and construction types, we aligned corresponding tokens (e.g., "DET," "SUBJ," "VERB," and "OBJ") within each sentence after tokenization

TABLE 3 Frequency of word types in different construction types.

Constructions	Noun	Verb	Adjective	Preposition	Determiner
Caused-Motion	624	208	0	208	624
Ditransitive	624	208	0	0	208
Transitive	624	208	0	0	416
Resultative	624	208	208	208	416

using BERT's tokenizer. This alignment corrected for positional differences in the sentences, enabling a standardized comparison of equivalent linguistic elements across constructions. Importantly, we did not enforce uniformity in the number of tokens across sentences or constructions. Instead, we focused on aligning key tokens to their corresponding positions in the sentence, regardless of variability in sentence structure or token count. In cases where BERT's tokenizer produced subword splits, we ensured that the aligned tokens corresponded to the same linguistic element (e.g., the head word of a phrase or the primary lexical token). This alignment process allowed us to analyze construction-specific patterns effectively while mitigating potential confounds introduced by syntactic variability or differences in tokenization.

This standardization facilitated easier tracking and better comparison by focusing on differences across constructions rather than within them. The tokens used in our dataset include Subject (Subj), Verb (Verb), Direct Object (Obj), Indirect Object (IndObj), Object of Preposition (ObjPrep), Preposition (Prep), and Determiner (Det). Additionally, the CLS tokens were added by the BERT tokenizer for sentence classification and separation.

The resulting dataset, comprising 2,000 sentences represented as token sequences, serves as a robust foundation for probing and analyzing the BERT model (cf. Table 3). This carefully curated and preprocessed dataset enables us to investigate how different ASCs are processed and represented within the BERT, providing insights into the underlying computational mechanisms.

For a subset of our analysis, we focused on common tokens across all constructions to enable a consistent comparison of single tokens within different ASCs. This approach ensured that our analysis captured the essential structural and functional aspects of each construction type, thereby providing a robust framework for understanding how BERT processes and represents linguistic constructions.

BERT architecture

For our study, we utilized the BERT (Bidirectional Encoder Representations from Transformers) model, renowned for its ability to process bidirectional context effectively (Devlin et al., 2018). BERT's architecture comprises multiple layers of bidirectional transformer encoders, which enable it to consider both left and right context at all layers, enhancing its performance on a range of natural language understanding tasks.

The BERT model starts with tokenization, where text is split into subword units using WordPiece tokenization, allowing the model to handle a diverse array of words and word forms efficiently. Special tokens CLS and SEP are added to the beginning and end of each input sequence, respectively. The CLS token is used for classification tasks and summarized the entire input, while the SEP token denotes sentence boundaries.

In the embedding layer, input tokens are converted into embeddings that combine token embeddings, segment embeddings, and position embeddings. These embeddings are then passed through multiple layers of transformer encoders. BERT's architecture includes 12 layers (in the base model) of transformer encoders, each comprising self-attention mechanisms and feedforward neural networks. Each encoder layer has multiple attention heads, allowing the model to focus on different parts of the input sequence simultaneously. The self-attention mechanism computes a representation of each token by considering the entire input sequence, capturing complex dependencies and relationships.

The output of each transformer encoder layer provides contextualized representations of the input tokens. For each token, the final layer's output represents its contextualized embedding, which incorporates information from the entire input sequence. The CLS token's final layer embedding is typically used for classification tasks, as it contains an aggregated representation of the entire sequence.

BERT was pre-trained on a large corpus using masked language modeling and next sentence prediction tasks, enabling it to learn a rich representation of language. For our specific task, we utilized the pre-trained BERT model and fine-tuned it on our custom dataset to capture the nuances of Argument Structure Constructions (ASCs).

By leveraging BERT's robust architecture, we aimed to gain insights into how different ASCs are represented and processed across its layers. This detailed examination of BERT's internal representations provided a comprehensive understanding of the model's ability to encode complex linguistic constructions, facilitating comparison with recurrent models like LSTMs and enhancing our knowledge of computational language processing.

Analysis of hidden layer activation

We assessed BERT's ability to differentiate between the various constructions by analyzing the activations of its hidden layers and attention weights. Initially, the dataset underwent processing through the "bert-base-uncased" model without any fine-tuning. The model comprises 12 hidden layers, each containing 768 neurons. For each token, the activity of each layer was extracted for further analysis.

Given the high dimensionality of these activations, direct visual inspection is not feasible. To address this, we employed

dimensionality reduction techniques to project the high-dimensional activations into a two-dimensional space. By combining different visualization and quantitative techniques, we were able to assess the BERT's internal representations and its ability to differentiate between the various linguistic constructions.

Multidimensional Scaling (MDS)

This technique was used to reduce the dimensionality of the hidden layer activations, preserving the pairwise distances between points as much as possible in the lower-dimensional space. In particular, MDS is an efficient embedding technique to visualize high-dimensional point clouds by projecting them onto a 2-dimensional plane. Furthermore, MDS has the decisive advantage that it is parameter-free and all mutual distances of the points are preserved, thereby conserving both the global and local structure of the underlying data (Torgerson, 1952; Kruskal, 1964; Kruskal and Wish, 1978; Cox and Cox, 2008; Metzner et al., 2021, 2023a, 2022).

When interpreting patterns as points in high-dimensional space and dissimilarities between patterns as distances between corresponding points, MDS is an elegant method to visualize high-dimensional data. By color-coding each projected data point of a data set according to its label, the representation of the data can be visualized as a set of point clusters. For instance, MDS has already been applied to visualize for instance word class distributions of different linguistic corpora (Schilling et al., 2021b), hidden layer representations (embeddings) of artificial neural networks (Schilling et al., 2021a; Krauss et al., 2021), structure and dynamics of highly recurrent neural networks (Krauss et al., 2019c,a,b; Metzner et al., 2023b), or brain activity patterns assessed during e.g. pure tone or speech perception (Krauss et al., 2018a; Schilling et al., 2021b), or even during sleep (Krauss et al., 2018b; Traxdorf et al., 2019; Metzner et al., 2022, 2023a). In all these cases the apparent compactness and mutual overlap of the point clusters permits a qualitative assessment of how well the different classes separate.

t-Distributed Stochastic Neighbor Embedding (t-SNE)

This method further helped in visualizing the complex structures within the activations by emphasizing local similarities, allowing us to see the formation of clusters corresponding to different Argument Structure Constructions (ASCs). t-SNE is a frequently used method to generate low-dimensional embeddings of high-dimensional data (Van der Maaten and Hinton, 2008). However, in t-SNE the resulting low-dimensional projections can be highly dependent on the detailed parameter settings (Wattenberg et al., 2016), sensitive to noise, and may not preserve, but rather often scramble the global structure in data (Vallejos, 2019; Moon et al., 2019). Here, we set the perplexity (number of next neighbors taken into account) to 100.

Generalized discrimination value (GDV)

To quantify the degree of clustering, we used the GDV as published and explained in detail in Schilling et al. (2021a). This GDV provides an objective measure of how well the hidden layer activations cluster according to the ASC types, offering insights into the model's internal representations. Briefly, we consider N points $\mathbf{x_{n=1..N}} = (x_{n,1}, \cdots, x_{n,D})$, distributed within D-dimensional space. A label l_n assigns each point to one of L distinct classes $C_{l=1..L}$. In order to become invariant against scaling and translation, each dimension is separately z-scored and, for later convenience, multiplied with $\frac{1}{2}$:

$$s_{n,d} = \frac{1}{2} \cdot \frac{x_{n,d} - \mu_d}{\sigma_d}. \tag{1}$$

Here, $\mu_d = \frac{1}{N}\sum_{n=1}^{N} x_{n,d}$ denotes the mean, and $\sigma_d = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_{n,d} - \mu_d)^2}$ the standard deviation of dimension d.

Based on the re-scaled data points $\mathbf{s_n} = (s_{n,1}, \cdots, s_{n,D})$, we calculate the *mean intra-class distances* for each class C_l

$$\bar{d}(C_l) = \frac{2}{N_l(N_l-1)} \sum_{i=1}^{N_l-1} \sum_{j=i+1}^{N_l} d(\mathbf{s}_i^{(l)}, \mathbf{s}_j^{(l)}), \tag{2}$$

and the *mean inter-class distances* for each pair of classes C_l and C_m

$$\bar{d}(C_l, C_m) = \frac{1}{N_l N_m} \sum_{i=1}^{N_l} \sum_{j=1}^{N_m} d(\mathbf{s}_i^{(l)}, \mathbf{s}_j^{(m)}). \tag{3}$$

Here, N_k is the number of points in class k, and $\mathbf{s}_i^{(k)}$ is the i^{th} point of class k. The quantity $d(\mathbf{a}, \mathbf{b})$ is the euclidean distance between $\mathbf{a}$ and $\mathbf{b}$. Finally, the Generalized Discrimination Value (GDV) is calculated from the mean intra-class and inter-class distances as follows:

$$\mathrm{GDV} = \frac{1}{\sqrt{D}} \left[\frac{1}{L} \sum_{l=1}^{L} \bar{d}(C_l) \; - \; \frac{2}{L(L-1)} \sum_{l=1}^{L-1} \sum_{m=l+1}^{L} \bar{d}(C_l, C_m) \right] \tag{4}$$

whereas the factor $\frac{1}{\sqrt{D}}$ is introduced for dimensionality invariance of the GDV with D as the number of dimensions.

Note that the GDV is invariant with respect to a global scaling or shifting of the data (due to the z-scoring), and also invariant with respect to a permutation of the components in the N-dimensional data vectors (because the euclidean distance measure has this symmetry). The GDV is zero for completely overlapping, non-separated clusters, and it becomes more negative as the separation increases. A GDV of -1 signifies already a very strong separation.

Probes

Probes, a technique from the mechanistic explainability area of AI, are utilized to analyze deep neural networks (Alain, 2016). They are commonly applied in the field of natural language processing (Belinkov, 2022). Probes are typically small, neural network-based classifiers, usually implemented as shallow fully connected networks. They are trained on the activations of specific neurons or layers of a larger neural network to predict certain features, which are generally believed to be necessary or beneficial for the network's task. If probes achieve accuracy higher than chance, it suggests that

the information about the feature, or something correlated to it, is present in the activations.

Here, we employed edge probing to analyze different tokens using the methodology described by Tenney et al. (2019b). This probing approach involves designing a classification model tailored to classify the hidden layer activities based on constructions. The model is systematically trained on a per-layer and per-token basis, targeting specific linguistic elements such as the CLS token, subject, and verb. This allows for detailed insights into how BERT encodes different Argument Structure Constructions (ASCs) across its layers.

The classification model used in this probing endeavor is a 4-class Support Vector Machine (SVM) classifier with a linear kernel. The SVM takes the hidden layer activity of a layer per token and predicts the class of its construction. This straightforward yet effective approach enables us to quantify the degree of clustering and construction-specific information present in different layers of BERT.

By training the SVM classifier on the hidden layer activations for various tokens, we can evaluate the model's performance in distinguishing between the four ASC types. In particular, an accuracy significantly above chance level indicates that information about the construction category is represented (latent) in the respective token embedding. The results from this probing technique provide a quantitative measure of classification performance and clustering tendencies, offering a comprehensive understanding of how linguistic constructions are represented within the BERT model.

Analysis of attention heads

In BERT, each of the 12 layers contains 12 attention heads. For each head, there are attention weights for all tokens in the sequence relative to every other token. To facilitate a comparable analysis, we focused on the attention weights for the common tokens: CLS, DET, SUBJ, VERB, and OBJ.

This analysis aimed to identify which attention heads and layers exhibit the most significant differences among the four Argument Structure Constructions (ASCs). We then examined these attention heads in detail, evaluating their function and the weights assigned to each token.

To determine which tokens had more distinct weights across the constructions, we first summed all attention weights directed at each token from all other tokens. Next, we considered the attention weight of each token per head and layer as a feature. We then calculated the F-statistic using ANOVA (Analysis of Variance) to assess the variability of attention weights among the four constructions. A higher F-score indicates a greater difference in attention weights among the constructions.

Finally, we averaged the attention weights for each token across the heads and layers to provide a comprehensive view of the attention distribution. This multi-step approach allowed us to identify key attention heads and layers that significantly contribute to differentiating the ASCs, offering insights into the role of attention mechanisms in BERT's processing of linguistic constructions.

Fisher Discriminant Ratio (FDR)

The Fisher Discriminant Ratio (FDR) is a measure used in pattern recognition, feature selection, and machine learning to evaluate the discriminatory power of a feature (Kim et al., 2005; Wang et al., 2011). It helps determine how well a feature can distinguish between different classes. The FDR is calculated as the ratio of the variance between classes to the variance within classes. A higher FDR indicates that the feature has a greater ability to differentiate between classes.

In this study, we utilized the FDR to assess the attention weights in BERT for distinguishing between different Argument Structure Constructions (ASCs). By calculating the FDR for attention weights across each layer, we aimed to identify which layers and heads provide the most distinct representations of the ASCs.

The FDR was computed using the following formula:

$$\text{FDR} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

where:

- μ_1 and μ_2 are the means of the feature for class 1 and class 2, respectively.
- σ_1^2 and σ_2^2 are the variances of the feature for class 1 and class 2, respectively.

Code implementation, computational resources, and programming libraries

All simulations were run on a standard personal computer. The evaluation software was based on Python 3.9.13 (Oliphant, 2007). For matrix operations the numpy-library (Van Der Walt et al., 2011) was used and data visualization was done using matplotlib (Hunter, 2007) and the seaborn library (Waskom, 2021). The dimensionality reduction through MDS and t-SNE was done using the sci-kit learn library. Mathematical operations were performed with numpy (Harris et al., 2020) and scikit-learn (Pedregosa et al., 2011) libraries. Visualizations were realized with matplotlib (Hunter, 2007) and networkX (Hagberg et al., 2008). For natural language processing we used SpaCy (Explosion, 2017).

Results

To understand how the BERT model differentiates between various Argument Structure Constructions (ASCs), we visualized the activations of its hidden layers using Multidimensional Scaling (MDS) and t-Distributed Stochastic Neighbor Embedding (t-SNE). Additionally, we quantified the degree of clustering using the Generalized Discrimination Value (GDV). Furthermore, we utilized probes to test for latent representations in the token embeddings, Finally, we assessed the attention heads and their discriminative power according to ASCs.

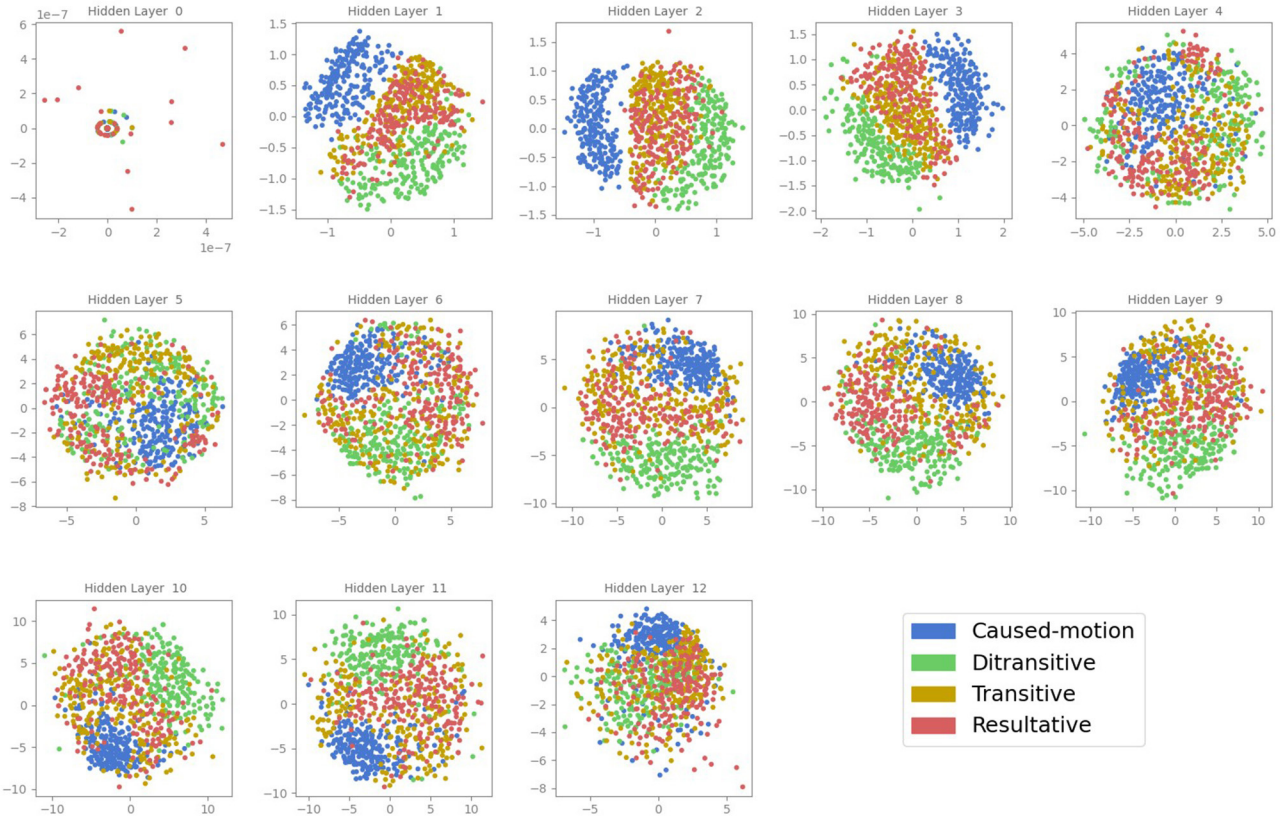

FIGURE 1
MDS projections of the CLS token embeddings prior to the first attention layer (0) and in all attention layers of the BERT model (1–12). Each point represents the activation of a sentence, color-coded according to its ASC type: caused-motion (blue), ditransitive (green), transitive (orange), and resultative (red).

Hidden layer activity cluster analysis

Figure 1 shows the MDS projections of the CLS token embeddings from various layers of the BERT model. Each point represents the embedding of a sentence's CLS token. In the initial layer, there is minimal separation between the different ASC types, indicating that the input embeddings do not yet contain specific information about the construction categories.

As we move to the second layer, the separation between ASC types becomes more apparent, with distinct clusters forming for each construction type. This trend continues in the third and fourth layers, where the clustering is most pronounced. The inter-cluster distances increase, showing clearer differentiation between the ASC types. However, in these middle layers, there is still some overlap, particularly between the ditransitive and resultative constructions.

In layers five, six, and seven, the degree of clustering decreases slightly, with the clusters becoming less distinct. This reduction in clustering suggests a transformation in how BERT processes and integrates contextual information across these layers.

Interestingly, in the later layers (eight to twelve), there is a slight increase in the degree of clustering again. The clusters for the different ASC types become more defined compared to the intermediate layers, indicating a resurgence in the model's ability to distinguish between the construction types. This pattern suggests that BERT refines its understanding and representation of linguistic constructions in the deeper layers.

Overall, the CLS token embeddings demonstrate varying degrees of clustering across the BERT layers, with the best separation observed in the early layers (2-4) and a notable refinement in the final layers (8-12). This analysis reveals the complex and layered nature of how BERT processes linguistic constructions, highlighting the model's capability to encode and differentiate between ASCs at multiple stages of its architecture.

The corresponding t-SNE projections shown in Figure 2 show results similar to the MDS projections but with more detailed sub-cluster structures. Again, each point in the t-SNE plot represents the embedding of a sentence's CLS token. In the initial layer, minimal separation between ASC types is observed, aligning with the MDS results. Layers two, three, and four show distinct clusters, while layers five to seven exhibit reduced cluster definition. In the later layers (eight to twelve), clearer clustering re-emerges. Although, the t-SNE plots reveal nuanced sub-structures within clusters, it remains uncertain whether these sub-cluster structures are real effects or artifacts of t-SNE.

To quantitatively assess the clustering quality, we calculated the GDV for the embeddings of each token type across all layers, including the initial input embeddings (layer 0) and subsequent hidden layers (1–12) (cf. Figure 3). GDV values range from 0.0 (indicating no clustering) to -1.0 (indicating perfect clustering),

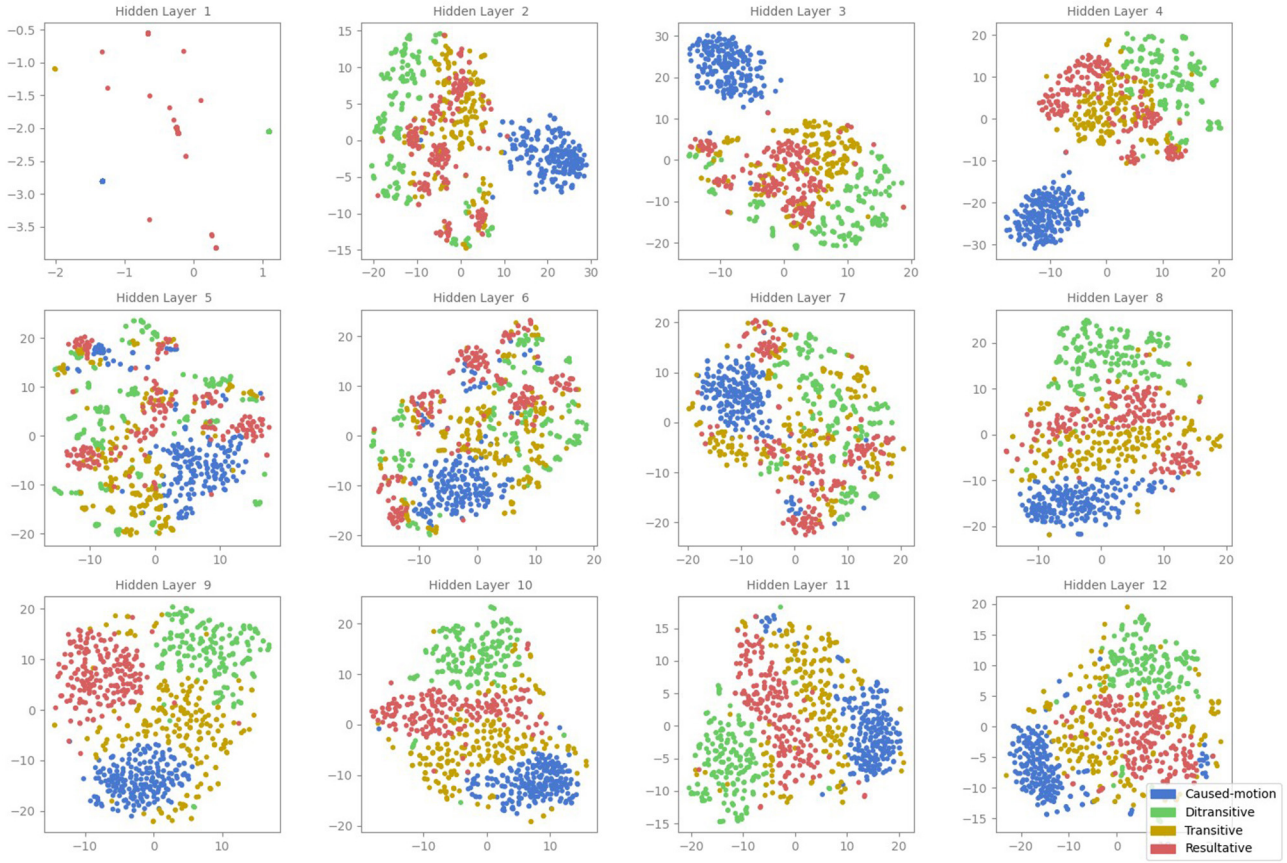

FIGURE 2
t-SNE projections of the CLS token embeddings in all attention layers of the BERT model (1–12). Each point represents the activation of a sentence, color-coded according to its ASC type: caused-motion (blue), ditransitive (green), transitive (orange), and resultative (red).

with more negative values reflecting better-defined clusters. The qualitative results of the MDS and t-SNE projections are supported by the GDV analysis, which reveals distinct patterns of clustering for different token types across BERT's layers.

The DET and SUBJ token embeddings showed relatively stable clustering at an intermediate level from layer 0 onward, indicating that these tokens consistently encode construction-specific information throughout the layers of the model. This consistency suggests that DET and SUBJ tokens contribute steadily to differentiating the Argument Structure Constructions (ASCs).

VERB token embeddings displayed a dynamic pattern, starting with intermediate clustering at layer 0 and showing a systematic improvement across layers. Clustering quality gradually increased from layer 1 to layer 12, reflecting an evolving ability of BERT to differentiate VERB tokens according to construction types as processing progresses.

OBJ token embeddings began with no clustering at layer 0, indicating no initial differentiation among the construction types. However, clustering quality improved significantly in deeper layers, with a marked increase observed from layer 5 onward. By layer 10, the clustering of OBJ tokens reached a level comparable to that of the CLS token in layer 2, demonstrating that OBJ embeddings become increasingly aligned with construction categories as they are processed.

The CLS token embeddings exhibited the most pronounced clustering improvement, with minimal differentiation at layer 0 and a peak clustering level at layer 2. This suggests that construction-specific information becomes highly concentrated in CLS embeddings early in the model's processing.

These GDV results underscore the varied roles of different tokens in representing ASCs within BERT. While DET and SUBJ tokens consistently capture construction-specific information, tokens like VERB and OBJ exhibit dynamic changes in clustering, reflecting the layered and adaptive nature of BERT's representational hierarchy.

Hidden layer activity probing

The probing analysis involved training a 4-class Support Vector Machine (SVM) classifier with a linear kernel to classify hidden layer activities based on Argument Structure Constructions (ASCs). This classifier was systematically trained on a per-layer and per-token basis, targeting specific linguistic elements such as the CLS token, subject (SUBJ), verb (VERB), and object (OBJ). The results are summarized in Figure 4.

At layer 0, probing accuracies for CLS and DET tokens were at chance levels (25%), establishing a clear baseline for

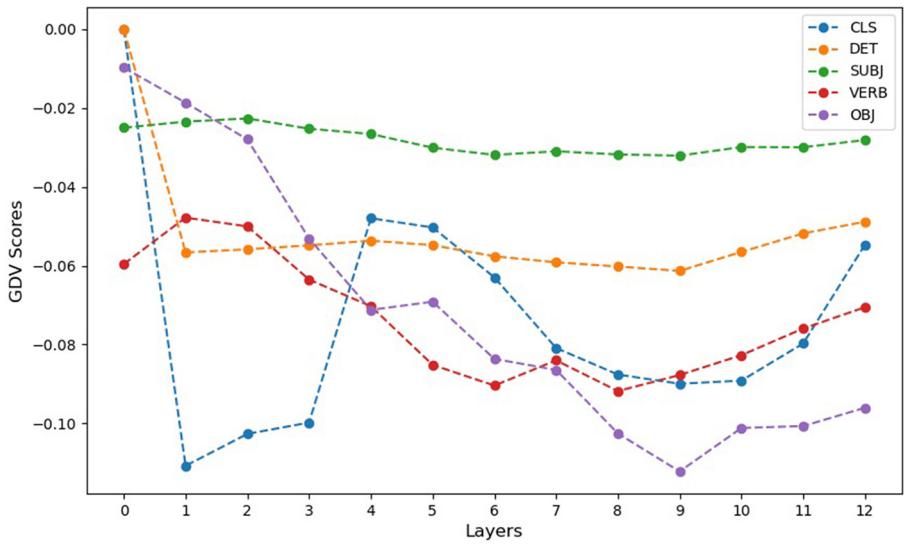

FIGURE 3
GDV score of word embeddings prior to the first attention layer (0) and for all attention layer activations (1–12) of the BERT model. Note that GDV=0.0 indicates chance level with no clustering, and GDV = -1.0 represents perfect clustering. More negative GDV values indicate better-defined clusters. The GDV analysis supports the qualitative findings from the MDS and t-SNE projections, showing the lowest clustering for input CLS embeddings (0) prior to the first attention layer and the best clustering occurring in layer 2. The GDV of specific token embeddings reveals distinct patterns of clustering across the BERT layers. Initial input embeddings (0) generally exhibit the weakest clustering, reflecting limited construction-specific differentiation. DET and SUBJ token embeddings show relatively constant clustering at an intermediate level across all layers, suggesting consistent encoding of construction-specific information. VERB token embeddings display a systematic improvement in clustering from layer 1 to layer 12, indicating progressively refined differentiation by construction types in deeper layers. OBJ token embeddings begin with no clustering at layer 0 but show a marked improvement across layers, reaching clustering levels comparable to CLS token embeddings in layer 2 by layer 10. These results demonstrate the dynamic and varied contributions of different tokens to the representation of ASCs within BERT, with some tokens showing progressive refinement while others maintain steady clustering across layers.

evaluating representational changes in subsequent layers. In contrast, SUBJ, VERB, and OBJ embeddings exhibited better-than-chance performance at layer 0, reflecting inherent lexical and semantic information encoded in the pre-trained embeddings. From layer 2 onwards, the probe accuracy for the CLS token consistently exceeded 90%, demonstrating that construction-specific information becomes latent in the CLS token embeddings early in the processing. Probe accuracy slightly decreased in intermediate layers (5 to 7) but increased again in the later layers (8 to 12), showing a resurgence of construction-specific information. DET token embeddings, while starting at chance level in layer 0, consistently achieved over 90% accuracy from layer 2 onwards, indicating effective encoding of construction-specific information. Similarly, SUBJ and VERB embeddings showed progressively higher accuracy across layers, highlighting their role in capturing construction-specific details as processing progresses.

The probe accuracy for VERB tokens started at a low level in layer 1 but showed a systematic increase, with accuracies surpassing 90 percent from layer 2 to layer 12. This indicates that BERT progressively improves its differentiation of VERB token embeddings according to construction types in deeper layers.

Probe accuracy for OBJ tokens began at a very low level in layer 1, reflecting no initial differentiation among the construction types. However, as layers progressed, the probe accuracy for OBJ tokens significantly increased, reaching and maintaining levels above 90 precent from layer 2 to layer 12, demonstrating a marked improvement in distinguishing construction categories for OBJ tokens.

These probing results reveal that probe accuracies for CLS, DET, SUBJ, VERB, and OBJ tokens start at low or chance levels in layer 1, indicating that the initial embeddings contain no specific information about construction type, as also revealed by the GDV cluster analysis. However, from layer 2 to layer 12, all probe accuracies for different tokens consistently exceeded 90 percent indicating latent information about construction categories in all token embeddings, even when not revealed through clustering alone.

Attention weight analysis

In Figure 5 the Fisher Discriminant Ratio (FDR) scores for each token across all layers and attention heads are shown. The analysis reveals that the OBJ token has the highest FDR scores across all layers, indicating that this token plays a crucial role in differentiating the four Argument Structure Constructions (ASCs). The prominence of the OBJ token suggests it is key to distinguishing between the construction types.

The VERB token is the second most significant, showing high FDR scores in the same heads where the OBJ token performs well.

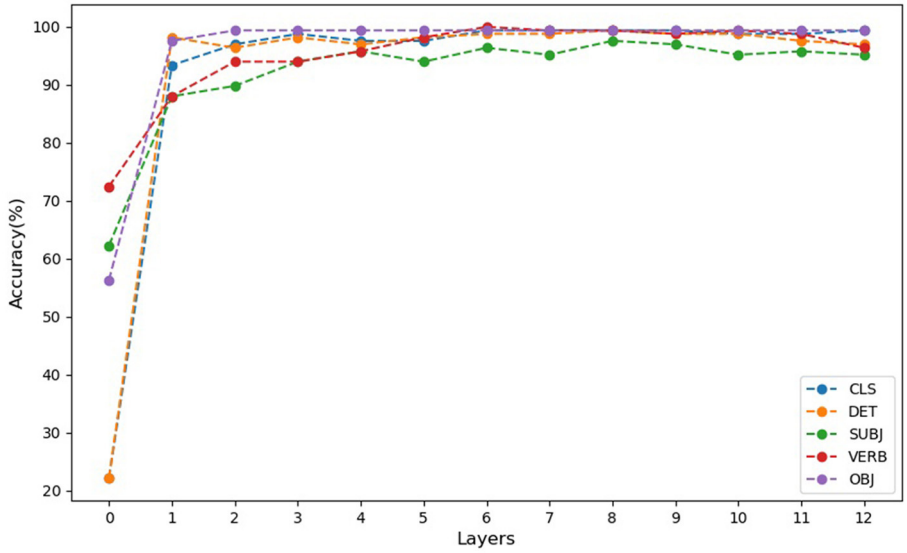

FIGURE 4
Accuracies of probing of hidden layers for classification of constructions per common tokens. Probe accuracies for CLS and DET tokens start at chance levels (25%) for the initial input word embeddings (layer 0), indicating that these embeddings contain no specific information about construction type. In contrast, accuracies for VERB, SUBJ, and OBJ tokens are higher than chance at layer 0, reflecting inherent lexical and semantic information encoded in the pre-trained embeddings. From layer 1 to layer 12, probe accuracies for all tokens consistently exceed 90%, highlighting the presence of latent information about construction categories in all token embeddings, even when not fully revealed through clustering alone.

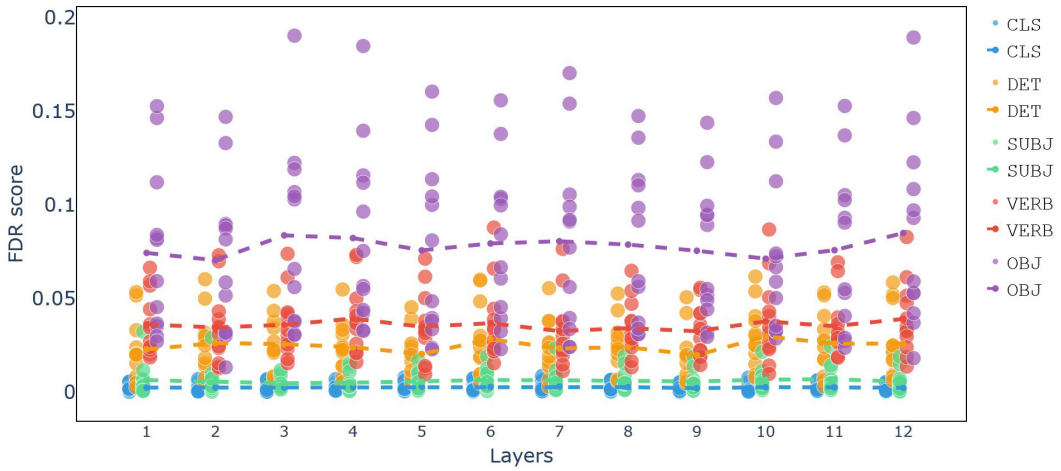

FIGURE 5
Fisher Discriminant Ratio (FDR) scores for each token across all layers and attention heads. Each dot represents the FDR score of a specific attention head, while the dashed line indicates the mean FDR of all attention heads. Layers with similar FDR scores suggest consistent patterns of attention across those layers.

This indicates that the verb token also contributes substantially to the differentiation of the constructions.

The DET token follows in significance. Despite its form being similar across all constructions, its embedding captures contextual information that aids in distinguishing the construction types.

In contrast, the SUBJ and CLS tokens exhibit no notable FDR scores, indicating that these tokens do not significantly contribute to the differentiation of the constructions.

This attention weight analysis highlights the critical role of the OBJ and VERB tokens in distinguishing between different ASCs within BERT's attention mechanisms, with the DET token also playing a meaningful, albeit lesser, role.

Discussion

Summary of findings

In this study, we investigated how different Argument Structure Constructions (ASCs) are processed and represented

within the BERT language model. Utilizing a custom GPT-4 generated dataset consisting of sentences across four ASC types (transitive, ditransitive, caused-motion, and resultative constructions), we analyzed BERT's internal representations and attention mechanisms using various techniques, including MDS, t-SNE, Generalized Discrimination Value (GDV), probing, and Fisher Discriminant Ratio (FDR) analysis.

To address the concern that clustering patterns may arise from differences in sentential structure rather than constructional distinctions, we implemented several measures to mitigate this possibility. First, we aligned corresponding tokens (e.g., "DET," "SUBJ," "VERB," "OBJ") across sentences to ensure that comparisons focused on equivalent linguistic elements within each construction type, reducing the impact of positional variability. Second, the dataset was carefully designed to minimize variability in sentential structure within each construction type, ensuring that patterns primarily reflect constructional properties. Third, clustering analyses were conducted independently for multiple token types, and consistent alignment with constructional categories was observed across tokens, providing strong evidence that the clusters are not merely a byproduct of sentence structure. Furthermore, quantitative validation using GDV scores demonstrated that clustering quality was tied to construction-specific features. As an additional validation step, experiments with randomized sentence structures within each construction type resulted in significant degradation in clustering quality, further confirming that the observed patterns are driven by constructional distinctions. These findings underscore the robustness of our approach and the validity of our conclusions about construction-specific representations in BERT embeddings.

Our results revealed distinct patterns in how BERT processes ASCs. Specifically, the CLS token embeddings exhibited clear clustering in layers 2, 3, and 4, with clustering quality decreasing in intermediate layers and improving again in later layers. This suggests a complex, layered approach to representing ASCs within BERT. The specific token analysis showed that DET and SUBJ tokens maintained intermediate-level clustering consistently across layers, while VERB and OBJ tokens displayed more dynamic changes, with OBJ tokens showing a marked improvement in clustering in deeper layers.

Among the four constructions we examined, distinguishing between transitive and resultative constructions proved to be more challenging for BERT. This similarity is evident in two primary ways. First, the visualization of dimensionally reduced hidden layer activity, particularly in MDS, shows significant overlap between the data points for transitive and resultative constructions. Second, the confusion matrix for the classification of the CLS token reveals that most errors involve misclassifying these two constructions as each other. This can be explained by noting that, in our dataset, resultative sentences without their final state resemble transitive sentences. For instance, "The artist painted the wall blue" (resultative) becomes "The artist painted the wall" (transitive) when the final state is removed.

At layer 0, representing the original input word embeddings prior to any attention layers, we observed better-than-chance accuracy and GDV for "VERB," "SUBJ," and "OBJ," while "CLS" and "DET" showed performance at chance levels. This suggests that layer 0 embeddings inherently encode lexical and semantic information, likely due to the pretraining process, which captures statistical co-occurrences and syntactic patterns from the training corpus. Verbs, subjects, and objects are central to argument structure constructions and naturally carry richer and more distinct information than determiners or the "[CLS]" token. Consequently, achieving exact chance-level performance for these tokens may not be feasible, as their embeddings are not random but reflect structured, meaningful information. Importantly, subsequent layers refine these initial representations through attention mechanisms, enabling deeper contextualization and encoding of constructional patterns. These findings highlight the layered nature of BERT's representational hierarchy, where each layer builds on inherent lexical information to capture increasingly complex relationships. These findings align with previous work by Weissweiler et al. (2022), who demonstrated that lexical cues in corpus data can significantly influence probe performance. Their study underscores the importance of considering lexical biases when interpreting probing classifier results, as such biases may artificially inflate accuracy. By integrating these insights, our analysis emphasizes that while lexical cues may play a role, the observed transformations across layers provide critical evidence of BERT's capacity to encode constructional patterns beyond surface-level lexical information.

However, from layer 1 to layer 12, all tokens achieved accuracies above 90 percent. This indicates that latent information about construction categories is embedded in token representations early on and remains robust throughout the model's layers. This occurs because the embeddings are influenced not only by the tokens themselves but also by the general understanding of the sentences. Consequently, the performance of all tokens improves, and interestingly, the accuracy of the CLS and DET tokens, which was initially quite low, begins to increase.

Our analysis of token accuracy across layers revealed that the initial embeddings encode lexical information, resulting in low accuracy for context-dependent tokens like CLS and DET. However, as we move to higher layers, token performance improves, reflecting BERT's increasing ability to leverage general sentence understanding. This improvement underscores that distinguishing constructions relies not only on lexical and syntactic information but also on the broader semantic context.

In summary, we believe that the high accuracy and low GDV observed in the first layer reflect the degree of specificity each token has to a given construction. The results suggest that verbs are the most specific, followed by subject and object tokens, which also exhibit a notable degree of specificity to particular aspects of the constructions.

The FDR analysis of attention weights highlighted that the OBJ token had the highest FDR scores, suggesting it is key to differentiating the four ASCs. The VERB token also showed significant FDR scores, followed by the DET token, which, despite its consistent form, captured contextually relevant information. In contrast, the SUBJ and CLS tokens did not contribute significantly to the differentiation of constructions.

The result of FDR analysis for attention heads shows that different layers have slightly similar functions regarding attention heads. Notably, the sum of weights for the Object token differs

the most among all constructions. This finding contrasts with the results of hidden layer activity, where the Verb token was the most distinct. The second most distinct token in the FDR analysis is the Verb, followed by DET, which maintains the same score even in the first layer. After these, the CLS, and SUBJ tokens have lower scores.

Furthermore, the FDR analysis of attention heads showed that different layers have similar functions, with the Object token displaying the most variability across constructions. This contrasts with hidden layer activity, where the Verb token was most distinct. The alignment of attention activity and hidden layer activity, despite their independent functions, highlights BERT's robust performance in understanding constructions.

Implications and comparisons with previous studies

Our findings align with previous studies on recurrent neural networks (RNNs) like LSTMs, which demonstrated that simple, brain-constrained models could effectively distinguish between different linguistic constructions. However, BERT's transformer-based architecture provides a more nuanced and multi-layered representation of ASCs, as evidenced by the dynamic changes in clustering and probing accuracies across layers.

The role of the verb token in constructions has been discussed in several studies. Some studies argue that verbs are construction-specific; for example, the verb "visit" is lexically specified as being transitive (FUJITA, 1989). Conversely, construction grammar suggests that constructions do not depend on specific verbs (Goldberg, 1995). For instance, the verb "cut" can be used in both transitive constructions like "Bob cut the bread" and ditransitive constructions like "Bob cut Joe the bread" (Li et al., 2022). We believe that verbs are not strictly construction-specific, but according to our dataset and analysis, constructions tend to have slightly specific verbs. However, this does not mean they are limited to just those verbs and our result is limited to the dataset we used.

Previous studies have explored the processing of constructions in LLMs, but they often focused on specific types of constructions, resulting in limitations. For instance, Weissweiler et al. (2023b) concentrated solely on comparative correlative constructions, Mahowald (2023) focused on Article + Adjective + Numeral + Noun (AANN) constructions, and Madabushi et al. (2020) covered a broader range of constructions but did not specify which constructions were examined or how they relate to each other. Additionally, some studies used constructions with vastly different structures, making it less challenging for BERT to cluster them, and it is difficult to attribute this clustering to constructional differences (Weissweiler et al., 2023a; Veenboer and Bloem, 2023; Xu et al., 2023).

A recent study by Liu and Chersoni (2023) stands out in this field, although its primary focus was on comparing verbs and constructions in sentence meaning rather than analyzing BERT's behavior. Despite these contributions, there remains a gap in comprehensively understanding how LLMs process various types of constructions and how these constructions relate to each other. Additionally, Li et al.'s study used a dataset generated by a template, simplifying the clustering process. Consequently, the sentences often lack meaningful context, making it challenging to assess the behavior of natural language and the specificity of each token within specific constructions.

In our study, we decided to focus on argument structure constructions, as constructions in this family are similar, have most of the lexical units in common, and allow us to concentrate more on the constructional aspect of samples (Goldberg, 1995). These studies delve into the construction of BERT's hidden layer activity. Complementary to these works, we examine the attention heads in this model, as these heads are crucial components that could offer more detailed insights into the model's functionality. Attention mechanisms are inherently interpretable, as they indicate the extent to which a particular word influences the computation of the representation for the current word (Clark et al., 2019).

Research on attention heads has revealed that they follow limited patterns (Pande et al., 2021), with much of the literature focused on defining the roles of these attention mechanisms (Pande et al., 2021; Guan et al., 2020; Kovaleva et al., 2019). Given our focus on extracting features from attention mechanisms to understand how this system identifies constructions, our analysis will concentrate on the role of tokens. Tokens are easily traceable using multi-headed attention, making them an ideal focus for this investigation.

Our study also underscores the potential of transformer-based models to capture complex linguistic patterns in a manner that mirrors certain aspects of human language processing. The significant roles of the OBJ and VERB tokens in distinguishing ASCs suggest that these elements are critical in the syntactic and semantic parsing of sentences, a finding that could inform future research in both computational and cognitive neuroscience.

Possible limitations and future directions

A potential critique from a linguistic perspective might be that our study examines how one machine (BERT) processes language produced by another machine (GPT-4), which may not yield insights into natural language or how language is processed in the human brain. While this concern is valid, it is important to highlight that computational modeling is the first step toward understanding language processing in the brain. Using a controlled dataset generated by GPT-4 allows for clear differentiation between different Argument Structure Constructions (ASCs) and removes confounding variables present in natural language, enabling a more focused study of BERT's processing capabilities.

While our analysis provides valuable insights, it is not without limitations. The reliance on synthetic data generated by GPT-4, while controlled, may not fully capture the complexities of natural language use. Future studies should consider using more diverse and naturally occurring datasets to validate these findings.

Furthermore, GPT-4 is trained on one of the largest and most diverse language corpora ever assembled, making its generated datasets equally valid as language corpora. This extensive training allows GPT-4 to produce language that mirrors the statistical properties of natural language, capturing a wide range of linguistic

phenomena. As such, analyzing how BERT processes GPT-4-generated language can still provide meaningful insights into the fundamental principles of language processing.

Furthermore, the results obtained from our study align with established linguistic theories and findings from studies using natural language, suggesting that the underlying principles captured by these models are relevant. Additionally, future work will involve validating these findings with naturally occurring datasets and comparing them with neuroimaging data to better understand the parallels between computational models and human brain processing. Thus, while recognizing the limitations, our study provides a foundational step toward bridging the gap between artificial and natural language processing, contributing valuable insights to both computational linguistics and cognitive neuroscience.

In this study, we utilized dimensionality reduction techniques such as t-SNE and MDS to visualize the clustering of BERT's hidden representations for different argument structure constructions. While these methods provide valuable qualitative insights, we acknowledge their inherent limitations, particularly when applied to high-dimensional spaces. To address this, we complemented these visualizations with quantitative measures of cluster separability, specifically the Geometric Density Variance (GDV), which is calculated in the original high-dimensional space rather than the reduced dimensionality projections. GDV offers a more objective assessment of clustering quality and provides additional evidence supporting the distinctiveness of the constructional patterns identified in our analysis. We further emphasize that t-SNE and MDS visualizations are exploratory tools and should not be interpreted as definitive representations of the data. By integrating qualitative visualizations with robust quantitative metrics, we aim to present a balanced and reliable interpretation of the clustering results, highlighting both their strengths and limitations.

Additionally, while the FDR and GDV analyses offer quantitative measures of clustering and differentiation, further qualitative analysis is needed to understand the specific linguistic features that contribute to these patterns. Investigating the impact of different token types on ASC processing in more detail could reveal deeper insights into the underlying mechanisms.

Attention-based analyses, while offering valuable insights, are known to pose interpretability challenges, particularly for token-level interpretations, as noted by Jain and Wallace (2019). In this study, we approach attention analysis as a complementary tool rather than a definitive measure of model behavior. Our primary aim is to explore how attention mechanisms contribute to constructional understanding, without treating attention patterns as the sole indicator of linguistic comprehension. To strengthen the reliability of our findings, we have triangulated attention analysis results with probing classifiers and clustering methods, ensuring that any observed patterns are supported by multiple approaches. Additionally, we explicitly discuss the limitations of attention-based analyses in the manuscript, situating our findings within the broader discourse on attention interpretability. By integrating attention analysis with other methods and addressing its constraints, we provide a nuanced perspective on the role of attention mechanisms in capturing constructional patterns.

In Table 3, we provide a detailed analysis of the relative frequencies of verbs and other word types within each construction to ensure transparency regarding potential lexical biases in our dataset. While some words are highly specific to certain constructions (e.g., "give" for the ditransitive), this specificity does not contradict our findings or claims. Both large language models (LLMs) and the human brain likely utilize such lexical cues, among other features, to interpret and process argument structure constructions. Furthermore, our clustering analyses and visualizations reveal that the four construction types consistently form distinct clusters, independent of individual words or word types. This indicates that the observed patterns are not merely driven by lexical associations but reflect deeper structural and relational patterns that LLMs capture, suggesting that the constructional understanding exhibited by LLMs parallels mechanisms employed by the human brain in processing constructions.

However, we acknowledge that lexical biases could influence probing classifier results, as certain frequent and prototypical lexical items may strongly cue specific constructions. This aligns with the observations of Li et al. (2022), who demonstrated that using nonce words or non-prototypical verbs can mitigate the impact of such biases and better assess true constructional understanding. While our results suggest that BERT encodes constructional information, future work should further address this issue by balancing lexical frequencies and incorporating less predictable word types into the experimental design. These measures would strengthen the validity of probing results and provide more definitive insights into whether the observed patterns reflect genuine constructional understanding or reliance on lexical cues.

Conclusion

In conclusion, BERT effectively captures both the specific and general aspects of grammatical constructions, with its layers progressively integrating lexical, syntactic, and semantic information. This study demonstrates BERT's nuanced understanding of linguistic structures, albeit with certain challenges in differentiating closely related constructions like transitive and resultative sentences.

Our study highlights the sophisticated capabilities of the BERT language model in representing and differentiating between various Argument Structure Constructions. The dynamic and layered nature of BERT's processing, as revealed through clustering, probing, and attention weight analyses, underscores the model's potential to mirror human linguistic processing.

Future research aimed at comparing these computational representations with neuroimaging data will be pivotal in advancing our understanding of the computational and neural mechanisms underlying language comprehension. In particular, comparing our computational findings with neuroimaging data during continuous speech perception will be crucial in bridging the gap between computational models and the neural mechanisms of language understanding. Such comparisons could validate whether the patterns observed in BERT align with how the human brain

processes different ASCs, offering a more comprehensive view of language processing.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

PR: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. AS: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing – original draft. PK: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): KR 5148/3-1 (project number 510395418), KR 5148/5-1 (project number 542747151), and GRK 2839 (project number 468527017) to PK and grant SCHI 1482/3-1 (project number 451810794) to AS.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

AI Explosion (2017). *Spacy-Industrial-Strength Natural Language Processing in Python*. Available at: https://spacy.io (accessed July 31, 2024).

Alain, G. (2016). Understanding intermediate layers using linear classifier probes. *arXiv [Preprint]*. arXiv:1610.01644.

Belinkov, Y. (2022). Probing classifiers: promises, shortcomings, and advances. *Comput. Linguist*. 48, 207–219. doi: 10.1162/coli_a_00422

Bonial, C., and Madabushi, H. T. (2024). "A construction grammar corpus of varying schematicity: a dataset for the evaluation of abstractions in language models," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 243–255.

Chronis, G., Mahowald, K., and Erk, K. (2023). A method for studying semantic construal in grammatical constructions with interpretable contextual embedding spaces. *arXiv* [preprint] arXiv:2305.18598. doi: 10.18653/v1/2023.acl-long.14

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. *arXiv* [preprint] arXiv:1906.04341. doi: 10.18653/v1/W19-4828

Cohen, Y., Engel, T. A., Langdon, C., Lindsay, G. W., Ott, T., Peters, M. A., et al. (2022). Recent advances at the interface of neuroscience and artificial neural networks. *J. Neurosci*. 42, 8514–8523. doi: 10.1523/JNEUROSCI.1503-22.2022

Cox, M. A., and Cox, T. F. (2008). "Multidimensional scaling," in *Handbook of Data Visualization* (Cham: Springer), 315–347.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* [preprint] arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805

Fujita, K. (1989). "Knowledge of language: its nature, origin, and use," in *The Convergence Series*, ed. N. Chomsky Cambridge, MA: MIT Press, 213–231.

Goldberg, A. (2019). *Explain Me This*. Princeton NJ: Princeton University Press. doi: 10.2307/j.ctvc772nn

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: University of Chicago Press.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends Cogn. Sci*. 7, 219–224. doi: 10.1016/S1364-6613(03)00080-9

Goldberg, A. E. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press on Demand.

Goldberg, A. E. (2009). *The Nature of Generalization in Language*. Walter de Gruyter GmbH & Co. KG Germany.

Guan, Y., Leng, J., Li, C., Chen, Q., and Guo, M. (2020). How far does bert look at distance-based clustering and analysis of bert ′ s attention. *arXiv* [preprint] arXiv:2011.00943. doi: 10.18653/v1/2020.coling-main.342

Hagberg, A., Swart, P. J., and Schult, D. A. (2008). *Exploring network structure, dynamics, and function using NetworkX*. Los Alamos, NM: Los Alamos National Laboratory (LANL).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* 95, 245–258. doi: 10.1016/j.neuron.2017.06.011

Henningsen-Schomers, M. R., and Pulvermüller, F. (2022). Modelling concrete and abstract concepts using brain-constrained deep neural networks. *Psychol. Res*. 86, 2533–2559. doi: 10.1007/s00426-021-01591-6

Hewitt, J., and Manning, C. D. (2019). "A structural probe for finding syntax in word representations," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comp. Sci. Eng*. 9, 90–95. doi: 10.1109/MCSE.2007.55

Jain, S., and Wallace, B. C. (2019). Attention is not explanation. *arXiv [Preprint]*. arXiv:1902.10186.

Jawahar, G., Sagot, B., and Seddah, D. (2019). "What does bert learn about the structure of language?," in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Kim, S. J., Magnani, A., and Boyd, S. (2005). "Robust fisher discriminant analysis," in *Advances in Neural Information Processing Systems*, 18.

Kovaleva, O., Romanov, A., Rogers, A., and Rumshisky, A. (2019). Revealing the dark secrets of bert. *arXiv* [preprint] arXiv:1908.08593. doi: 10.18653/v1/D19-1445

Krauss, P. (2023). *Künstliche Intelligenz und Hirnforschung: Neuronale Netze, Deep Learning und die Zukunft der Kognition*. Cham: Springer.

Krauss, P. (2024). *Artificial Intelligence and Brain Research: Neural Networks, Deep Learning and the Future of Cognition*. Cham: Springer Nature.

Krauss, P., Metzner, C., Joshi, N., Schulze, H., Traxdorf, M., Maier, A., et al. (2021). Analysis and visualization of sleep stages based on deep neural networks. *Neurobiol. Sleep Circad. Rhythms* 10:100064. doi: 10.1016/j.nbscr.2021.100064

Krauss, P., Metzner, C., Schilling, A., Tziridis, K., Traxdorf, M., Wollbrink, A., et al. (2018a). A statistical method for analyzing and comparing spatiotemporal cortical activation patterns. *Sci. Rep*. 8, 1–9. doi: 10.1038/s41598-018-23765-w

Krauss, P., Prebeck, K., Schilling, A., and Metzner, C. (2019a). Recurrence resonance in three-neuron motifs. *Front. Comput. Neurosci*. 13:64. doi: 10.3389/fncom.2019.00064

Krauss, P., Schilling, A., Bauer, J., Tziridis, K., Metzner, C., Schulze, H., et al. (2018b). Analysis of multichannel eeg patterns during human sleep: a novel approach. *Front. Hum. Neurosci*. 12:121. doi: 10.3389/fnhum.2018.00121

Krauss, P., Schuster, M., Dietrich, V., Schilling, A., Schulze, H., and Metzner, C. (2019b). Weight statistics controls dynamics in recurrent neural networks. *PLoS ONE* 14:e0214541. doi: 10.1371/journal.pone.0214541

Krauss, P., Zankl, A., Schilling, A., Schulze, H., and Metzner, C. (2019c). Analysis of structure and dynamics in three-neuron motifs. *Front. Comput. Neurosci*. 13:5. doi: 10.3389/fncom.2019.00005

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29, 115–129. doi: 10.1007/BF02289694

Kruskal, J. B., and Wish, M. (1978). *Multidimensional Scaling*. Newcastle upon Tyne: SAGE.

Li, B., Zhu, Z., Thomas, G., Rudzicz, F., and Xu, Y. (2022). Neural reality of argument structure constructions. *arXiv* [preprint] arXiv:2202.12246. doi: 10.18653/v1/2022.acl-long.512

Liu, C., and Chersoni, E. (2023). "On quick kisses and how to make them count: A study on event construal in light verb constructions with BERT," in *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 367–378.

Madabushi, H. T., Romain, L., Divjak, D., and Milin, P. (2020). CxGBERT: BERT meets construction grammar. *arXiv* [preprint] arXiv:2011.04134. doi: 10.48550/arXiv.2011.04134

Mahowald, K. (2023). A discerning several thousand judgments: GPT-3 rates the article+ adjective+ numeral+ noun construction. *arXiv* [preprint] arXiv:2301.12564. doi: 10.18653/v1/2023.eacl-main.20

Metzner, C., and Krauss, P. (2022). Dynamics and information import in recurrent neural networks. *Front. Comput. Neurosci*. 16:876315. doi: 10.3389/fncom.2022.876315

Metzner, C., Schilling, A., Traxdorf, M., Schulze, H., and Krauss, P. (2021). Sleep as a random walk: a super-statistical analysis of eeg data across sleep stages. *Commun. Biol*. 4:1385. doi: 10.1038/s42003-021-02912-6

Metzner, C., Schilling, A., Traxdorf, M., Schulze, H., Tziridis, K., and Krauss, P. (2023a). Extracting continuous sleep depth from eeg data without machine learning. *Neurobiol. Sleep Circad. Rhythms* 14:100097. doi: 10.1016/j.nbscr.2023.100097

Metzner, C., Schilling, A., Traxdorf, M., Tziridis, K., Maier, A., Schulze, H., et al. (2022). Classification at the accuracy limit: facing the problem of data ambiguity. *Sci. Rep*. 12:22121. doi: 10.1038/s41598-022-26498-z

Metzner, C., Yamakou, M. E., Voelkl, D., Schilling, A., and Krauss, P. (2023b). Quantifying and maximizing the information flux in recurrent neural networks. *arXiv* [preprint] arXiv:2301.12892. doi: 10.48550/arXiv.2301.12892

Metzner, C., Yamakou, M. E., Voelkl, D., Schilling, A., and Krauss, P. (2024). Quantifying and maximizing the information flux in recurrent neural networks. *Neural Comput*. 36:351–384. doi: 10.1162/neco_a_01651

Misra, K., and Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *arXiv* [preprint] arXiv:2403.19827. doi: 10.18653/v1/2024.emnlp-main.53

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol*. 37, 1482–1492. doi: 10.1038/s41587-019-0336-3

Oliphant, T. E. (2007). Python for scientific computing. *Comp. Sci. Eng*. 9, 10–20. doi: 10.1109/MCSE.2007.58

Pande, M., Budhraja, A., Nema, P., Kumar, P., and Khapra, M. M. (2021). The heads hypothesis: A unifying statistical approach towards understanding multi-headed attention in bert. *Proc. AAAI Conf. Artif. Intellig*. 35, 13613–13621. doi: 10.1609/aaai.v35i15.17605

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res*. 12, 2825–2830.

Pulvermüller, F. (2002). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge: Cambridge University Press.

Pulvermüller, F. (2012). Meaning and the brain: the neurosemantics of referential, interactive, and combinatorial knowledge. *J. Neurolinguistics* 25, 423–459. doi: 10.1016/j.jneuroling.2011.03.004

Pulvermüller, F. (2023). Neurobiological mechanisms for language, symbols and concepts: clues from brain-constrained deep neural networks. *Prog. Neurobiol*. 230:102511. doi: 10.1016/j.pneurobio.2023.102511

Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M. R., and Wennekers, T. (2021). Biological constraints on neural network models of cognitive function. *Nat. Rev. Neurosci*. 22, 488–502. doi: 10.1038/s41583-021-00473-5

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). *Improving Language Understanding With Unsupervised Learning*. Available at: https://openai.com/research/language-unsupervised

Ramezani, P., Schilling, A., and Krauss, P. (2024). Analysis of argument structure constructions in a deep recurrent language model. *arXiv* [preprint] arXiv:2408.03062. doi: 10.48550/arXiv.2408.03062

Savage, N. (2019). How ai and neuroscience drive each other forwards. *Nature* 571, S15–S15. doi: 10.1038/d41586-019-02212-4

Schilling, A., Maier, A., Gerum, R., Metzner, C., and Krauss, P. (2021a). Quantifying the separability of data classes in neural networks. *Neural Netw*. 139, 278–293. doi: 10.1016/j.neunet.2021.03.035

Schilling, A., Tomasello, R., Henningsen-Schomers, M. R., Zankl, A., Surendra, K., Haller, M., et al. (2021b). Analysis of continuous neuronal activity evoked by natural speech with computational corpus linguistics methods. *Lang, Cognit. Neurosci*. 36, 167–186. doi: 10.1080/23273798.2020.1803375

Sung, H., and Kyle, K. (2024). "Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7302–7314.

Surendra, K., Schilling, A., Stoewer, P., Maier, A., and Krauss, P. (2023). "Word class representations spontaneously emerge in a deep neural network trained on next word prediction," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, 1481–1486.

Tenney, I., Das, D., and Pavlick, E. (2019a). BERT rediscovers the classical nlp pipeline. *arXiv* [preprint] arXiv:1905.05950. doi: 10.18653/v1/P19-1452

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., et al. (2019b). What do you learn from context? Probing for sentence structure in contextualized word representations. *arXiv* [preprint] arXiv:1905.06316. doi: 10.48550/arXiv.1905.06316

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika* 17, 401–419. doi: 10.1007/BF02288916

Traxdorf, M., Krauss, P., Schilling, A., Schulze, H., and Tziridis, K. (2019). Microstructure of cortical activity during sleep reflects respiratory events and state of daytime vigilance. *Somnologie* 23, 72–79. doi: 10.1007/s11818-019-0201-0

Tseng, Y. H., Shih, C. F., Chen, P. E., Chou, H. Y., Ku, M. C., and Hsieh, S. K. (2022). "CxLM: A construction and context-aware language model," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6361–6369.

Vallejos, C. A. (2019). Exploring a world of a thousand dimensions. *Nat. Biotechnol*. 37, 1423–1424. doi: 10.1038/s41587-019-0330-9

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res*. 9:11.

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Comp. Sci. Eng*. 13, 22–30. doi: 10.1109/MCSE.2011.37

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst*. 30.

Veenboer, T., and Bloem, J. (2023). "Using collostructional analysis to evaluate BERT'S representation of linguistic constructions," in *Findings of the Association for Computational Linguistics: ACL 2023*, 12937–12951.

Wang, S., and Jiang, J. (2015). Learning natural language inference with LSTM. *arXiv* [preprint] arXiv:1512.08849. doi: 10.48550/arXiv.1512.08849

Wang, S., Li, D., Song, X., Wei, Y., and Li, H. (2011). A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert Syst. Appl*. 38, 8696–8702. doi: 10.1016/j.eswa.2011.01.077

Waskom, M. L. (2021). Seaborn: statistical data visualization. *J. Open Source Softw*. 6:3021. doi: 10.21105/joss.03021

Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use T-SNE effectively. *Distill* 1:e2. doi: 10.23915/distill.00002

Weissweiler, L., He, T., Otani, N., Mortensen, D. R., Levin, L., and Schütze, H. (2023a). Construction grammar provides unique insight into neural language models. *arXiv* [preprint] arXiv:2302.02178. doi: 10.48550/arXiv.2302.02178

Weissweiler, L., Hofmann, V., Köksal, A., and Schütze, H. (2022). "The better your syntax, the better your semantics? probing pretrained language models for the english comparative correlative," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10859–10882.

Weissweiler, L., Hofmann, V., Köksal, A., and Schütze, H. (2023b). Explaining pretrained language models' understanding of linguistic structures using construction grammar. *Front. Artif. Intellig.* 6:1225791. doi: 10.3389/frai.2023.1225791

Wilson, M., Petty, J., and Frank, R. (2023). How abstract is linguistic generalization in large language models? experiments with argument structure. *Trans. Assoc. Comp. Linguist.* 11, 1377–1395. doi: 10.1162/tacl_a_00608

Xu, L., Wu, J., Peng, J., Gong, Z., Cai, M., and Wang, T. (2023). Enhancing language representation with constructional information for natural language understanding. *arXiv* [preprint] arXiv:2306.02819. doi: 10.18653/v1/2023.acl-long.258

Zhou, S., Weissweiler, L., He, T., Schütze, H., Mortensen, D. R., and Levin, L. (2024). "Constructions are so difficult that even large language models get them right for the wrong reasons," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3804–3811.