ARTICLE | Article in Press

Transformer language models reveal distinct patterns in aphasia subtypes and longitudinal change

S. Saeed Ahmadi · Julius Fridriksson · Dirk B.den Ouden

Received: 7 December 2025 / Accepted: 20 October 2026
© The Author(s) 2026

Abstract

Language impairments in aphasia are characterized by various representational disruptions that may be reflected in discourse production. This research examines the capacity of transformer-based language models, particularly GPT-2, to serve as a computational framework for analyzing variations in aphasic narrative speech. A longitudinal dataset of narrative speech samples collected at six time points from individuals with aphasia (N = 47; baseline cross-sectional sample N = 57) was utilized as part of an intervention study. All transcripts were processed via the GPT-2 language model to obtain activation values from each of the 12 transformer layers. Statistically significant differences in activation values across aphasia subtypes were found at every layer (all $p < .001$), with the most pronounced effects in the deeper layers. Pairwise Tukey HSD tests revealed consistent distinctions between Broca's aphasia and both Anomic and Wernicke's aphasia, suggesting a shared activation profile between the latter two. Longitudinal tests revealed significant changes over time, especially in the final three layers (10–12). These findings suggest that transformer-based activation patterns reflect meaningful variation in aphasic discourse and may align with clinical observations. Overall, GPT-2 provides a scalable tool to model representational dynamics in aphasia and may offer exploratory insights into language representation in aphasic discourse.

Introduction

Aphasia is an acquired language disorder that typically occurs after a stroke and significantly affects communication, quality of life, and social participation[1]. While lesion characteristics, neuronal integrity, and treatment modalities affect the intricate process of language change after stroke, scalable methods for monitoring longitudinal change in discourse remain limited.

Aphasia is not a homogeneous condition; it comprises various subtypes that differ in clinical presentation and neural underpinnings[2]. The classification of aphasia types remains controversial. Broca's aphasia is a nonfluent type of aphasia characterized by agrammatic speech production, although comprehension remains generally intact. Wernicke's aphasia is characterized by fluent yet semantically empty speech and compromised comprehension[3]. Other subtypes of aphasia further highlight the heterogeneity of language deficits following stroke. It is important to consider these distinctions, as collapsing across subtypes may obscure clinically meaningful

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



https://doi.org/10.1038/s41598-026-77518-1

Published online: 30 October 2026

scientific reports

variation in recovery patterns and neural reorganization. By explicitly examining differences within subtypes, studies can identify whether computational markers generalize across aphasia or reflect subtype-specific impairment, ultimately supporting more accurate diagnostics and individualized treatment planning.

Recent advancements in artificial intelligence, particularly in natural language processing (NLP), have introduced novel methods for characterizing language deficits. NLP enables computational analysis of fluency, coherence, and syntax in clinical speech, offering new avenues for assessing language impairments, and progress in this area has accelerated with the development of large language models. Unexpectedly, the internal representations of these models have even been found to align with human brain activation during language processing[4–7].

Transformers, especially models such as GPT-2, process language via multi head self-attention and deep hierarchical structures that encode linguistic and semantic information across layers[8,9]. Different attention heads and layers have been suggested to specialize in particular linguistic phenomena[10]. The architectural characteristics of the GPT-2 make it a valuable instrument for investigating language impairments in aphasia at a representational level, given the demonstrated correlations between layer activation in the GPT-2 and neural activity[4]. In this context, the activation values extracted from each layer represent model-derived summary measures of the layer's hidden-state responses to linguistic input. These measures are exploratory features derived from the model's internal representations rather than direct evidence of specific linguistic functions. Previous studies have suggested that deeper layers may be more sensitive to higher-level semantic and discourse-related information, whereas surface layers are more sensitive to lexical and syntactic characteristics[11–13]. Importantly, human programmers have not specifically coded these levels to encode certain linguistic functions. During training on large-scale text corpora, the model adjusts its internal weights to capture statistical regularities of language. As a result, any specialization observed in a particular layer is not predefined but emerges through learning[11,14].

Within this architecture, large language models encode text across successive levels, where representations in different layers may more strongly reflect distinct linguistic features such as lexical, syntactic, or semantic information. The surface layers primarily address superficial patterns, including word forms, punctuation, and basic part-of-speech tags. The middle layers begin to encode grammatical structures, such as rules, clause demarcations, and sentence constructions, such as how one might evaluate phrase structure in continuous speech. The deeper layers emphasize semantic significance and discourse-level characteristics, addressing broader context, reference resolution, and speaker purpose[12,15]. This tiered organization provides a framework for interpreting layer-wise analyses of aphasic and typical speech.

The OpenAI GPT structure is based on multi head self-attention[16]. A key advantage of self-attention is its interpretability, which allows researchers to examine which input elements are treated as most relevant for text parsing[17]. The small version of GPT-2 includes 12 layers and 12 attention heads.

In this study, we aimed to investigate whether the internal layer activation patterns of GPT-2 differ between individuals with various types of aphasia and neurotypical individuals. Specifically, we examine (1) whether the representational patterns in aphasic speech differ from typical (unimpaired) language patterns, (2) whether these differences vary across aphasia subtypes, and (3) how these patterns change over the course of a longitudinal intervention study.

If significant associations are discovered between large-language-model activation patterns and standardized language assessment scores, these patterns may serve as exploratory computational correlates of aphasic language variation. Recent work has shown that surprisal-based indices from GPT-style models discriminate aphasia presence and subtype more accurately than traditional linguistic measures do[18]. By obtaining these markers from precise, naturalistic speech samples, clinicians could (i) create objective, low-effort tools for diagnosis and monitoring, (ii) enhance subtype classification for personalized treatment, and (iii) may support future research on prediction of therapy responsiveness. Furthermore, by bridging the gap between neurolinguistics and language model studies, this research contributes to the theoretical understanding of the reconfiguration of language following a stroke.

We analyzed narrative speech collected from an intervention study in individuals with aphasia over multiple time points, extracting GPT-2 hidden-state activations. Statistical analyses were used to evaluate longitudinal changes and characterize differences across aphasia subtypes based on model-derived representations.

Methods

This study used the GPT-2 across six time points to examine longitudinal changes in internal language representations in individuals with aphasia enrolled in an intervention study aimed at identifying predictors of response to treatment (POLAR dataset[19]).

Participants

The participants were selected from a larger dataset, which included 102 individuals with poststroke aphasia[19]. From this group, we included the 57 individuals who produced at least 100 words in their speech sample at the first time point. In addition, we included 10 non-aphasic participants without a history of stroke who served as neurologically typical controls. For the longitudinal analyses, a subset of 47 aphasic individuals was selected on the basis of having sufficient narrative output (at least 100 words per story) and attending assessment sessions at all six time points. Within this longitudinal subsample of 47 participants, subtype distribution was as follows: Broca's (n=17), Wernicke's (n=4), Conduction (n=8), Anomic (n=16), Global (n=2). The participant demographics and clinical characteristics are presented in Appendix A, Table A1. Unless otherwise specified, baseline and cross-sectional analyses were conducted using the full aphasia sample (N = 57), with the 10 neurotypical controls included where applicable. Longitudinal and timepoint-averaged analyses were restricted to the 47 participants who met the narrative-length criterion and completed assessments at all six timepoints. The sample used for each analysis is specified in the corresponding section.

Narrative elicitation and transcription

Narrative discourse samples were collected at six time points: (1) baseline (pretreatment), (2) following the initial treatment phase, (3) post rest period, (4) after the second treatment phase, (5) one-month post-treatment, and (6) six months post-treatment. At each time point, participants were asked to retell the story of *Cinderella* using a wordless picture book to ensure standardized elicitation[20]. All audio samples were transcribed manually by trained graduate students in speech-language pathology following the AphasiaBank transcription protocol[21], although only textual transcription (without error coding) was used in this study. Transcripts were reviewed for accuracy and consistency, and all words were converted to lowercase to provide uniformity in input to the model.

Language model and feature extraction

We employed the pretrained GPT-2 small model (12 layers) to analyze the transcribed narratives. For each transcript, we extracted token-level hidden state activations from all 12 GPT-2 layers. To control for differences in transcript length, all narratives were truncated to a common sequence length corresponding to the minimum token count across the analyzed samples. Thus, the same number of tokens contributed to the extracted hidden-state representations for each transcript. Because this procedure was anchored to the shortest transcript, longer narratives were proportionally more truncated than shorter narratives. Consequently, the extracted representations reflect a length-matched initial portion of each narrative rather than the complete narrative in cases where truncation occurred. Prior to tokenization, AphasiaBank-style disfluency markers, including repetitions, false starts, and error codes, were removed to prevent annotation-specific conventions from contributing variation to the extracted representations. Consequently, the resulting activation measures reflect the linguistic content

retained in the normalized narratives rather than explicitly coded disfluency or error information. Token-level activations were then averaged within each layer to produce a 768-dimensional hidden-state vector per layer for each transcript. Finally, for statistical analysis and visualization, a scalar summary value was computed by averaging across the 768 hidden-state dimensions within each layer. Because GPT-2 hidden states contain both positive and negative values, this signed average may reflect partial cancellation across dimensions and should therefore be interpreted as an exploratory model-derived summary measure rather than a direct measure of activation magnitude or semantic processing.

This summary approach follows precedent in work linking transformer representations to language and neural data, which commonly rely on pooled or averaged embeddings as a tractable, model-derived index of representational engagement[22,7], rather than attempting to interpret individual hidden-state dimensions, which lack consistent semantic meaning across training runs or model checkpoints. The mean signed activation was adopted because it yields a single, layer-specific scalar suitable for group- and time-level comparisons, consistent with the exploratory aims of this study. We do not interpret this measure as a proxy for activation magnitude or a specific linguistic computation, but strictly as an index of overall model engagement at each layer.

Data analysis

We implemented a multitiered statistical framework via R[23]. Our analysis pipeline consisted of six steps:

Baseline group comparison of GPT-2 layer activations

To investigate whether baseline layer-wise hidden-state activations of the GPT-2 differ between individuals with aphasia and neurotypical controls, we extracted mean activation values from each of the 12 transformer layers for all participants for all Cinderella narrative discourse samples. This baseline comparison used the full N = 57 individuals with aphasia and N = 10 neurotypical controls. We conducted nonparametric Wilcoxon rank-sum tests for each layer to compare the activation distributions across the two groups. This method enabled us to ascertain whether layers in the GPT-2 model encapsulated group-specific linguistic representations at the baseline point.

Layer-wise activation differences across aphasia subtypes at baseline

To examine whether various subtypes of aphasia displayed unique representational profiles at baseline in the GPT-2, we performed a layer-wise analysis of hidden-state activations at baseline (Time Point 1). Kruskal–Wallis rank-sum tests were conducted for each of the 12 transformer layers to assess variations in the mean activation values across five aphasia subtypes: Broca's, Wernicke's, Conduction, Anomic, and Global aphasia, as defined with the WAB-R[24], and one control group. This analysis included the full aphasic sample (N = 57) and 10 neurotypical controls. Nonparametric tests were used because of the small sample sizes of several subtypes and the potential for non-normal activation distributions. For layers showing a significant omnibus group effect, Dunn's post hoc pairwise comparisons were performed with Benjamini–Hochberg correction for multiple comparisons.

Group-level activation profile

To examine representational patterns in aphasic speech, we conducted a group-level analysis of GPT-2 hidden-state activations. At each time point (1–6) and for each of the 12 model layers, we extracted token-level hidden states from the transcripts of 47 individuals with aphasia. These activations were first averaged within each transcript to obtain a transcript-level mean. Next, transcript-level means were averaged across participants, resulting in a 6 (time points) × 12 (layers) matrix of mean activation values. In addition, to statistically assess changes in activation over time and across model layers, we fitted a linear mixed-effects model with participant-level random intercepts:

$$\text{Activation} \sim \text{Layer} \times \text{Time} + (1|\text{Participant})$$

Layer activation patterns across aphasia subtypes

To examine variation in representational patterns across standard aphasia subtypes, as defined with the WAB-R, GPT-2 hidden-state activations were averaged across all six timepoints for each participant, yielding a single timepoint-averaged activation value per participant per layer and removing the repeated-measures structure. This analysis was restricted to the N = 47 aphasic participants retained across all six timepoints. Group differences were evaluated using separate one-way ANOVAs for each of the 12 layers, with aphasia subtype as the between-subjects factor and timepoint-averaged activation as the dependent variable. Significant omnibus effects were followed by Tukey's HSD pairwise comparisons. Each participant therefore contributed a single timepoint-averaged activation value per layer, removing the repeated-measures structure. Inspection of model residuals indicated no substantial violations of ANOVA assumptions; therefore, one-way ANOVA with Tukey's HSD was used for these comparisons. These analyses characterize subtype differences averaged across the longitudinal period and do not directly assess subtype-specific changes over time.

Longitudinal dynamics of GPT-2 layer activations over time

To examine how representational patterns changed across the six longitudinal time points, we fitted linear mixed-effects models for each GPT-2 layer, using activation as the dependent variable and time points as a categorical fixed effect. Aphasia subtypes were included as between-subject predictors, allowing us to test for interactions between subtypes and patterns of change over time. A random intercept for participants was included to account for within-subject repeated measures. This approach enabled us to assess longitudinal trends in activation while reducing the multiple comparison burden inherent in individual-level regressions. In addition, we used Wilcoxon signed-rank tests to compare time point 1 and time point 6 activation values separately for each of the 12 GPT-2 layers, assessing changes in layer–layer activation between Time Points 1 and 6. This nonparametric test was chosen because of nonnormal distributions and the presence of outliers in the GPT-2 activation data. Given the use of only two time points and the absence of assumptions regarding linearity or equal variances, the Wilcoxon tests provided a more robust alternative to regression. Effect sizes (r) were computed for each layer-specific comparison, and the results were subsequently corrected for multiple comparisons to ensure statistical validity.

Correlation with clinical language scores

We assessed the cross-sectional relationship between baseline GPT-2 layer activation and aphasia severity using Pearson correlations. For each of the 12 layers, baseline (Time Point 1) activation was correlated with baseline WAB-R AQ scores across the 57 participants, yielding one observation per participant per layer.

To control for alpha inflation resulting from the large number of layer-wise and group-level statistical tests conducted across analyses, all reported p values were adjusted via Benjamini–Hochberg false discovery rate (FDR) correction[25]. Specifically, FDR adjustment was applied across the 12 GPT-2 layers within each analysis. This approach ensures that significant effects reflect robust differences rather than chance findings due to multiple comparisons.

Model diagnostics were conducted for the linear mixed-effects models reported above. All models converged without warnings, and none showed singular fits. For the Layer × Time model, fixed effects accounted for a substantial proportion of variance (marginal $R^2 = 0.89$), with the participant-level random intercept contributing modestly beyond this (conditional $R^2 = 0.91$). Residual inspection revealed a departure from normality (Shapiro–Wilk $W = 0.70$, $p < .001$), driven primarily by heteroscedasticity across layers, as Layers 11–12 exhibited both larger mean activation and larger variance than earlier layers. Given the large, balanced sample underlying this

model and the magnitude of the reported effects, we consider the fixed-effects inferences reasonably robust to this violation, but we report it here rather than assume normality was met.

Results

Baseline group comparison of GPT-2 layer activations

Wilcoxon rank-sum comparisons of GPT-2 hidden-state activations between the aphasia and control groups at Time Point 1 indicated that although the overall activation patterns were similar (with deeper layers showing greater activation than surface layers), there were statistically significant changes across layers 1, 2, 5, 8, 10, 11, and 12 of the transformer. Specifically, activations were higher in layers 12, 11, and 8 in the aphasia group than in the control group, whereas they were lower in layers 10, 5, 2, and 1 (all adjusted p-values < 0.05). This suggested that, at baseline, layer-wise representational profiles for the Cinderella narrative differed between aphasic and neurotypical participants Fig. 1.

Layer-wise activation differences across aphasia subtypes at baseline

Statistically significant variations in GPT-2 activations were detected among all groups throughout all 12 layers (adjusted $p < .05$), with notably pronounced effects in layers 5, 9, 10, 11, and 12. Subtype sample sizes were: Broca's (n=21), Wernicke's (n=4), Conduction (n=11), Anomic (n=19), Global (n=2), and control (n=10). Post hoc analysis indicated that Broca's aphasia consistently differed from the Anomic, Conduction, and Global subtypes across multiple deeper layers. Furthermore, differences between Anomic and Conduction aphasia were noted in layers 1, 5, and 9. Compared with non-aphasic subjects, only the Broca, Anomic, and Conduction aphasia subtypes presented considerable abnormalities, particularly in layers 5, 9, 10, and 11.

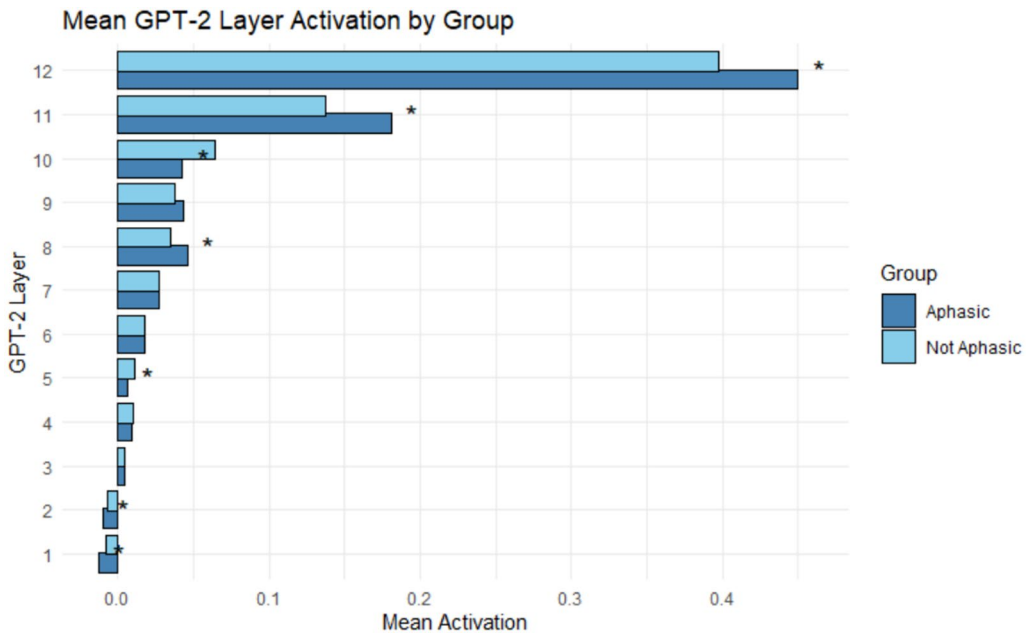


Fig. 1. Mean GPT-2 layer activation in aphasic and non-aphasic groups. Bar plot showing the average GPT-2 layer activation across 12 layers for individuals with aphasia (blue) and healthy controls (skyblue). Each bar represents the mean activation value for one group at each layer.

Group-level activation profile

Averaged GPT-2 hidden-state activations were calculated across all patients for each model layer (1–12) and each time point (time points 1–6). These values are visualized in Fig. 2 as a heatmap. Higher activation values were observed in deep layers across all time points. The overall layer-wise activation profile remained consistent throughout the six-month period. To statistically quantify these observations, we fitted a linear mixed-effects model with a Layer × Time interaction. The analysis revealed a significant main effect of Layer ($F(11, 3266) = 4847.8$, $p < 2\times10^{-16}$), indicating substantial differences in activation values across the 12 layers. There was also a significant main effect of time ($F(5, 3266) = 33.8$, $p < 2\times10^{-16}$) and a significant layer × time interaction ($F(55, 3266) = 18.9$, $p < 2\times10^{-16}$), suggesting that temporal activation trajectories differed by layer. Model diagnostics indicated clean convergence and no singular fit, though residuals showed heteroscedasticity across layers (see Methods). Post hoc analyses revealed that Layers 10, 11 and 12 had consistently greater activation and the greatest variability across time (Layer 12 = 0.0066, Layer 11 = 0.0049, Layer 10 = 0.0027).

Layer activation patterns across aphasia subtypes

To assess whether GPT-2 layer-wise activations varied by aphasia subtype, we conducted twelve one-way ANOVAs (one for each of the 12 layers) using aphasia subtype as the independent variable and mean activation values as the dependent variable. These analyses revealed significant main effects of subtype at all layers (all $p < 0.001$), indicating consistent subtype-related differences in activation values (Fig. 3). No time variable was included in these models, as timepoint-averaged activations were used for this analysis. Post hoc comparisons via Tukey's HSD test ($\alpha = 0.05$) revealed specific differences between subtypes, particularly between Broca's and Anomic

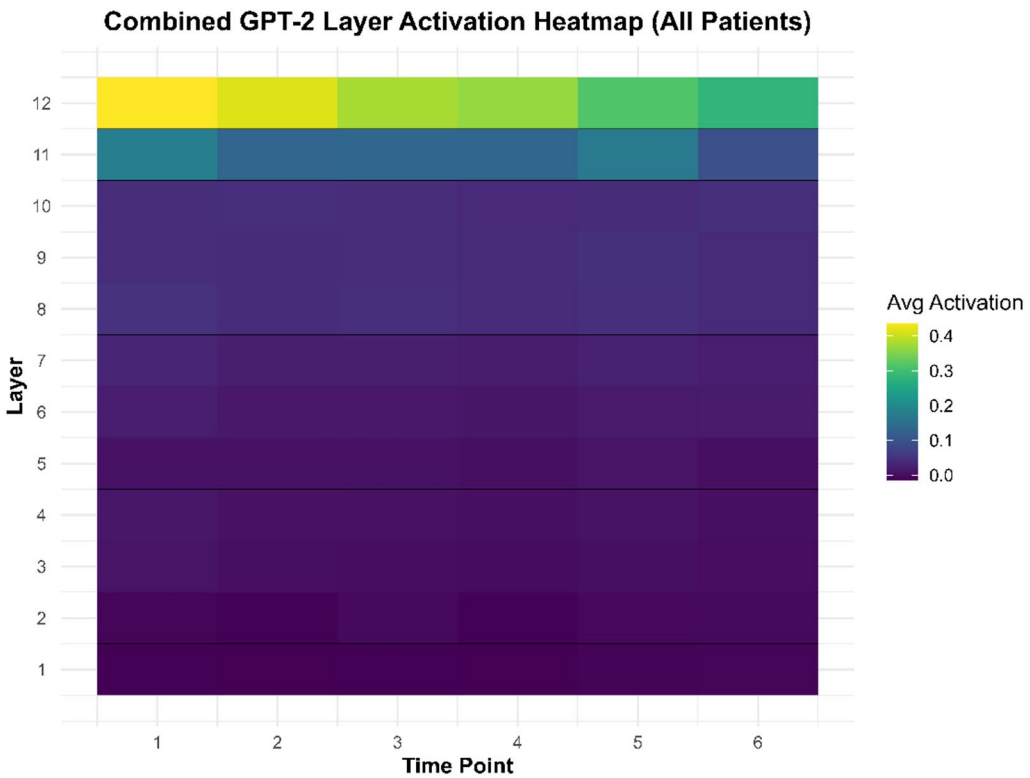


Fig. 2 Combined GPT-2 layer activation heatmap. Heatmap showing the average GPT-2 layer activation across all patients and time points. Each tile represents the mean hidden state activation for one layer (y-axis) during a specific time point (x-axis). Color intensity corresponds to activation value, with deeper layers showing consistently higher activation.

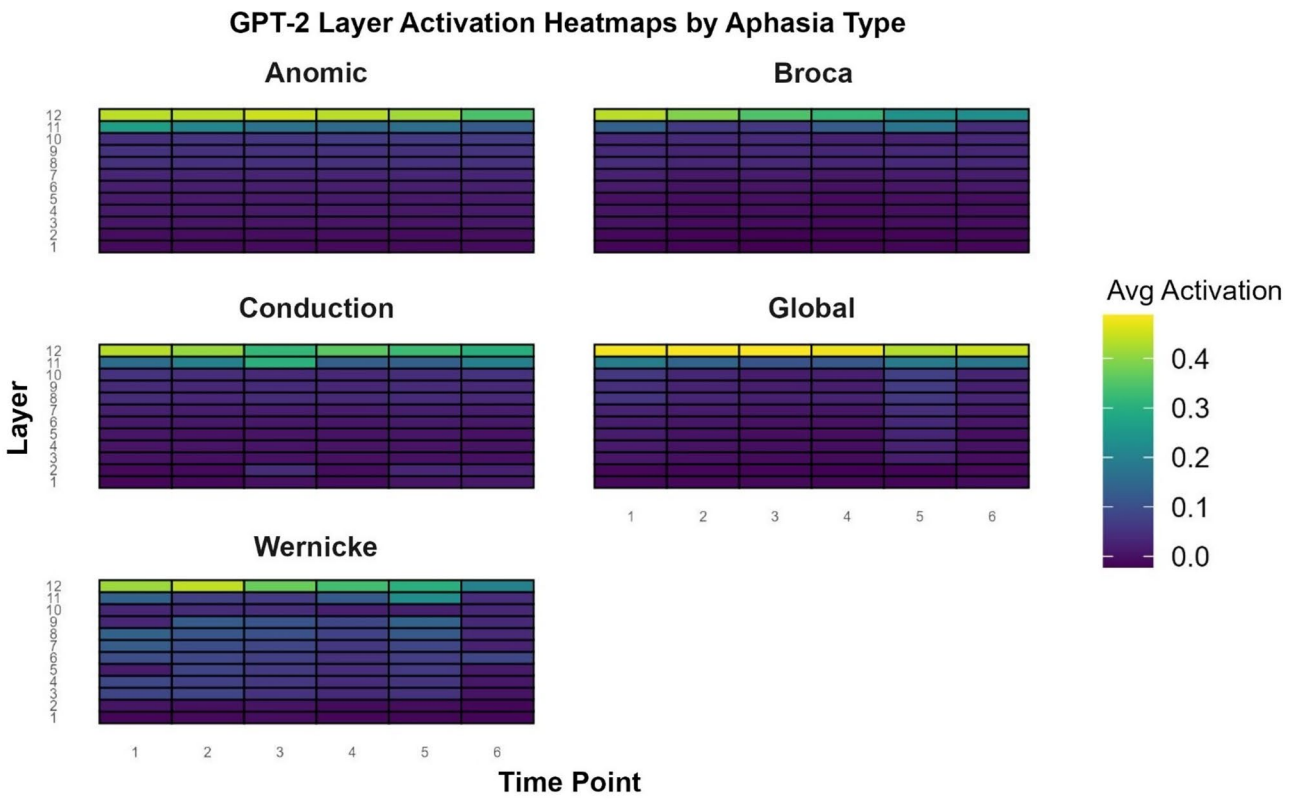


Fig. 3 GPT-2 layer activation heatmaps by aphasia subtype. Each panel shows average activation across time point (x-axis) and GPT-2 layers (y-axis) for one aphasia type. Color intensity represents mean activation. Different patterns across subtypes suggest major differences in representational engagement between groups.

aphasia and between Wernicke's and Broca's aphasia. A full list of comparisons and adjusted p values is provided in Appendix A Table A3.

Longitudinal dynamics of GPT-2 layer activations over time

To quantify where significant changes in layer activation occurred over time, we first fitted linear mixed-effects models for each of the 12 GPT-2 layers with time points (1-6) as fixed effects and participants as random effects. F tests on the time effect revealed that several deeper layers exhibited significant longitudinal modulation. These effects remained significant after False Discovery Rate (FDR) correction at $q < .05$, indicating that activation patterns in later layers changed systematically across the longitudinal study period.

To further characterize the magnitude and direction of these effects, we conducted Wilcoxon signed-rank tests to compare activation between the first and last time points across patients, regardless of direction. The individual trajectories exhibited both increases and decreases in activation; however, the most significant group-level changes were noted in the final three layers of GPT-2. Layer 12 demonstrated the greatest effect size ($r = 0.84$), followed by Layer 11 ($r = 0.67$) and Layer 10 ($r = 0.24$), with all the results achieving statistical significance after adjustment for multiple comparisons. Conversely, earlier layers (Layers 1–9) presented smaller and statistically nonsignificant median effect sizes, with r values predominantly less than 0.24. Figure 4 presents the effect sizes derived from the Wilcoxon tests. These results indicated the magnitude of change over time in model-derived activation measures. Individual participants showed both increases and decreases in deeper-layer activations; the effect sizes reflect group-level patterns of change rather than uniform directional shifts. The deeper layers

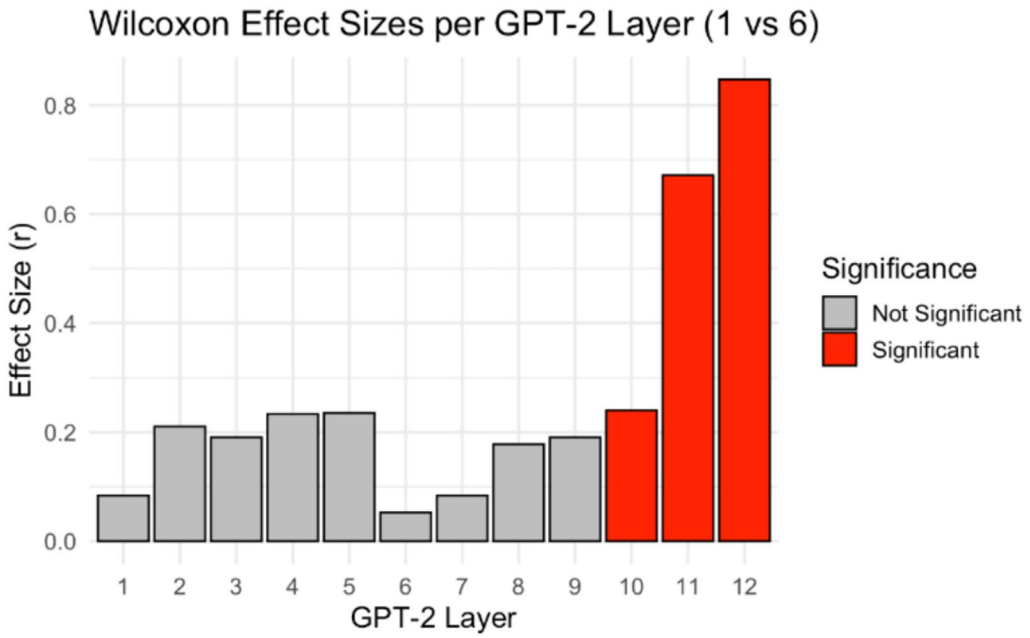


Fig. 4 Wilcoxon effect sizes per GPT-2 layer (1 vs 6). Wilcoxon effect sizes (r) for each GPT-2 layer comparing activations between Time Point 1 and Time Point 6. Bars are colored by significance after FDR correction. Deeper layers (10–12) showed the largest and statistically significant changes over time.

of GPT-2 showed statistically significant modulation of activation value during the longitudinal period, and the clinical significance of these changes remains to be established.

Correlation with clinical language scores

Layer-wise correlation analyses demonstrated significant positive associations between GPT-2 activation and WAB-R AQ scores across most layers, especially in the early and deep layers (e.g., Layer 2: $r = 0.52$, $p < 0.001$; Layer 10: $r = 0.51$, $p < 0.001$). These correlations reflect associations between model-derived activation measures and clinical severity scores. The direction and magnitude of these relationships should be interpreted cautiously, as higher activation does not necessarily indicate better language ability; activation value may reflect compensatory processing, input complexity, or other factors not directly related to clinical improvement. In Fig. 5, layers whose activation significantly correlates with aphasia severity are marked.

To further characterize the relationship between aphasia severity and GPT-2-derived activation values, we examined the continuous association between WAB-R AQ scores and layer-wise activation values across participants. Several layers, including both early and deeper layers, exhibited positive associations between activation values and WAB-R AQ scores. These findings indicate that associations with aphasia severity were distributed across multiple depths of the model rather than being restricted to deeper layers (Fig. 6).

Discussion

Overview of key findings

This study identified statistically significant differences in GPT-2 layer activations patterns between aphasic and neurotypical individuals across specific layers (1, 2, 5, 8, 10, 11, and 12). Aphasia subtypes as defined by the WAB-R showed distinct activation profiles, particularly in deeper layers. Activation values in deeper layers

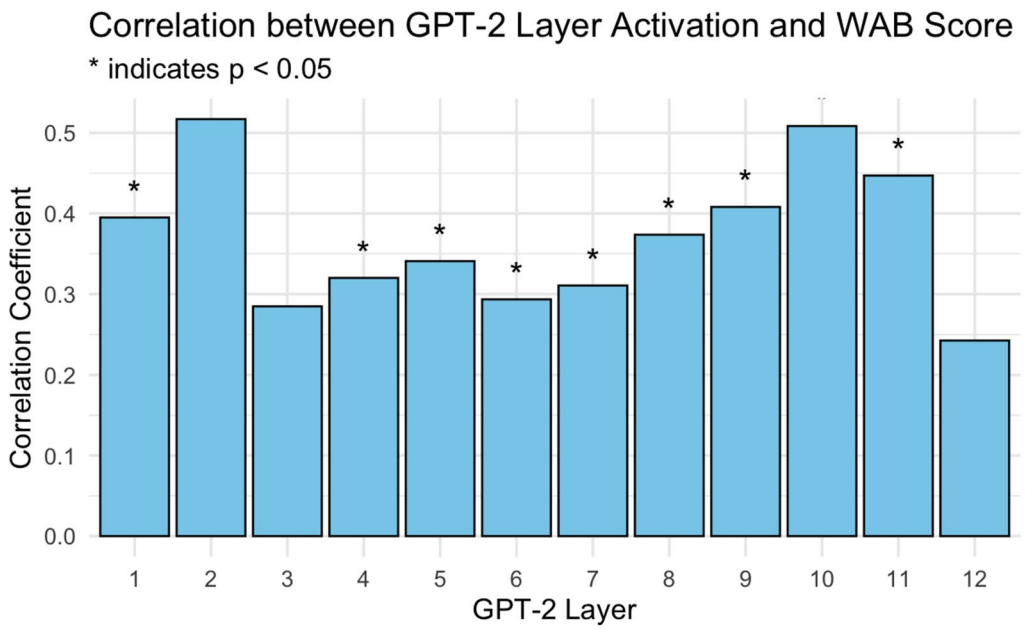


Fig. 5 Correlation between GPT-2 layer activation and WAB-R AQ score. The bar plot shows the Pearson correlation coefficients between mean activation in each GPT-2 layer and patients' WAB-R AQ scores. Asterisks (*) indicate statistically significant correlations at $p < 0.05$. Significant correlations were observed across both early (e.g., Layer 2) and deep (e.g., Layers 10-11) layers, indicating that activation values were significantly associated with aphasia severity at multiple depths of the model rather than in deep layers alone. Each bar represents the Pearson correlation coefficient for one GPT-2 layer, computed across participants at baseline.

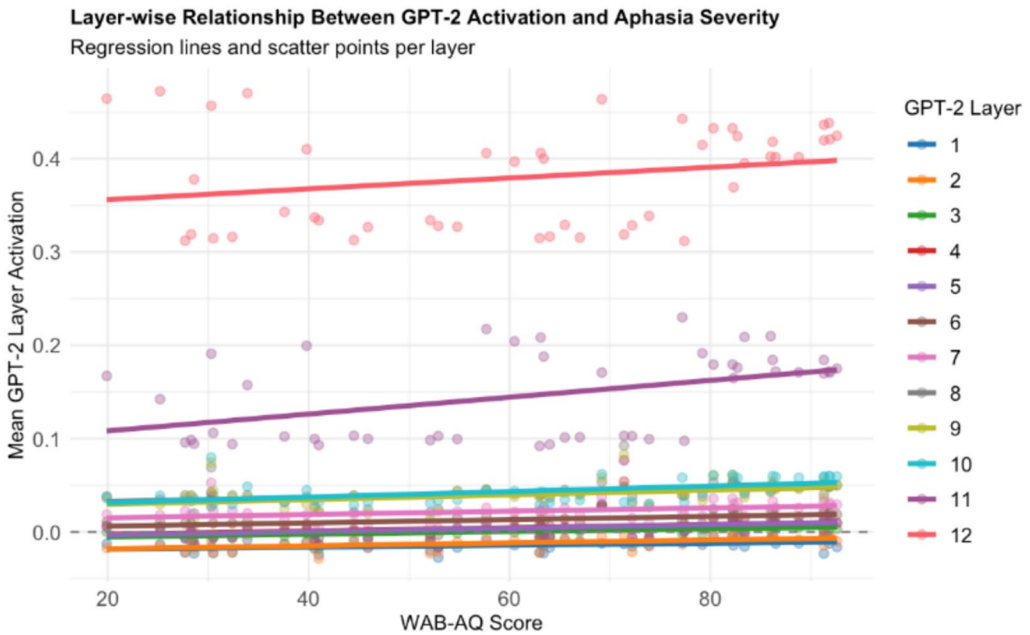


Fig. 6 Layer- wise relationship between GPT-2 activation and aphasia severity. Each point represents an individual participant. Colored regression lines illustrate layer-wise linear associations between GPT-2 activation values and WAB-R AQ scores. This visualization shows how activation values relate continuously to WAB-R AQ scores across participants, with the relationship for each layer quantified using Pearson's r.

(10–12) changed significantly from baseline to 6-month follow-up, with Layer 12 showing the largest effect size. Activation measures were positively associated with WAB-R AQ scores at multiple levels. These findings represent exploratory observations about how transformer language models process aphasic discourse and should not be interpreted as direct evidence of neural mechanisms or clinical improvement without further validation.

Layer activation in typical and aphasic speech

Our study revealed that the general pattern of layer activation in GPT-2 during speech processing is largely consistent between aphasic and neurotypical speakers, especially in the model's deeper layers. However, notable differences were observed in layers 1, 2, 5, 8, 10, 11, and 12, suggesting that, compared with neurotypical controls, aphasic subjects demonstrated modified representational depth at baseline. These differences in activation patterns suggest that the GPT-2 model processes aphasic narratives differently from typical speech. The functional significance of these differences, whether they reflect disrupted linguistic processing, reduced semantic abstraction, or other factors, remains speculative and requires direct validation through probing tasks or other interpretability methods. We caution against over-interpreting layer-wise differences as direct evidence of specific linguistic functions in aphasic populations, as the mapping between layer depth and linguistic function varies across models, tasks, and populations[14,13].

In narrative-discourse samples of people with aphasia, changes in deeper-layer hidden-state activations may reflect sensitivity to contextual and semantic properties of the input[22,7], although the present analyses do not directly establish the linguistic computations represented by these layers. Deeper-layer activation may reflect continued sensitivity to contextual or semantic properties of aphasic discourse,however, this interpretation is indirect and was not tested using layer-specific probing or ablation methods. Accordingly, activation differences in these layers should be interpreted as differences in model-derived representational states rather than direct evidence of a specific linguistic computation.

Subtype-specific patterns

Layer-wise analyses revealed that most significant longitudinal changes occurred in the deepest layers of GPT-2, particularly Layer 12. These layers showed consistent decreases in activation over time at the group level (except in Wernicke's aphasia), despite individual variability. As these layers have been suggested to be more sensitive to contextual and semantic information, although the precise functional role of individual layers remain uncertain, their decreased engagement may reflect changes in model-derived representations over time, although the functional interpretation of these activation changes remains uncertain.

Importantly, changes in the scalar layer-wise activation measure do not have a straightforward functional interpretation. Because this measure reflects the signed mean across hidden-state dimensions, higher or lower values cannot be directly equated with greater or reduced neural-like processing, computational effort, or processing efficiency. Rather, these changes indicate differences in the model-derived representational states associated with the narratives. Accordingly, functional interpretations of the observed activation patterns should be considered exploratory.

While all subtypes demonstrate a general trend toward greater activation in deeper GPT-2 layers, which is consistent with their role in semantic integration and discourse-level processing[22,13], pairwise statistical comparisons revealed distinct subtype-specific activation profiles across these layers.

The differences in layer-activation levels between different aphasic subtypes (by WAB-R classification) allow us to speculate on the possible nature of these patterns. Notably, as detailed in the results section, individuals with Broca's aphasia exhibited significantly greater activation in the deeper layers of GPT-2 (10–12) than other subtypes did. Individuals with Broca's aphasia exhibited significantly greater activation values in deeper GPT-2 layers (10–12) than several other subtypes. One possible interpretation is that these differences reflect sensitivity to linguistic information that remains available in syntactically sparse narratives. However, linking these layer-wise

differences specifically to semantic processing or reconstruction is speculative and was not directly tested in the present study. This interpretation should therefore be regarded as a hypothesis for future investigation rather than an established mechanism.

In contrast, Wernicke's aphasia exhibited markedly different activation profiles, with significantly lower deep-layer activation than Broca's aphasia and a more even distribution across the middle to deep layers. One possible, speculative interpretation is that this dispersed pattern relates to the fluent but semantically incoherent quality often described in Wernicke's aphasia; however, with n=4, this pattern should be treated as a hypothesis-generating observation rather than a confirmed subtype signature.

Anomic and conduction aphasia both showed a gradual decrease in activation in Layers 11–12, but conduction aphasia consistently exhibited greater activation than anomic aphasia across several layers. One possible interpretation is that these differences may be associated with phonological monitoring or discourse-level reformulation, processes previously implicated in conduction aphasia[26]. However, the present analyses do not directly measure these processes, and this interpretation should therefore be considered hypothesis-generating rather than mechanistically established.

Global aphasia showed an early deep-layer activation pattern; however, with only two participants in this subgroup, this observation is highly preliminary and should not be interpreted as characteristic of Global aphasia more broadly. Any functional interpretation here is speculative and offered only as a hypothesis for future study. Nonetheless, this activation exhibited minimal variation and only a slight decrease over time, indicating a relatively stable pattern in the model-derived activation measure across the longitudinal period.

The subtype-specific patterns indicate that GPT-2-derived representations appear sensitive to broad differences in aphasic discourse patterns while also identifying nuanced differences in the engagement of subtypes with deep layers over time. This pattern suggests transformer-derived measures may be a useful exploratory tool for characterizing discourse-level variation in aphasic speech, pending further validation. Because aphasia subtype and severity are not statistically independent in this sample (e.g., Global and Wernicke's aphasia are associated with lower WAB-R AQ scores), the subtype models reported here did not include severity as a covariate. Accordingly, the present analyses cannot determine the extent to which observed subtype effects reflect subtype-specific differences versus differences associated with overall aphasia severity.

In addition, our analyses identified statistically significant associations between GPT-2 activation measures and both aphasia subtype and WAB-R AQ scores. These exploratory associations suggest that transformer language models capture differences related to clinical presentation. However, validation of these measures as clinical tools would require prospective testing, comparison with established discourse variables, and demonstration of predictive utility in independent samples. Future research should systematically evaluate whether these model-derived measures provide clinical information beyond existing assessment methods.

Limitations and future directions

A major limitation of the present study is that, although GPT-2 has demonstrated alignment with brain data, the precise linguistic functions of its individual layers remain somewhat speculative. Our analyses depend on the current literature connecting deeper layers to discourse-level semantics; nonetheless, further controlled investigation is needed. The present analyses should therefore be interpreted primarily as exploratory representational analyses rather than clinically validated biomarkers. The speech transcripts did not include acoustic or prosodic data, which could be valuable in future multimodal methodologies. Future research should incorporate neurotypical comparison groups, broaden analysis to multilingual environments, and utilize layer probing approaches to more accurately correlate linguistic deficiencies with model internals. Future research should prioritize and integrate automated speech recognition techniques with large language models. In addition, aphasia subtype classification may partially reflect differences in severity, lesion burden, and chronicity. Therefore, subtype-related findings should be interpreted cautiously, particularly for smaller subtype groups. Additionally, we did not directly

control for or measure the effects of treatment intensity, responder status, or individual differences in therapy response. The longitudinal dataset is derived from an intervention study, yet treatment-related variables are not explicitly modeled. This limits our ability to attribute observed activation changes to recovery processes versus other time-related factors. Furthermore, transcripts were tokenized at the text level without acoustic or prosodic information, which may limit the generalizability of findings to clinical speech assessment that typically includes these dimensions. Relatedly, disfluency markers and error codes were removed during transcript normalization. Because these features can themselves be clinically informative with respect to aphasia subtype and severity, their removal means that the present model-derived measures do not capture explicitly coded disfluency- and error-related information. Future work should determine whether retaining or explicitly modeling these features alters the observed activation patterns. The use of a common sequence length required truncation of narratives longer than the shortest transcript. Consequently, information occurring later in longer narratives was not represented in the extracted hidden-state measures. Because narrative length may vary with language severity and fluency, this procedure could have differentially affected participants and may have contributed to observed subtype- or severity-related differences. Future studies should examine alternative approaches to controlling transcript length that preserve information from across the full narrative.

Conclusion

This research presents a new approach for the automated and objective analysis of aphasic speech via transformer-based language model representations. This method circumvents the need for manual linguistic coding by utilizing only textual transcripts, though it still relies on manually produced speech transcriptions, providing extensive analysis of discourse-level patterns in aphasia. This study suggests that GPT-2-derived representations may capture subtype- and severity-related variation in aphasic discourse as well as longitudinal changes in these model-derived representations. These findings provide preliminary evidence that transformer-derived summary measures may reflect variation relevant to clinical presentation; whether these measures have utility for monitoring progress in clinical practice remains untested and would require prospective validation. Additional research is necessary to elucidate the distinct functional roles of each model layer, which may facilitate targeted computational evaluations and tailored language interventions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-77518-1.

Author contributions S.A. contributed to the conceptualization of the study, data processing, computational analyses, interpretation of results, manuscript drafting, and revisions. J.F. contributed to study design, data collection supervision, clinical interpretation, and critical manuscript review. D.d.O. contributed to methodological guidance, interpretation of linguistic analyses, and substantive manuscript revisions. All authors reviewed and approved the final manuscript.

Funding National Institute on Deafness and Other Communication Disorders (DC014664; PI J. Fridriksson).

Data availability The dataset analyzed in this study consists of identifiable clinical speech recordings collected from individuals with post-stroke aphasia. Due to the presence of potentially identifiable personal and health-related information, the raw speech data cannot be made publicly available in order to protect participant privacy and confidentiality. Processed numerical activation matrices derived from the speech data, along with all analysis scripts used in this study, are available from the corresponding author (S. Saeed Ahmadi) upon reasonable request, subject to approval by the relevant ethics committee.

Declarations

Competing interests The authors declare no competing interests.

Ethics The study involved secondary analysis of clinical speech data and was approved by the Institutional Review Board of the University of South Carolina. All procedures were conducted in accordance with relevant ethical guidelines. Informed consent was obtained from all participants or their legally authorized representatives.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Brady, M. C., Kelly, H., Godwin, J., Enderby, P. & Campbell, P. Speech and language therapy for aphasia following stroke. *Cochrane Database Syst. Rev.***6**, CD000425 (2016).
2. Zhao, X. et al. Network-based statistics distinguish anomic and Broca aphasia. *ArXiv, arXiv***2302**(03250), v03252 (2023).
3. Goodglass, H. *Understanding Aphasia* (Academic Press, 1993).
4. Caucheteux, C., Gramfort, A. & King, J.-R. Deep language algorithms predict semantic comprehension from brain activity. *Sci. Rep.***12**(1), 16327 (2022).
5. Caucheteux, C. & King, J.-R. Brains and algorithms partially converge in natural language processing. *Commun. Biol.***5**(1), 134 (2022).
6. Gleichgerrcht, E. et al. Neural bases of elements of syntax during speech production in patients with aphasia. *Brain Lang.***222**, 105025 (2021).
7. Schrimpf, M. et al. The neural architecture of language: integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U. S. A.***118**(45), e2105646118 (2021).
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems30* (2017).
9. Yeh, C. et al. Attentionviz: a global view of transformer attention. *IEEE Trans. Vis. Comput. Graph.***30**(1), 262–272 (2023).
10. Vig, J., & Belinkov, Y. Analyzing the structure of attention in a transformer language model. *arXiv preprint* arXiv:1906.04284. (2019).
11. Geva, M., Schuster, R., Berant, J., & Levy, O. Transformer feed-forward layers are key-value memories. *arXiv preprint* arXiv:2012.14913 (2020).
12. Jawahar, G., Sagot, B., & Seddah, D. *What does BERT learn about the structure of language?* Association for Computational Linguistics (2019).
13. Tenney, I., Das, D., & Pavlick, E. BERT rediscovers the classical NLP pipeline. *arXiv preprint* arXiv:1905.05950. (2019).
14. Rogers, A., Kovaleva, O. & Rumshisky, A. A primer in BERTology: What we know about how BERT works. *Trans. Assoc. Comput. Linguist.***8**, 842–866 (2021).
15. Peters, M. E., Neumann, M., Zettlemoyer, L., & Yih, W.-t. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint* arXiv:1808.08949. (2018).
16. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. Improving language understanding by generative pretraining. (2018).
17. Belinkov, Y. & Glass, J. Analysis methods in neural language processing: a survey. *Trans. Assoc. Comput. Linguist.***7**, 49–72 (2019).
18. Cong, Y., LaCroix, A. N. & Lee, J. Clinical efficacy of pretrained large language models through the lens of aphasia. *Sci. Rep.***14**(1), 15573 (2024).
19. Kristinsson, S. et al. Predicting outcomes of language rehabilitation: prognostic factors for immediate and long-term outcomes after aphasia therapy. *J. Speech Lang. Hear. Res.***66**(3), 1068–1084 (2023).
20. Saffran, E. M., Berndt, R. S. & Schwartz, M. F. The quantitative analysis of agrammatic production: procedure and data. *Brain Lang.***37**(3), 440–479 (1989).
21. Macwhinney, B., Fromm, D., Forbes, M. & Holland, A. AphasiaBank: methods for studying discourse. *Aphasiology***25**(11), 1286–1307 (2011).
22. Caucheteux, C., Gramfort, A., & King, J.-R. Disentangling syntax and semantics in the brain with deep networks. International conference on machine learning (2021).
23. Team, R. C. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2025).
24. Kertesz, A. Western aphasia battery—revised (2007).

25. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc.: Series B (Methodological)***57**(1), 289–300 (1995).
26. Halai, A. D., Woollams, A. M. & Lambon Ralph, M. A. Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic poststroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex***86**, 275–289 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

S. Saeed Ahmadi[1] · Julius Fridriksson[1] · Dirk B.den Ouden[1,2]

✉ S. Saeed Ahmadi
ahmadis@email.sc.edu

[1] Department of Communication Sciences and Disorders, University of South Carolina, Columbia, SC, USA

[2] Department of Communication Sciences and Disorders, Chapman University, Irvine, CA, USA