Aphasic Neurons: Functional Specialization in Large Language Models

Anonymous ACL submission

Abstract

This study investigates the internal representations of large language models (LLMs) through the lens of aphasia research. By introducing *Aphasic Neurons* and analyzing their layer-wise distributions and overlap across linguistic error types and five aphasia subtypes in three models, we provide empirical evidence that LLMs represent different aphasia subtypes and linguistic error types through largely distinct sets of neurons organized in functionally separable modules. These findings demonstrate that the analytical framework developed through aphasia research offers a novel perspective for investigating the modular organization of LLMs.

1 Introduction

Large language models (LLMs) have increasingly shown the ability to replicate human-like patterns in linguistic behavior (Cai et al., 2024), and it has been suggested that their internal representations possess a degree of modular organization (Qiu et al., 2024; AlKhamissi et al., 2025; Wang et al., 2025). However, the nature of this modularity remains largely unclear. We argue that insights from aphasia research, which has long illuminated the modular organization of human language processing, offer a valuable framework for investigating the internal representations of LLMs.

To begin with, aphasia refers to an acquired language disorder resulting from brain damage, in which previously acquired linguistic abilities are selectively impaired (Benson and Ardila, 1996). Crucially, aphasia symptoms vary considerably between individuals, and depending on the area of brain damage, patients may present with distinct profiles of linguistic impairment (Damasio, 1992). These profiles have been systematically classified into subtypes, including well-known types such as Broca's aphasia and Wernicke's aphasia, using standardized clinical assessments (Goodglass and Kaplan, 1983; Kertesz, 1982). Aphasia research has long served as a crucial window into the modular organization of human language processing (Geschwind, 1965; Dronkers et al., 2017), as the selective dissociation of linguistic abilities across these subtypes provides compelling evidence that language processing is organized in functionally separable modules (Hickok and Poeppel, 2000; Pylkkänen, 2019; Dronkers and Ivanova, 2023). Motivated by these insights, Wang et al. (2025) demonstrated that ablating specific layers and attention heads in LLMs can reproduce language production impairments that resemble different aphasia subtypes, suggesting that the modularity observed in aphasic symptoms may have computational parallels within LLMs. However, these studies focus on reproducing output-level behaviors, and it remains unexplored how such modularity is reflected in the internal representations of LLMs.

In this study, we examine whether LLMs internally represent aphasic utterances in subtype- and error-specific patterns, that is, whether distinct sets of neurons respond selectively to different aphasia subtypes and linguistic error types. This approach applies the logic of functional dissociability from aphasia research, whereby selective impairment of specific linguistic abilities across subtypes reveals the modular organization of language processing, to the analysis of LLM internal representations. To investigate these patterns at the neuronal level, we build on the concept of *expert units* introduced by Suau et al. (2020). We employ Average Precision (AP) and Jaccard similarity to evaluate the specificity and independence of these neurons (see Section 3.1 for details). We further examine their existence, layer-wise distributions, and degree of overlap across subtypes, providing novel empirical evidence on how modular organization of language processing is reflected in the internal representations of LLMs. Beyond this empirical finding, we demonstrate that the analytical framework developed through aphasia research offers a novel per-

spective for investigating the modular organization of LLMs.

2 Related work

2.1 Aphasia subtype

Aphasia presents with various patterns of language impairment depending on the location of brain damage, and systematic differences between subtypes have long served as evidence for the modular organization of language processing(Geschwind, 1965; Dronkers et al., 2017). Table 1 summarizes the five subtypes examined in this study.

Broca's aphasia and Wernicke's aphasia represent the most widely studied dissociation in aphasia research (Broca, 1861; Wernicke, 1874; Lichtheim, 1885; Berker et al., 1986): the former is characterized by non-fluent speech with relatively preserved comprehension, while the latter presents with fluent but meaningless speech along with severely impaired comprehension. Beyond this broad distinction, subtypes have been further differentiated by additional dimensions such as repetition ability and naming. Among fluent subtypes, conduction aphasia is characterized by severely impaired repetition despite preserved fluency and comprehension, while anomic aphasia presents with selective difficulty in word retrieval and naming with otherwise preserved language functions (Goodglass and Kaplan, 1983; Le et al., 2024). On the non-fluent side, transcortical motor (transmotor) aphasia is notable for its preserved repetition despite non-fluent speech, with the primary difficulty in initiating spontaneous speech (Le et al., 2024). These selective patterns of impairment and preservation instantiate the logic of double dissociation (Teuber, 1955; Fodor, 1983; Shallice, 1988), which has served as a key tool in neuropsychological research for inferring the modular organization of cognitive systems.

2.2 Computational Approaches to Aphasic Research

Recent studies have begun to examine whether LLMs can capture the linguistic characteristics of aphasic speech. Cong et al. (2024a) demonstrated that surprisal derived from pre-trained LLMs can serve as an effective index to detect aphasia, assess severity, and classify subtypes. Subsequently, Cong et al. (2024b) showed that such metrics are generalized to English and Mandarin Chinese speakers. Similarly, Rezaii et al. (2024) showed that LLMs can classify subtypes of primary progressive aphasia with 78% precision from short speech samples. These findings suggest that LLMs are sensitive to the characteristic linguistic features of aphasic speech.

Aphasia Type	Key Characteristics
Broca's	Difficulty producing fluent speech with relatively preserved ability to understand language; struggles with repeating words and saying the names of objects (naming)
Wernicke's	Produces fluent but meaningless speech; severely impaired ability to understand, repeat, and naming
Conduction	Fluent speech with good comprehension, but severe difficulty repeating words; frequent sound-level errors in speech
Anomic	Fluent speech with preserved comprehension and repetition, but marked difficulty finding words and naming
Transmotor	Difficulty starting speech spontaneously (initiating spontaneous speech), but able to repeat and understand; non-fluent output

Table 1: Characteristics of five aphasia subtypes analyzed in this study. Descriptions are based on Goodglass and Kaplan (1983) and Le et al. (2024).

Taking a step further into the model itself, Wang et al. (2025) demonstrated that ablating specific layers and attention heads in LLMs can reproduce language production impairments that resemble different aphasia subtypes. This suggests that the modularity observed in aphasic symptoms may have computational parallels within LLMs.

Although these studies have provided novel information on how LLMs can contribute to aphasia research, the question of how LLMs internally represent and process the linguistic features of aphasic speech has not been directly addressed. In other words, existing approaches remain at the behavioral or output level. Meanwhile, research on LLM interpretability has shown that individual neurons exhibit functional specialization for specific linguistic concepts or modules (Durrani et al., 2020; Suau et al., 2020; Dai et al., 2022). Accordingly, we apply the logic of functional dissociability established in aphasia research to the analysis of internal representations of LLMs, advancing our understanding of how language processing is organized within LLMs.

3 Experiments

3.1 Methods: How to detect "*Aphasic Neurons*"

We follow the approach of Suau et al. (2020) to identify neurons specialized in capturing the speech patterns of individuals with aphasia, which we term

Aphasic Neurons. They proposed a method that treats each neuron within a pre-trained Transformer model as a binary classifier and quantifies its specialization using Average Precision (AP). AP is a widely used evaluation metric in information retrieval and machine learning, defined as the area under the Precision–Recall curve (Davis and Goadrich, 2006; Manning et al., 2008). AP integrates precision and recall across different classification thresholds. Specifically, sentences are first ranked by their prediction scores, and the threshold is swept from the highest to the lowest score, computing precision and recall at each threshold. In other words, AP is computed as the mean precision over recall values from 0 to 1.

In this study, we denote the responses of neuron m to all sentences associated with concept c (totaling $N_c = N_c^+ + N_c^-$ sentences) as

$$\mathbf{u}_{mc} \in R^{N_c},$$

and the corresponding binary labels as

$$\mathbf{b}_c \in \{0, 1\}^{N_c}.$$

We target the gate_proj layer for neuron-level analysis, as it plays a central role in selective information processing in gated activation functions (Shazeer, 2020). For each input sentence, we extract the output of the gate_proj layer in each Transformer block using a forward hook, excluding PAD token positions, and apply max pooling across the token dimension to obtain a single scalar activation value per neuron per sentence.

The activation values in $\mathbf{u}_{mc}$ are used directly as prediction scores without any preprocessing or transformation. By treating each neuron as a binary classifier and using $\mathbf{u}_{mc}$ as prediction scores, we compute AP as follows:

$$\mathrm{AP}_{mc} = \mathrm{AP}(\mathbf{u}_{mc}, \mathbf{b}_c) \in [0, 1]. \quad (1)$$

Thus, a neuron with a high AP value is capable of accurately distinguishing between sentences that contain a particular linguistic feature of an aphasic speaker and those that do not. Following Suau et al. (2020), who defined *expert units* as the top 1% of neurons with the highest AP values for a given concept, we identified the top 1% neurons with the highest AP values for each target concept as *Aphasic Neurons*.

3.2 Experimental Setup

3.2.1 Datasets

In this study, we utilized a corpus of English speech produced by individuals with aphasia, sourced from AphasiaBank (MacWhinney et al., 2011). AphasiaBank is a large-scale database of aphasic utterances annotated in the CHAT format. It contains detailed information for each utterance, including the speaker's gender, age, diagnosed aphasia subtype, error types, and accompanying gestures and fillers (Forbes et al., 2012).

The aphasia subtypes analyzed were restricted to five categories for which sufficient sample sizes were available: anomic, Broca's, conduction, transmotor, and Wernicke's. Global aphasia and isolation aphasia, which had only a few data points, were excluded from the analysis. Speakers whose aphasia diagnosis had not been definitively confirmed (notaphasicbywab) were likewise excluded. In addition to these speaker-level exclusions, shorter utterances can introduce noise into analysis. Therefore, we excluded sentences with fewer than ten words from the analysis.

3.2.2 Models

We conducted two distinct experiments. In Experiment 1, we used three large language models of comparable parameter counts, namely Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Gemma2-9B (Gemma Team et al., 2024). Based on these results, Experiment 2 focused on the model with the highest performance and conducted further analyses using it. All experiments were conducted with a maximum input sequence length of 128 tokens and a batch size of 4.

3.2.3 Evaluation Metrics

Average Precision (AP) Score The AP score is used to quantify the degree to which individual neurons discriminate speech in aphasia. Details are provided in Section 3.1.

Jaccard Similarity Jaccard similarity measures the overlap between two sets of neurons and is defined as follows:

$$\mathrm{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Here, A and B denote the sets of *Aphasic Neurons* identified for two different aphasia subtypes or linguistic error types, each represented as a set

of (layer index, unit index) pairs. A higher Jaccard similarity value means greater overlap.

Exclusivity Rate The exclusivity rate is the proportion of *Aphasic Neurons* linked to a specific type that are not shared with any other type. A higher exclusivity rate indicates that a more distinctive, type-specific neuron set was formed for the studied category (Suau et al., 2020).

3.3 Experiment 1: Layer-wise Distribution of Error Type-specific Neurons

In Experiment 1, we sought to detect *Aphasic Neurons* that could discriminate among three broad linguistic error types in aphasic speech: phonological, semantic, and morphological. We also investigated how distinct or overlapping these types are in their neural representations. Three open-source models with similar parameter sizes were used in the experiment. Applying the methodology in section 3.1, we detected *Aphasic Neurons* for each error type: phonological, semantic, and morphological. For each type, we used a binary classification framework. For example, utterances with phonological errors were treated as positive examples, while those with other error types were negative. Neurons strongly activated in this setting were called *Aphasic Neurons*. Model performance was evaluated in two ways: (1) the Top-K mean AP score and (2) the mean Jaccard similarity across error type pairs. Models with high AP scores and low inter-type Jaccard similarity were considered to accurately and independently capture each error type.

3.4 Experiment 2: Characteristics of Aphasia Subtype-specific Neurons

In Experiment 2, building on the results of Experiment 1, we identified *Aphasic Neurons* for speech from five aphasia subtypes—Broca's aphasia, Wernicke's aphasia, conduction aphasia, anomic aphasia, and transmotor aphasia—and analyzed the independence of neurons among subtypes and their layer-wise distribution. To ensure fair comparisons between subtypes, the number of positive and negative examples was balanced; the number of positive samples was unified to match that of transmotor aphasia, which had the fewest data points (1,006 samples under the condition min_words=10). Negative examples were sampled equally from the remaining four aphasia subtypes, excluding the positive subtype. As in Experiment 1, AP scores and Jaccard coefficients were used to assess the degree of overlap in *Aphasic Neurons* among the different subtypes.

4 Results

4.1 Experiments 1: Layer-wise Distribution of Error Type-specific Neurons

As shown in Table 2, Llama achieved the highest Top-K mean AP score (0.7785) among the three models. It also exhibited the lowest mean Jaccard similarity across error types (0.2201). Gemma attained an AP score nearly as high as Llama's (0.7782), but its inter-type Jaccard similarity was the highest (0.2561). This suggests that Gemma failed to represent the three linguistic error types as sufficiently independent neuron sets (Figure 1). Mistral yielded intermediate values on both metrics (AP: 0.7672; Jaccard: 0.2207).

Based on these results, we determined that Llama-3.1-8B-Instruct was the most capable model in terms of both classification accuracy and inter-type separability and accordingly adopted Llama for all subsequent analyses in Experiment 2.

We further examined the layer-wise distribution of *Aphasic Neurons* for each error type in Llama-3.1-8B-Instruct, as illustrated in Figure 2. The neurons responsive to semantic errors were broadly distributed across shallow layers (around layer 0) through intermediate layers. In contrast, neurons associated with morphological errors exhibited a pronounced concentration in the final layer (layer 31), showing a markedly distinct distributional pattern relative to the other two types of errors. Neurons responsive to phonological errors, meanwhile, presented an intermediate distributional profile: like semantic errors, they were distributed across middle layers, yet like morphological errors, they also showed notable concentration in the final layer.

4.2 Experiment 2: Characteristics of Aphasia Subtype-specific Neurons

Using Llama-3.1-8B-Instruct, we detected *Aphasic Neurons* for each of the five aphasia subtypes — anomic, Broca's, conduction, transmotor, and Wernicke's — under the min_words = 10 condition. The AP scores for each subtype are presented in Table 3.

All five aphasia subtypes yielded mean AP scores above the random baseline of 0.500. Broca's aphasia showed the highest mean AP (0.646), fol-

Model	Sym.	Mean AP
Llama 3.1-8B	Phono.	**.784**
	Sem.	.758
	Morph.	**.794**
Mistral 7B	Phono.	.767
	Sem.	.751
	Morph.	.784
Gemma 2-9B	Phono.	.783
	Sem.	.759
	Morph.	.793

Table 2: *Aphasic Neurons* detection performance (min_words= 10, top-k = 1%). AP denotes mean average precision; Bold marks the best score per metric.

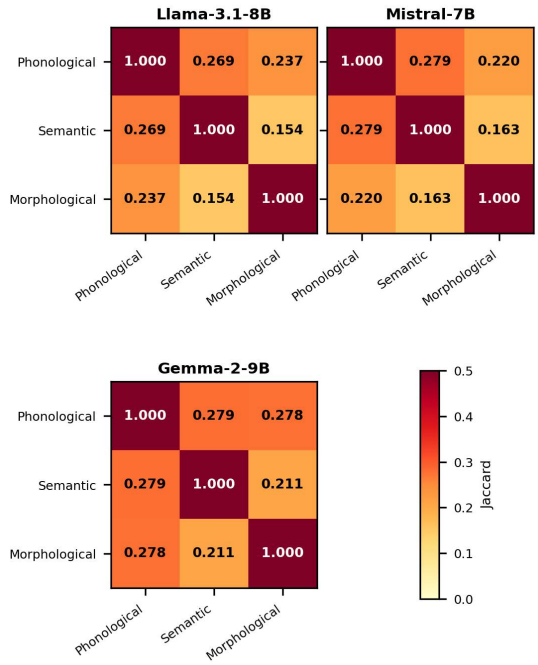

Figure 1: Pairwise Jaccard similarity across phonological, semantic, and morphological symptoms. Llama-3.1-8B is used in Experiments 2.

Aphasia Type	Mean AP	Max AP
Anomic	0.590	0.660
Broca's	0.646	0.727
Conduction	0.561	0.601
Transmotor	0.592	0.659
Wernicke's	0.550	0.600
baseline	0.500	—

Table 3: Average Precision (AP) scores for *Aphasic Neurons* per aphasia subtype (min_words = 10). Positive and negative samples were balanced to match the smallest subtype (transmotor aphasia; n = 1,006), with negative samples drawn equally from the remaining four subtypes. Mean AP is computed over the top-k = 4,588 neurons; Max AP is the highest AP score among them.

lowed by transmotor (0.592), anomic (0.590), conduction (0.561), and Wernicke's (0.550). Broca's aphasia also exhibited the highest Max AP of 0.727,

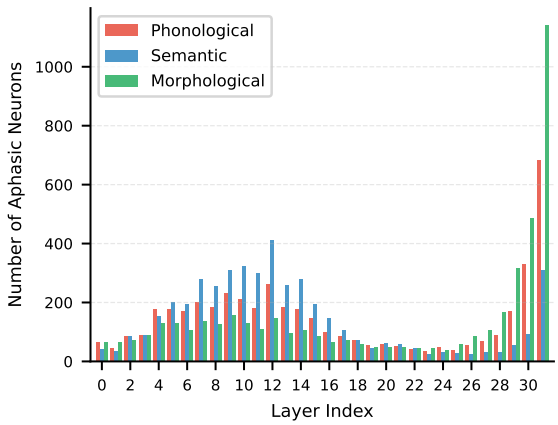

Figure 2: Layer-wise distribution of *Aphasic Neurons* across phonological, semantic, and morphological language impairment conditions.

suggesting that certain individual neurons achieve notably high classification accuracy for this subtype. The variation in mean AP scores between subtypes indicates that some subtypes were captured more than others, with Broca's aphasia showing the most prominent neuron-level specialization.

In addition, we analyze the degree to which the neurons identified for each aphasia subtype overlap with those of other subtypes. As illustrated in Figure 3, the vast majority of neurons for each subtype were exclusive to that subtype alone. The exclusivity rates, in descending order, were as follows: Broca's (100.0%), transmotor (95.7%), anomic (90.3%), Wernicke's (87.7%), and conduction (82.7%). Notably, Broca's aphasia showed no overlap with any other subtype. Based on these observations, Figure 4 presents the results of a Jaccard similarity-based evaluation of inter-subtype neuron overlap. Consistent with the high exclusivity rates reported above, overall overlap between subtypes was low; nevertheless, the conduction–Wernicke pair exhibited markedly greater overlap than any other combination (intersection: 554 neurons; Jaccard: 0.0643). The anomic–conduction pair showed the second highest overlap (intersection: 239 neurons; Jaccard: 0.0267), followed by anomic–transmotor (intersection: 197 neurons; Jaccard: 0.0219). In contrast, Broca's aphasia showed no overlap with any other subtype, demonstrating that it features a specialized group of neurons. All remaining subtype pairs yielded Jaccard values below 0.005.

An examination of the layer-wise distribution of the 554 neurons shared between the conduction

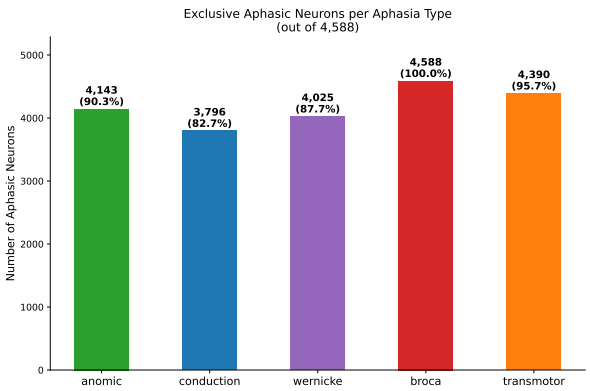

Figure 3: Exclusive *Aphasic Neurons* per Aphasia Subtype. Each bar shows the number of neurons that belong exclusively to that subtype (i.e., not shared with any other subtype) out of the top-k = 4,588 neurons selected per type.

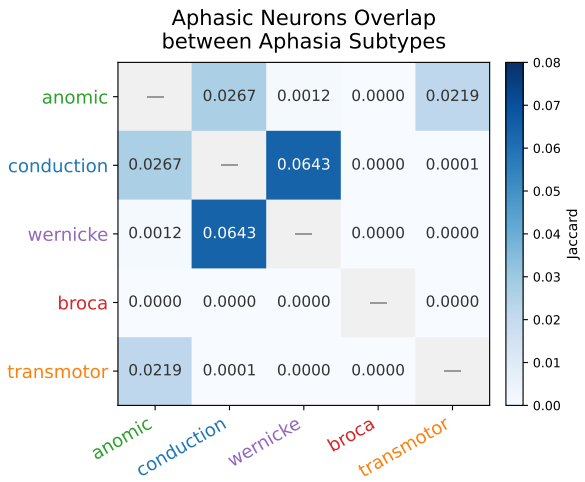

Figure 4: Pairwise Jaccard Similarity of *Aphasic Neurons* across Aphasia Subtypes. Higher values indicate greater overlap between the two subtypes' neuron sets. The conduction–wernicke pair shows the highest overlap (0.0643), while anomic shows relatively low overlap with all other subtypes.

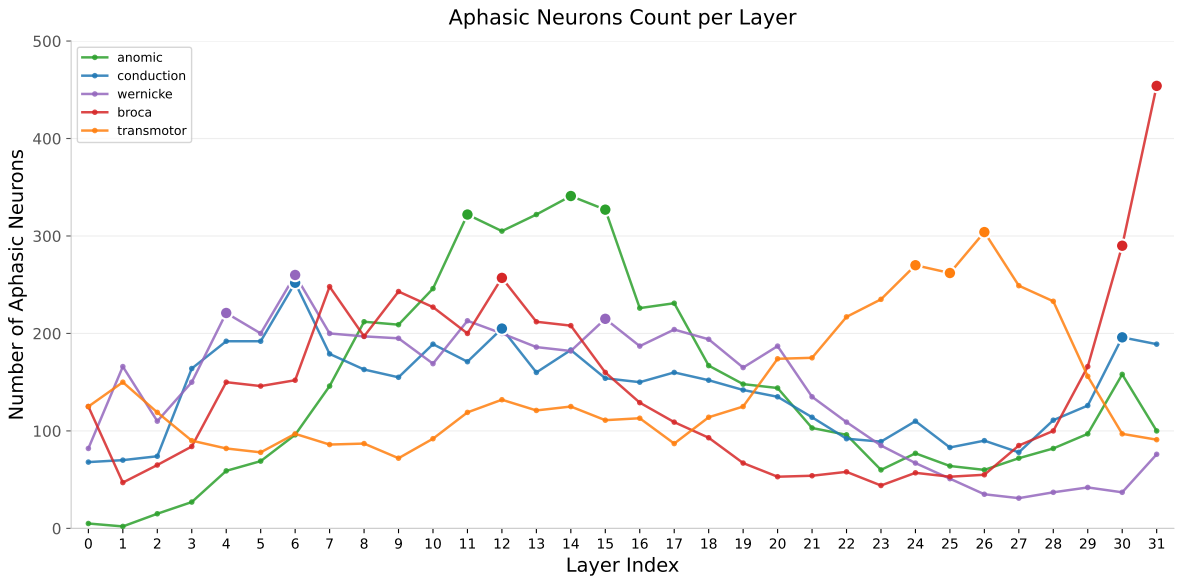

Figure 5: Layer-wise Distribution of *Aphasic Neurons* in Llama-3.1-8B-Instruct. Larger markers indicate the top 3 layers with the highest neuron counts per subtype.

and Wernicke subtypes revealed a wide distribution across layers, with a relative concentration in the early-to-middle layers 5–11, accounting for approximately 37.7% of overlapping neurons, followed by the middle layers 12–19 at 31.6%. The layer-wise distributions of the neurons for each aphasia subtype are shown in Figure 5. Anomic aphasia showed a characteristic distributional profile with a peak in the middle layers, particularly around layers 11–15. Broca's aphasia exhibited a pronounced peak at the final layer (layer 31), with additional concentrations at layers 30 and 12. Conduction aphasia showed its highest concentration at layer 6, followed by layers 12 and 30. Transmotor aphasia showed a concentration in the later layers, particularly layers 24–26. Wernicke's aphasia showed peaks in layers 6 and 4, with an additional elevation around layer 15, resulting in a comparatively distributed profile compared with the other subtypes.

5 Discussion

The results of Experiment 1 demonstrate that *Aphasic Neurons* were identified across all three models, suggesting that functionally specialized neurons for aphasic speech patterns exist as a generalizable phenomenon that is not specific to a particular model architecture. Although the overall pattern was consistent, some variation was observed in the layer-wise distribution between models. As shown in Appendix A, the distribution in Gemma is concentrated in the middle layers, in contrast to the pronounced peaks in the final layers observed in Llama and Mistral, possibly reflecting Gemma's distinct architectural properties, including its use of GeGLU (Shazeer, 2020) and its greater number of layers.

An examination of the layer-wise distribution of error-type-specific neurons in the selected model revealed patterns that may reflect how linguistic processing is organized across layers. Among the three types of errors, morphological errors showed the clearest distribution pattern, with a pronounced concentration in the final layers (layer 31: 1,142 neurons; layer 30: 485 neurons), consistent with the interpretation that morphological processing occurs in layers immediately preceding output generation. The distributions for phonological and semantic errors were less clearly differentiated, although some tendencies were observed: phonological errors showed concentration in the final layers alongside an additional cluster around layer 12, while semantic errors peaked in the intermediate layers. While prior research suggests that surface-level representations are captured in earlier layers and semantic processing emerges in deeper layers (Jawahar et al., 2019; Tenney et al., 2019; Rogers et al., 2020), the present findings only partially align with this pattern, suggesting that the layer-wise organization of linguistic processing may differ when input involves linguistically atypical features such as aphasic speech. These layer-wise distributional patterns suggest that linguistic processing within LLMs is functionally differentiated across layers, with phonological, semantic, and morphological error types engaging distinct regions of the network rather than being uniformly distributed.

In Experiment 2, which focused on aphasia subtypes, the exclusivity rate exceeded 82% across all subtypes, and the neurons for each aphasia type were distributed independently of each other. This finding suggests that the LLM represents each subtype of aphasia through a distinct set of neurons.

While the majority of neurons were exclusive to each subtype, some degree of overlap between subtypes was also observed. Most notably, the conduction–Wernicke pair showed markedly greater overlap than any other combination (Jaccard: 0.0643; intersection: 554 neurons), suggesting that these two subtypes share, at least in part, common internal representations within the LLM. The 554 overlapping neurons were concentrated in the early-to-middle layers (layers 5–12), accounting for approximately 37.7% of the shared neurons, a distributional pattern consistent with the intermediate-layer peaks observed for semantic errors in Experiment 1 (layers 9, 10, and 12). To further examine the linguistic characteristics these shared representations reflect, we calculated the proportion of phonological, semantic, and morphological errors for each aphasia subtype in the AphasiaBank corpus used in this study. The results revealed that conduction aphasia exhibited a semantic error rate of 35.9%, the second highest among all subtypes after Wernicke's aphasia. Although conduction aphasia is typically characterized by repetition deficits and phonological errors, this finding suggests that the shared internal representations between the two subtypes may reflect common semantic error patterns encoded within the LLM. Taken together, these findings suggest that the internal representations of LLMs are organized in a way that reflects not only broad categorical distinctions but also finer-grained shared structures across functionally related linguistic patterns.

In contrast, Broca's aphasia exhibited a 100% exclusivity rate alongside the highest Max AP among all subtypes (0.727), suggesting that it occupies an entirely unique representational space within the LLM. This complete absence of neuron-level overlap with other subtypes is consistent with the markedly distinct linguistic profile of Broca's aphasia, which showed the highest phonological error rate (45.0%) and overall error rate (0.459 errors/utterance) among all subtypes.

A coherent picture emerges when the layer-wise distributions of Experiments 1 and 2 are considered together. The subtype-specific distributions observed in Experiment 2 show a natural correspondence with the error-type distributions observed in Experiment 1, suggesting that the internal representations of LLMs are organized in a hierarchically structured manner that reflects the linguistic char-

acteristics of each subtype. For instance, Broca's aphasia, characterized by the highest phonological error rate, showed a pronounced peak at the final layer, consistent with the phonological error distribution in Experiment 1. Similarly, anomic aphasia, primarily characterized by word-retrieval difficulty, showed its neurons peaking in the middle layers, consistent with the semantic error distribution. Notably, the phonological errors of Broca's and conduction aphasia have been argued to differ in character, with Broca's aphasia marked by sound substitutions and conduction aphasia additionally characterized by transpositions (Monoi et al., 1983; Bastiaanse et al., 1994), and their distinct layer-wise profiles may reflect these fine-grained differences in error type.

These findings, considered across Experiments 1 and 2, suggest that the internal representations of LLMs are organized in functionally separable modules that correspond to distinct linguistic error types and aphasia subtypes. The identification of *Aphasic Neurons* and their layer-wise distributions reveals that different aspects of linguistic processing are encoded in different parts of the network, with semantic processing reflected in intermediate layers and phonological and morphological processing concentrated in deeper layers. Furthermore, the patterns of overlap and exclusivity observed across subtypes suggest that LLM internal representations exhibit a form of modularity that is functionally analogous to the dissociability of linguistic abilities documented in aphasia research. More broadly, the analytical logic developed in aphasia research offers a productive lens through which to examine the internal organization of LLMs, opening new avenues for interpretability research that go beyond behavioral or output-level analyses.

6 Conclusion

This study aimed to advance the interpretability of LLMs by applying the analytical logic developed in aphasia research to the analysis of their internal representations. To this end, we introduced *Aphasic Neurons* and analyzed their layer-wise distributions and overlap across linguistic error types and aphasia subtypes, offering a new perspective on the modular organization of language processing within LLMs.

The results provide empirical evidence that LLMs represent different aphasia subtypes and linguistic error types through largely distinct sets of neurons organized in functionally separable modules. The layer-wise distributions of *Aphasic Neurons* revealed a hierarchical structure in which semantic processing is reflected in intermediate layers and phonological and morphological processing is concentrated in deeper layers, a pattern that held consistently across both error types and aphasia subtypes. Furthermore, the patterns of overlap and exclusivity observed across subtypes suggest that the modularity observed in LLM internal representations is not arbitrary but reflects functionally meaningful distinctions analogous to those documented in aphasia research.

Building on these empirical findings, this study demonstrates that the analytical framework developed through aphasia research offers a novel and productive perspective to investigate the modular organization of LLMs. By applying the logic of functional dissociability to the analysis of internal representations, we show that insights from neurolinguistic research can inform interpretability research in ways that go beyond behavioral or output-level analyses.

Limitations

The present study has several limitations. First, the models compared in Experiment 1 differ in architectural design, including layer count and activation function, making it difficult to fully disentangle the effects of architecture from those of linguistic processing. Second, the identification of *Aphasic Neurons* in this study is based solely on activation patterns, and whether ablating or manipulating these neurons affects linguistic processing remains to be examined. Third, the top-1% threshold for neuron selection was adopted from Suau et al. (2020) to ensure comparability with prior work, but its optimality for detecting aphasic speech patterns remains an open question.

Future work

Extending the analysis to models with more uniform architectural designs would help clarify whether the observed patterns reflect genuine linguistic processing rather than model-specific characteristics. Additionally, ablation studies would allow direct verification of whether the identified *Aphasic Neurons* are functionally involved in linguistic processing, providing stronger evidence for the modular organization of LLMs.

References

Badr AlKhamissi, Greta Tuckute, Antoine Bosselut, and Martin Schrimpf. 2025. The LLM language network: A neuroscientific approach for identifying causally task-relevant units. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10887–10911, Albuquerque, New Mexico. Association for Computational Linguistics.

Roelien Bastiaanse, Dicky Gilbers, and Klarien Van Der Linde. 1994. Sonority substitutions in broca's and conduction aphasia. *Journal of Neurolinguistics*, 8(4):247–255.

D. Frank Benson and Alfredo Ardila. 1996. *Aphasia: A Clinical Perspective*. Oxford University Press, New York.

Ennis Ata Berker, Ata Husnu Berker, and Aaron Smith. 1986. Translation of broca's 1865 report: Localization of speech in the third left frontal convolution. *Archives of Neurology*, 43(10):1065–1072.

Paul Broca. 1861. Remarques sur le siège de la faculté du langage articulé; suivies d'une observation d'aphémie. *Bulletins de la Société Anatomique de Paris*, 6:330–357.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. 2024. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 37–56, Bangkok, Thailand. Association for Computational Linguistics.

Yan Cong, Arianna N. LaCroix, and Jiyeon Lee. 2024a. Clinical efficacy of pre-trained large language models through the lens of aphasia. *Scientific Reports*, 14:15573.

Yan Cong, Jiyeon Lee, and Arianna LaCroix. 2024b. Leveraging pre-trained large language models for aphasia detection in English and Chinese speakers. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 238–245, Mexico City, Mexico. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Antonio R. Damasio. 1992. Aphasia. *New England Journal of Medicine*, 326(8):531–539.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240. ACM.

Nina F. Dronkers and Maria V. Ivanova. 2023. The neuroscience of language and aphasia. In Gregory G. Brown, Bruce Crosson, Kathleen Y. Haaland, and Tarrant Z. King, editors, *APA handbook of neuropsychology: Neuroscience and neuromethods*, pages 139–158. American Psychological Association.

Nina F. Dronkers, Maria V. Ivanova, and Juliana V. Baldo. 2017. What do language disorders reveal about brain-language relationships? from classic models to network approaches. *Journal of the International Neuropsychological Society*, 23(9-10):741–754.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. *Preprint*, arXiv:2010.02695.

Jerry A. Fodor. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press, Cambridge, MA.

Margaret M. Forbes, Davida Fromm, and Brian MacWhinney. 2012. Aphasiabank: A resource for clinicians. *Seminars in Speech and Language*, 33(3):217–222.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Norman Geschwind. 1965. Disconnexion syndromes in animals and man. *Brain*, 88(3):585–644.

Harold Goodglass and Edith Kaplan. 1983. *Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins, Philadelphia, PA.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Gregory Hickok and David Poeppel. 2000. Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, 4(4):131–138.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Andrew Kertesz. 1982. *Western Aphasia Battery*. Grune and Stratton, New York.

Huykien Le, Forshing Lui, and Mickey Y. Lui. 2024. Aphasia. In: StatPearls [Internet]. StatPearls Publishing, Treasure Island (FL). Updated October 29, 2024.

Ludwig Lichtheim. 1885. On aphasia. *Brain*, 7(4):433–484.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. 2011. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Hisako Monoi, Yoko Fukusako, Motonobu Itoh, and Sumiko Sasanuma. 1983. Speech sound errors in patients with conduction and broca's aphasia. *Brain and Language*, 20(2):175–194.

Liina Pylkkänen. 2019. The neural basis of combinatory syntax and semantics. *Science*, 366(6461):62–66.

Zihan Qiu, Zeyu Huang, and Jie Fu. 2024. Unlocking emergent modularity in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2638–2660, Mexico City, Mexico. Association for Computational Linguistics.

Neguine Rezaii, Yao Xiong, Jamie Reilly, and Murray Grossman. 2024. Artificial intelligence classifies primary progressive aphasia from connected speech. *Brain*, 147(9):3070–3082.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Tim Shallice. 1988. *From Neuropsychology to Mental Structure*. Cambridge University Press, Cambridge.

Noam Shazeer. 2020. GLU variants improve transformer. *Preprint*, arXiv:2002.05202.

Xavier Suau, Luca Zappella, and Nicholas Apostoloff. 2020. Finding experts in transformer models. *Preprint*, arXiv:2005.07647.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Hans-Lukas Teuber. 1955. Physiological psychology. *Annual Review of Psychology*, 6:267–296.

Chengcheng Wang, Zhiyu Fan, Zaizhu Han, Yanchao Bi, and Jixing Li. 2025. Emergent modularity in large language models: Insights from aphasia simulations. *Preprint*, bioRxiv:2025.02.22.639416.

Carl Wernicke. 1874. *Der aphasische Symptomencomplex*. Cohn and Weigert, Breslau.

A Layer-wise Distribution of Aphasic Neurons Across Models

Figure 6 shows the layer-wise distribution of *Aphasic Neurons* across phonological, semantic, and morphological error types for all three models (min_words = 10).

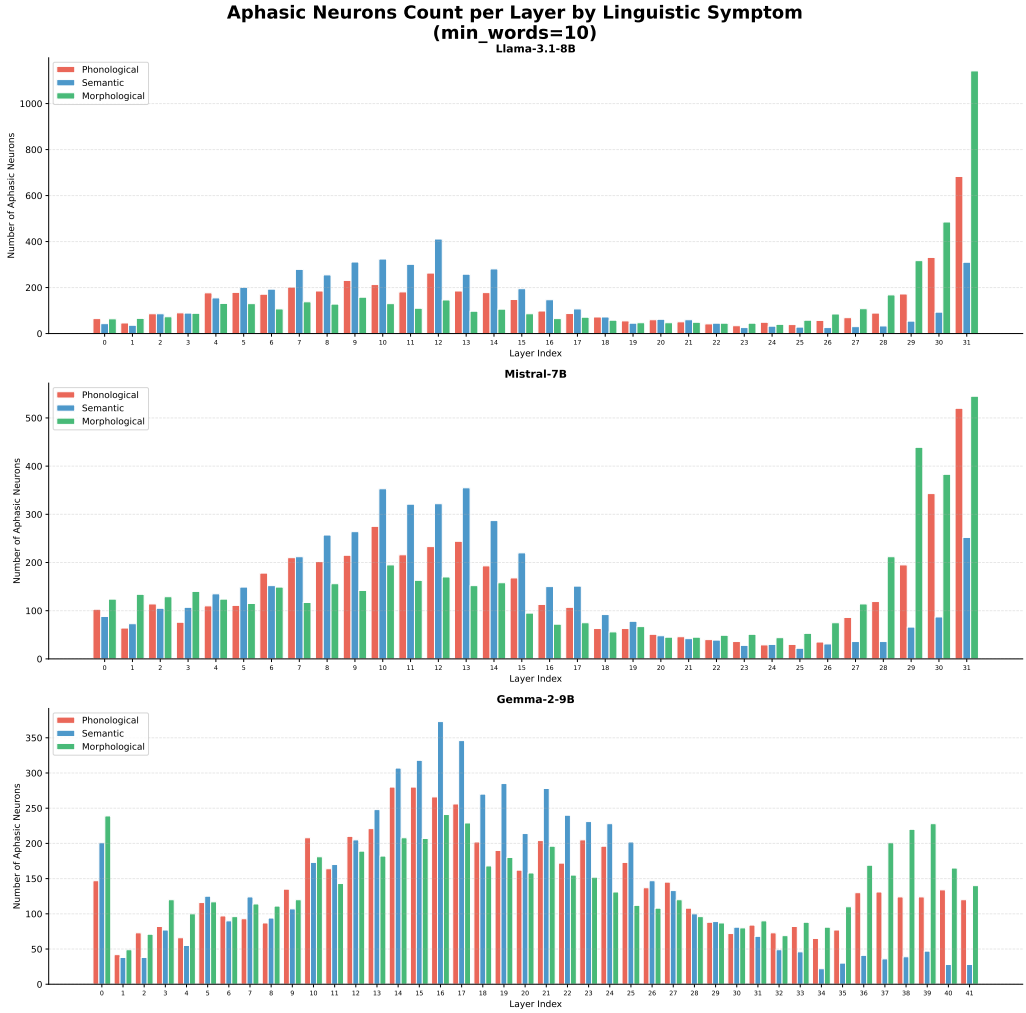

Figure 6: Layer-wise distribution of *Aphasic Neurons* across phonological, semantic, and morphological language impairment conditions for all three models (min_words = 10).