Proceedings of CONF-SEML 2026 Symposium: Importance of Machine Learning Methods and Analysis in Engineering
DOI: 10.54254/2755-2721/2026.AS32075

Interpretability of Natural Language Processing Models and Alignment with Human Language Mechanisms

Yang Fan

School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China
yangf4ai@gmail.com

Abstract. Despite significant progress in Large Language Models (LLMs) in recent years, their inner mechanisms for processing linguistic representations remain unclear. This lack of transparency limits the understanding of the models and hinders efforts to improve their reliability and alignment with human cognition. By reviewing existing literature, the paper analyzes the comparison between neuroimaging-based methods like fMRI and EEG and Natural Language Processing (NLP) models, revealing similarities in language processing between artificial neural networks and the human brain. The results indicate that, although there are commonalities in language processing, current models have yet to fully align with human perception. And future studies could advance the field of neural AI by incorporating causal interventions and multimodal data integration, promoting the biological plausibility and interpretability of AI systems.

Keywords: Large Language Models (LLMs), Natural Language Processing (NLP), Cognitive Neuroscience, Interpretability

1. Introduction

In recent years, large language models (LLMs) built on the Transformer architecture have brought about transformative breakthroughs in the field of Natural Language Processing (NLP). As a result, these models perform exceptionally well in tasks such as text generation, machine translation, and intricate question answering, while also revealing emergent "comprehension" capabilities [1,2]. Yet, this rapid growth in capability has been accompanied by a growing lack of transparency regarding the models' internal workings. Despite the GPT series, BERT, and their variants consistently setting new performance benchmarks, they still fundamentally function as "black box" systems. This lack of transparency obstructs our scientific comprehension of how linguistic information is represented and processed within artificial neural networks, raising significant concerns about trustworthiness, robustness, and alignment safety [3-6]. At the same time, cognitive neuroscience uses fMRI, EEG, and behavioral studies to investigate the underlying mechanisms of language processing in the brain. The intersection of neuroscience and AI helps design more biologically realistic neural networks, while NLP offers models to test cognitive processes [7-10]. By analyzing activity patterns in artificial neural networks and the human brain in language processing, similarities in information processing can be revealed, pointing to clear directions for future cross-disciplinary research. This

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.AS32075

paper outlines NLP techniques that align with neuroscientific frameworks, explores current methods to linking model representations with brain activity, and identifies key challenges in integrating artificial neural networks with neuroscientific models. By incorporating these emerging methods, this paper seeks to build a more structured, causally coherent, and biologically grounded framework for explanation.

2. Explainable NLP methods and human language processing mechanisms

2.1. Human language processing mechanisms

Through techniques such as fMRI, EEG, and MEG, cognitive neuroscience has revealed the spatial distribution and temporal dynamics of language processing in the brain, thus clarifying the roles of different brain regions in syntactic and semantic processing. By using neuroimaging and behavioral experiments, it explores the organization of linguistic processes across time and space in the brain.

Traditionally, language functions have been localized to Broca's area, which is responsible for syntax and production, and Wernicke's area, which is responsible for semantics and comprehension, both located in the left hemisphere. However, modern neuroimaging studies reveal a more extensive and distributed "language network" spanning multiple regions across the temporal, frontal, and parietal lobes. Huth et al. conducted an fMRI study using natural language stimuli and found that semantic information is arranged in a detailed pattern across various cortical areas [11]. Different cortical regions show sensitivity to specific semantic categories, such as social, visual, numerical. This indicates that the brain does not process "language" as an isolated module but rather integrates it extensively into perceptual and cognitive networks.

EEG and MEG, with their millisecond-level temporal resolution, complement the high spatial resolution of fMRI by uncovering the ultrafast dynamics of language processing. For instance, the N400 is a negative wave that typically appears around 400 milliseconds after stimulus presentation and is highly sensitive to semantic violations, such as the statement "He drank a gulp of cement," reflecting the brain's process of semantic integration and expectation checking. In contrast, the P600 is commonly associated with syntactic processing difficulties. And these electrophysiological markers indicate that the brain operates as a highly predictive system during language processing, continually forecasting subsequent inputs based on context. This is similar to how large language models predict the next word based on context.

2.2. Interpretability methods in NLP models

Explainable NLP improves the transparency of language models by uncovering how they encode and process language structures. With the widespread application of deep learning technologies, particularly the Transformer architecture, interpretability methods have evolved from simple weight analysis to more complex probing and attribution techniques.

In this process, Madsen et al. classified post-hoc interpretability methods, which are techniques applied after model training, primarily distinguishing them based on the level of abstraction in local explanations and the amount of information contained in the model components [12]. Specifically, feature attribution methods interpret models by measuring the impact of each token in the input sequence on the final prediction. Gradient-based methods, like Integrated Gradients and Saliency Maps, assess the importance of different parts by calculating the gradient of the output with respect to the input, akin to identifying "receptive fields" in neuroscience. In addition, perturbation-based methods such as LIME and SHAP infer the contribution of different parts by masking or replacing

DOI: 10.54254/2755-2721/2026.AS32075

portions of the input and observing changes in the output, which is similar to "lesion studies" in neuroscience, where brain region damage is analyzed to assess functional loss.

Moreover, attention analysis is another method that directly visualizes self-attention weights in Transformer models. Although this approach is controversial, it is still widely used, especially when demonstrating how a model "focuses" on specific parts of the context, providing valuable insights. Another common method is probe/diagnostic classifiers. This technique involves training a simple linear classifier on frozen model layers to detect whether the representations in those layers contain specific linguistic properties, such as part-of-speech tags or syntactic tree depth. This approach is similar to decoding neural signals in neuroscience to identify specific cognitive states. Example-based explanations reveal the model's decision-making process by identifying the training samples that have the greatest impact on the current prediction. Influence function analysis identifies the samples that most affect the model's output, thus demonstrating the model's behavior.

3. Bridging constructs in NLP interpretability and cognitive neuroscience

3.1. Representation alignment

The main link between NLP and neuroscience is representational alignment, comparing language model embeddings with brain responses. This theory suggests that, despite physical differences, the internal representations of artificial neural networks and the biological brain should align when performing the same task.

3.1.1. Encoding models and hierarchical correspondences

Encoding models predict neural activity based on text features, while decoding models reconstruct linguistic meaning from brain signals or eye-tracking data. Existing studies indicate that when using narrative or sentence comprehension datasets, the intermediate layers of large Transformer models align most closely with the activity of higher cortical areas involved in semantic integration, such as the superior temporal sulcus (STS) and angular gyrus (AG). Specifically, shallow layers (1st to 4th) of models such as BERT are primarily associated with activity in the auditory cortex and language areas, which handle phonological and morphological processing; the middle layers (5th to 8th) correlate highly with activity in syntactically dense regions; and deep layers (9th to 12th) show a strong association with the default mode network (DMN) and brain regions involved in processing long narrative comprehension [10]. And this suggests that language models may implicitly capture a hierarchical structure similar to human language processing, from perceptual features to syntactic structures, and onto semantic and pragmatic integration.

3.1.2. Representational similarity analysis and challenges

Representational Similarity Analysis (RSA) compares the representational structures of different systems through information geometry [9]. RSA assumes that when the brain produces similar neural activation patterns for "cat" and "dog," the model's embedding representations for these two words should also be similar (e.g., with high cosine similarity). This consistency suggests that there may be alignment between the brain and the model in terms of information representation. RSA allows for cross-modal and cross-system comparisons of different data types, such as neural firing rates with fMRI BOLD signals, or the representations of the brain with those of artificial intelligence models. This makes RSA an essential tool for connecting neuroscience and NLP, revealing the similarities between the brain and models in language processing. But current methods

DOI: 10.54254/2755-2721/2026.AS32075

rely on linear mappings, which do not capture the nonlinear nature of neural activity. Additionally, limitations in dataset size, individual differences, and the low signal-to-noise ratio of fMRI data hinder the generalizability of findings. Future research could explore more complex alignment frameworks, like nonlinear transformations or task-specific constraints, to better understand the differences between artificial and biological systems. Investigating whether representational shifts during model fine-tuning align with neural plasticity in human language learning could be a key area for further study.

3.2. Attention and cognitive relevance

The "attention" mechanism in the Transformer architecture adjusts the importance of words in the input sequence while processing the current word, similar to human attention.

3.2.1. The "attention as gaze" hypothesis

Studies have shown a significant correlation between the model's attention weights and human eye-tracking data. When reading, humans tend to linger longer on ambiguous or information-dense words. Similarly, the model assigns higher attention to such words when processing them, aiding in the interpretation of the internal state. For instance, in a coreference resolution task, humans tend to look back at the word "trophy" when processing "it," and the model also assigns higher attention to "trophy" in this context. Moreover, the high "surprise" or attention given to a word by the model often matches the amplitude of the N400 effect in human EEG. This suggests that the model's "prediction error" signal shares a similar function with the brain's predictive coding mechanism.

3.2.2. Comparison with human attention

Some argue that "attention is not explanation," suggesting that attention weights function merely as a "gatekeeper" for the model's internal information flow, and do not directly reflect the importance of individual features [13]. In some cases, words that receive high attention may contribute little to the output prediction and could even be statistical artifacts. Besides, critics point out that attention weights are intermediate computational steps and may not accurately reflect semantic importance [13]. For example, a model might allocate 90% of attention to commas or separators, while only 10% goes to content words, yet the actual prediction is mainly driven by the latter. Thus, visualizing attention alone can be misleading, as it often highlights processing bottlenecks rather than the semantic core.. This is similar to human visual blind spots, where some information may be noticed but not consciously perceived, or deeply processed without direct gaze [7].

3.3. Probing and behavioral evidence

The probing method involves training a small supervised classifier based on the frozen internal states of a LLM to predict whether specific linguistic features, such as tense, plurality, or syntactic tree depth, are encoded in the model's representations. This helps determine whether the language model encodes specific linguistic structures in its representations.

3.3.1. The psycholinguistic analogy

The probing method is similar to behavioral tests in psycholinguistics. In psychology, since the "syntactic module" in the brain cannot be directly observed, its existence is typically inferred by

DOI: 10.54254/2755-2721/2026.AS32075

analyzing response times to grammatical versus ungrammatical sentences. Similarly, in AI, since "syntactic vectors" cannot be directly accessed, probing is employed to evaluate whether syntactic information can be linearly decoded from the model's internal representations. Indeed, existing research has demonstrated that as models increase in size, they progressively develop specialized subspaces to manage different linguistic tasks. For example, some layers specialize in processing morphological features, others in syntactic analysis, while additional layers manage the overall discourse structure. This "functional specialization" is similar to the modular organization of the human brain cortex, where regions like the superior temporal gyrus and inferior frontal gyrus are involved in distinct language functions [1].

3.3.2. Limitations of probing

Although the probing method can reveal whether information is decodable, its effectiveness remains debated. In particular, a successful successful probing task may indicate that information is decodable, but this does not necessarily mean that the model actually utilizes this information. Besides, control tasks proposed by Hewitt and Liang suggest that some robust detectors can even "learn" tasks from randomly initialized networks, showing that detection performance may reflect the detector's inherent capacity rather than the quality of the model's representations [14].

Similarly, human behavioral responses do not always reveal the underlying mechanisms of the brain. For example, humans may be able to intuitively make correct syntactic judgments but often cannot explicitly explain the rules behind these judgments (i.e., implicit knowledge). Therefore, a promising direction is to design unified evaluation frameworks that directly compare the model's probing results with human experimental data, bridging the gap between computational evidence and cognitive evidence. For instance, by comparing a model's performance on psycholinguistic benchmark tasks with human participant behavioral data, a more accurate understanding of the similarities and differences between model and human cognition can be achieved. As shown in Table 1, these comparisons reveal both commonalities and limitations in the language processing mechanisms of NLP methods and cognitive neuroscience.

Table 1. Bridging constructs between NLP interpretability and cognitive neuroscience

Bridging Construct	NLP Method	Cognitive Parallel	Main Limitation
Representation Alignment	Encoding Models, Decoding Models, Rep. Similarity Analysis (RSA)	fMRI/EEG data mapping, Computational models of brain activity	Strong linear assumption, Low signal-to-noise ratio, Individual variability, Correlation ≠ causation
Attention & Cognitive Relevance	Attention visualization, Attention head analysis	Eye-tracking, ERPs (e.g., N400, P600), Reading time experiments	"Attention is not explanation", Correlation, not causation, Statistical artifact vs. cognitive process
Probing & Behavioral Evidence	Probing tasks, Diagnostic classifiers	Psycholinguistic experiments, Behavioral measures (e.g., reaction time, acceptability)	"Decodable" ≠ "Used", Probe may be too powerful (learns the task itself, not the representation)

4. Existing challenges and future directions

The main issue faced by current research is the over-reliance on correlations [10,15]. For example, finding that a certain layer of BERT is correlated with fMRI activity in the left temporal lobe does not prove that the brain is performing the same computational processes as the Transformer model. It merely suggests that the representational structures of both are similar under the same input. To advance the research, it is essential to address the issue of causal interpretability. Ablation studies,

DOI: 10.54254/2755-2721/2026.AS32075

intervention methods, and neurofeedback are key to achieving this goal. Ablation studies assess performance changes by disabling specific neurons or attention heads in the model, revealing their functions. Intervention methods alter the model's activation patterns to test whether they can trigger the desired output. Neurofeedback uses brain data to guide model training, enabling the model to learn representations similar to the brain's and also predict the brain's responses [7].

However, the issue of noise in biological data cannot be ignored. fMRI signals are slow with low signal-to-noise ratio, while EEG, though faster, has poor spatial resolution. Individual differences also make standardizing brain activity challenging. Thus, future research needs to develop metrics that account for individual differences and gather more neural response data from natural language contexts, especially involving long-term contexts like listening to podcasts or reading books, rather than isolated sentences [7]. Besides, LLMs rely solely on text data for training, disconnecting their representations from physical perception. Human language is closely tied to perception; mentioning "apple" activates not only language areas but also sensory regions such as vision and taste [10,16]. Future multimodal large language models (MLLMs), combining video, audio, and text, may offer more integrated representations similar to human brain activity. At present, there is no standardized framework in neural AI research. Different labs use different datasets, like the Harry Potter fMRI dataset and the Pereira dataset, and models, with no cross-laboratory comparison standard. Future work should establish a unified evaluation system to test models across multiple modalities (fMRI, EEG, eye-tracking) and language levels (phonetics, syntax, semantics, discourse). Such a system would provide a standardized "brain score" for each new AI model, helping advance the field of neural AI [7].

5. Conclusion

This paper explores the intersection of explainable NLP and cognitive neuroscience, highlighting that aligning computational representations with neural data may reveal similarities in language understanding between machines and humans. This finding provides a new perspective for further understanding the relationship between artificial and human intelligence. However, current research faces several challenges, particularly in terms of causal interpretability and biological data noise. Existing NLP models still struggle to align with human perception, and a standardized framework for cross-laboratory comparisons remains lacking. Future research can boost model interpretability and efficiency through causal analysis, multimodal techniques, and a unified evaluation system. In addition, interdisciplinary collaboration will offer opportunities to develop intelligent systems more closely aligned with human cognitive mechanisms, advancing the application and development of AI technologies across various fields.

References

[1] Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent abilities of large language models. arXiv preprint arXiv: 2206.07682.
[2] Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are emergent abilities of large language models a mirage? Advances in Neural Information Processing Systems, 36, 55565-55581.
[3] Liu, Y., Yao, Y., Ton, J.F., et al. (2024). Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. arXiv: 2308.05374.
[4] Kumar, P., & Mishra, S. (2025). Robustness in large language models: A survey of mitigation strategies and evaluation metrics. arXiv: 2505.18658.
[5] d. Cerqueira, J.S., Kemell, K.K., Rousi, R., et al. (2025). Mapping trustworthiness in large language models: A bibliometric analysis bridging theory to practice. arXiv: 2503.04785.

DOI: 10.54254/2755-2721/2026.AS32075

[6] Aljohani, M., Hou, J., Kommu, S., & Wang, X. (2025). A comprehensive survey on the trustworthiness of large language models in healthcare. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2025 (pp. 6720-6748).
[7] Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. Neuron, 95(2), 245-258.
[8] Li, J., Zaikin, A., Zhang, X., & Chen, S. (2022). Editorial: Closed-loop iterations between neuroscience and artificial intelligence. Frontiers in Systems Neuroscience, 16, 1002095.
[9] EmergentMind. (2025). NeuroAI: Bridging Brain and Machine. https: //www.emergentmind.com/topics/neuroai
[10] Shankar, S., Pan, Y., Jiang, H., et al. (2025). Bridging brains and machines: A unified frontier in neuroscience, artificial intelligence, and neuromorphic systems. arXiv: 2507.10722.
[11] Huth, A.G., de Heer, W.A., Griffiths, T.L., et al. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532, 453-458.
[12] Madsen, A., Reddy, S., & Chandar, S. (2023). Post-hoc interpretability for neural NLP: A survey. ACM Computing Surveys, 55(8), 1-42.
[13] Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 3543-3556.
[14] Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. In the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2733-2743.
[15] Momennejad, I. (2023). A rubric for human-like agents and NeuroAI. Philosophical Transactions of the Royal Society B, 378(1869), 20210446.
[16] Gao, R. (2024). Beyond #neuroai, we need #ai4neuro and we're going to suck at it (at first). https: //www.rdgao.com/blog/2024/08/14