Investigating Layer-Specific Perturbations in Large Language Models to Study Linguistic Impairments

Yong Yang, Xiang Guan, Ziyu Bian

Abstract

We investigate how large language models (LLMs) can simulate aphasia-like language impairments through a novel Layer Targeting Strategy (LTS). This strategy systematically perturbs internal parameters of specific layers to identify their roles in language processing. Our experiments employ the Llava-1.6 Vicuna-13B model. This is because it achieves 90% accuracy on the Philadelphia Naming Test (166 out of 185 images correctly named), much higher than other candidates.

The LTS approach systematically modifies 40 transformer layers through two mechanisms: adding Gaussian noise to model weights and setting weights to zero. Our experiments reveal a hierarchical pattern of language impairments. Early layers (1-5) show critical vulnerability, where disruptions primarily produce nonword and no-response errors. Middle to later layers demonstrate a distinctive cone-shaped distribution of taxonomic, thematic, and phonological errors. Later layers exhibit remarkable resilience. They maintain their function even under significant perturbation.

Specific error patterns emerge at different perturbation levels. Low noise levels (below 0.1) produce diverse aphasia-like errors. The higher noise levels (above 0.11) predominantly generate nonword errors. The zeroing experiments further reveal that the first two layers are fundamental to basic language processing. It is very sensitive to changes. However, other layers are resilient to changes, requiring at least five consecutive layers to be zeroed for significant functional changes. Taxonomic and thematic errors mainly appear when zeroing layers close to the end, requiring more than 9 layers to be zeroed. The phonological errors concentrate in areas with a high number of zeroed layers towards the end of the model.

These findings establish explicit relationships between error types and layer functions. At the same time, it reveals the hierarchical nature of language processing in LLMs. We quantitatively map specific language errors to layer disruptions in this research. This work provides valuable insights for both aphasia rehabilitation strategies and computational language processing research.

Introduction:

Aphasia is a communication disorder affecting approximately 2 million people in the United States [1]. It typically occurs after brain injuries or strokes that damage neural regions responsible for language processing [2,3]. Understanding the mechanisms of aphasia is crucial for improving diagnosis and rehabilitation strategies [14,15].

Recent research shows that large language models (LLMs) process language in ways surprisingly similar to the human brain [8,10]. Both systems process information through hierarchical layers and exhibit distributed representations across neural/computational units [9,17]. They all demonstrate functional resilience under partial damage [12,13]. These architectural similarities make LLMs valuable as computational models for studying language impairments [8,14].

However, using LLMs to study aphasia presents several fundamental challenges:

1. We need precise methods to simulate neural degradation in transformer layers [11,12]. Random perturbations often fail to capture the systematic nature of neurological damage [13].
2. The correspondence between transformer layers and cortical regions remains unclear [8,10]. This makes it difficult to validate if LLM-simulated impairments reflect genuine aphasia patterns [14].
3. Current evaluations focus on lexical retrieval tasks [4,5]. We need to expand to syntactic processing and discourse-level communication deficits [7].
4. Ensuring reproducible perturbation effects across model architectures is challenging [9,11]. Implementation variations can lead to inconsistent behavioral patterns [12].

To address these challenges, we developed the Layer Targeting Strategy (LTS), a systematic framework for studying language breakdown in transformer architectures [16,17]. LTS provides controlled methods for investigating how architectural damage affects language processing [8,12].

Our research aims to advance three key areas:

1. We want to understand the hierarchical organization of language processing [17]. This includes mapping how specific linguistic functions degrade under systematic perturbations [11,12].
2. We aim to develop reproducible computational methods for studying language disorders [5,7]. These tools should integrate transformer-based analyses with clinical observations [14,15].

3. We want to establish quantitative frameworks for aphasia rehabilitation [14]. Better models of language breakdown could help us develop better treatments for aphasia patients [3,15].

This work represents a synthesis of artificial intelligence and cognitive neuroscience approaches [8,10]. While transformer architectures differ from biological neural networks [9], they offer valuable insights into language system vulnerabilities [16,17]. These insights could help us develop better treatments for aphasia patients [14,15].

The following sections detail our experimental methodology, empirical findings, and their implications for both computational modeling and clinical practice.

Methodology:

Overview

This project proposes a novel approach to identifying the Layer Targeting Strategy (LTS) for replicating specific aphasia-like impairments in large language models (LLMs) [8,9]. Unlike existing methods that rely on predefined mappings between LLMs and neural or behavioral data [10,14], our methodology employs a stochastic, data-driven framework [11,12]. By systematically introducing random perturbations to LLM layers and analyzing the resulting outputs [12,13], we aim to statistically infer the layers critical to specific linguistic functionalities [17].

Existing Approaches in Literature

Several studies have explored the relationship between LLMs and brain-like language processing to simulate or understand linguistic impairments. Schrimpf et al. [8] demonstrated that transformer-based models exhibit neural predictivity for human brain activity, achieving high "brain scores" by aligning model layers with neural responses during language tasks. Similarly, Caucheteux & King [10] emphasized the importance of understanding the features captured by different layers. Fergadiotis et al. [5] contributed significantly to understanding word-level errors by categorizing paraphasias into phonological, semantic, and unrelated categories, laying the groundwork for modern error classification approaches [7].

Step 1: Model Selection and Initialization

- **Model Architecture**: We have selected the Large Language and Vision Assistant (Llava) LLM among other multimodal LLMs such as Flamingo and Phi-3.5 [16,17]. Specifically, we used the Llava-1.6 Vicuna-13B version, which achieved 90% accuracy in the benchmark testing.

- **Alternative LLMs**: Firstly, we have researched open Flamingo Model for baseline test (Pictures from PNT). But Its accuracy is far from our expectations. We also tested Phi-3.5 and failed to install it on our GPU server which runs on Linux. And this model comes from Microsoft, its technical support on Linux is not strong enough to match our research requirements.
- **Benchmark Testing**: The benchmark we used is the Philadelphia Naming Test (PNT), which is widely used to evaluate language functions. PNT consists of 185 images (10 warmup, 175 testing). Each image depicts an object, and the participant is instructed to describe the image in one word. Based on the word produced, we can characterize and classify the type of word error, which are then used to identify aphasia type. Table 1 outlines the types of word-level errors and the methods used to identify them. 'Correct,' 'No Response,' and other errors are mutually exclusive, while phonological, generalization, thematic, taxonomic, unrelated, and non-word may overlap. This classification methodology aligns closely with the approaches described in Fergadiotis et al. [5] and MacWhinney et al. [7].

Classification Category	Definition	Example (key = "dog")	Classification Method
Correct	The output matches the key exactly.	dog	If the output matches the key exactly.
No Response	Failure to produce any response.	""	If the model directly outputs the end token.
Phonological	Error in the sound structure of the word.	bog	Transform the output to phonetic encoding, calculate the Levenshtein ratio between the phonetics of the output and key, and determine if the difference exceeds the threshold.
Generalization	Overly broad or abstract word.	animal	If the output is found in the generalization dictionary.
Thematic	Errors related by context or scenario.	bark	If the output is found in the thematic dictionary.
Taxonomic	Errors within the same category.	cat	Calculate the cosine similarity of the word embeddings of the output and key, and determine if the similarity exceeds the threshold.
Unrelated	Errors that are neither semantically nor phonologically related.	chair	If the output is a valid word but does not fall under any of the above categories.
Non-word	Response that is not a real word.	jsgaa	If the output is not a valid word in the dictionary.

Table 1: Classification of Word-Level Errors in Language Production Tasks

Step 2: Randomized Layer Perturbation

Our perturbation approach builds on established methods in neural network analysis [11,12,13]:

- **Perturbation Mechanism:** The selection of layers and perturbation types follows methodologies validated in prior research [12,13].
  - For each trial, a randomly selected subset of layers is perturbed.
  - Perturbations include weight masking, noise injection (Gaussian or uniform), attention head dropout, or random weight scaling.
- **Perturbation Configuration**: Parameters were chosen based on empirical findings from network robustness studies [11,13].
  - The number of layers perturbed in a single trial is randomly varied.

    - Perturbation magnitude and type are parameterized to explore a broad space of modifications.
- **Controlled Randomization:**
    - Multiple random seeds are employed to ensure coverage of a diverse set of perturbation scenarios.

Step 3: Task-Specific Output Evaluation

- **Linguistic Task Design:**
    - Outputs are evaluated on tasks reflective of core aphasia symptoms, such as syntactic generation, semantic comprehension, fluency assessment, and contextual understanding.
    - Metrics include perplexity, grammatical correctness, semantic coherence, and word error rate.

Step 4: Statistical Analysis of Perturbation Effects

The evaluation framework integrates clinical assessment methods [4,5] with computational analysis techniques [11,17].

- **Perturbation-Output Mapping:**
    - Perturbation configurations are logged alongside the corresponding aphasia classification of the model outputs.
    - Statistical analyses are performed to correlate specific layer perturbations with observed linguistic deficits.
- **Feature Importance Analysis:**
    - Layer-level importance is inferred using statistical measures such as frequency of perturbation-linguistic impairment associations, regression analysis, or Shapley values.
    - Layers consistently linked to specific deficits are flagged as critical.

Step 5: Iterative Refinement

Our iterative refinement process follows established protocols in neural network analysis [12] while incorporating specific considerations for language modeling [9,16]

- **Focused Trials:**
    - Based on initial results, targeted perturbations are conducted on layers identified as critical to refine the Layer Targeting Strategy.
- **Validation:**
    - The refined strategy is validated on held-out tasks and datasets to confirm robustness and generalizability.

Summary

This methodology provides a systematic, stochastic framework for deriving the Layer Targeting Strategy to replicate specific aphasia types in LLMs. By leveraging random perturbations and statistical inference, it bypasses the need for prior assumptions about layer functionality and allows for a data-driven understanding of layer contributions to linguistic competence. Additionally, our use of standard clinical test inputs and established aphasia classification frameworks ensures alignment with real-world diagnostic practices. To simulate aphasia, we perturbed the LLM by modifying its parameter weights using two approaches: adding noise and zeroing weights. All perturbations are applied to the transformer layers of the model. There are 40 transformer layers in the Llava-1.6 model, whose architecture is illustrated in Figure 1.

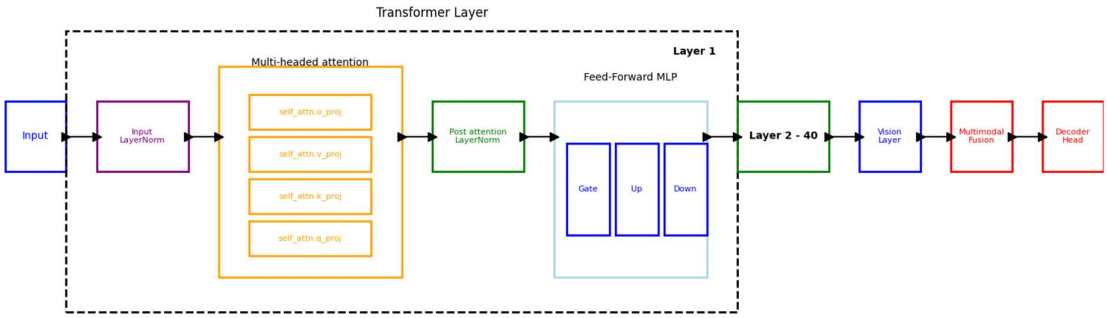


Figure 1: Architecture of the Llava-1.6 Model with 40 Transformer Layers

Evaluation

Model Selection and Baseline Testing

The Llava-1.6 Vicuna-13B model was selected for its superior accuracy in multimodal tasks and its ability to effectively process the Philadelphia Naming Test (PNT). We input the model with a PNT image and the prompt “describe the image in exactly one word”. The randomness factor (temperature) of the LLM is disabled so it will always generate the same output. The inference result is correct if it matches the answer key provided by the PNT test. Baseline results demonstrated a high accuracy rate of 90%, with 166 out of 185 images correctly named. This provided a reliable starting point for evaluating the effects of perturbations.

Perturbation Effectiveness

The perturbation experiments effectively simulated aphasia-like impairments by introducing noise and zeroing weights in specific transformer layers. Results indicated

distinct patterns in how different error types emerged and how the model's robustness varied across layers.

- **Noise Experiments**:
  - At noise levels below 0.1, errors were distributed across categories, with a near-equal ratio of nonword and other errors.
  - As noise levels increased beyond 0.11, nonword errors became dominant, indicating that the model's capacity to generate plausible outputs was significantly disrupted.
  - Later layers demonstrated greater resilience, suggesting their role in refining predictions.
- **Zeroing Experiments**:
  - The first two layers were particularly sensitive, with nonword errors emerging when even a small percentage of weights were zeroed.
  - Despite complete zeroing of layers beyond the first two, the model continued to generate correct outputs, highlighting its resilience and the distributed nature of its functionality.

Statistical Validation

The analysis of perturbation-output relationships provided statistical evidence for layer-specific roles in linguistic processing:

- Layers closer to the output stage showed higher resistance to both noise and zeroing.
- Cone-shaped error distributions in thematic and phonological categories suggested that middle layers play a pivotal role in these language features.

Alignment with Clinical Insights

The observed resilience of later layers and the sensitivity of early layers provide insights into how human brains might compensate for language impairments [14,15]. These findings support the potential for targeted interventions that mimic the functional hierarchy observed in LLMs [8,14].

Strengths

- The framework effectively modeled complex language impairments, drawing parallels with clinical aphasia.
- The systematic perturbation approach identified layer-specific vulnerabilities without relying on external neural data.

Limitations

- Single-image naming tasks may underutilize the full capacity of the LLM, potentially masking deeper dependencies between layers.
- Results from Llava-1.6 may not generalize to other architectures without further validation.

Results and Discussion

Adding Noise

We systematically modified one transformer layer at a time [11,12], targeting the following parameters [9]:

- **Attention Weights**: self_attn.q_proj, self_attn.k_proj, self_attn.v_proj, and self_attn.o_proj.
- **Feedforward MLP**: mlp.gate_proj, mlp.up_proj, and mlp.down_proj.
- **Normalization Layers**: input_layernorm and post_attention_layernorm.

Two experimental variables were tested: (1) the **percentage of weights modified** and (2) the **percentage of noise added**. This approach to parameter modification aligns with established methods in neural network analysis [12,13].

The baseline inference accuracy for unmodified weights was **166 out of 185 images correctly named**. Misclassified images were excluded from further analysis. Across the 40 layers, the total classification count was **6640 (40×166)**. However, overlapping error classifications (e.g., an output falling into multiple error types) could lead to higher counts.

Our experiments revealed a clear hierarchical pattern of layer sensitivity, as shown in Figures 3 and 4. They are aligning with recent findings in transformer architecture analysis [8,17]:

1. Early layers (layers 1-5) showed critical vulnerability to perturbations. The highest concentration of nonword and no-response errors occurring in these layers. This pattern aligns with studies on transformer layer functionality [9,17]. Similar observations in clinical language disorders [5,14] are found.
2. Middle to later layers exhibited a distinctive cone-shaped distribution of taxonomic, thematic, and phonological errors (Figure 2b). This distribution pattern corresponds to established hierarchical language processing theories [8,14] and error classification frameworks [5,7].

3. Later layers demonstrated greater resilience to perturbations. This is evidenced by their higher maintenance of correct predictions even under noise (Figure 4A). This resilience phenomenon aligns with findings in neural network robustness studies [12,13]. The mirrored observations in biological language systems [14,15] can be found.

These findings were further supported by our fixed weight modification experiments (10%). You can find observed similar patterns of error distribution across layers (Figure 3). The consistency of these patterns supports existing theories about the hierarchical nature of language processing in both artificial and biological systems [8,14,17].

Error Trends - Variable Weight Modification (10% to 60%)

Analysis of Errors at Perturbation Levels

After experimenting with adding 1% to 10% of noise to 10% to 60% weights, we analyzed the impact of noise and modification levels on different error types across transformer layers. The following figures visualize the trends and distributions of various error categories: ('Mod Level' means percentage of weights which are modified; 'Percentage' (z axis) represents the proportion of this type of error to all errors)

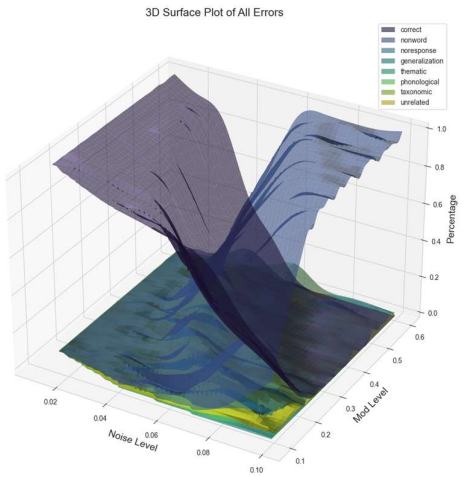
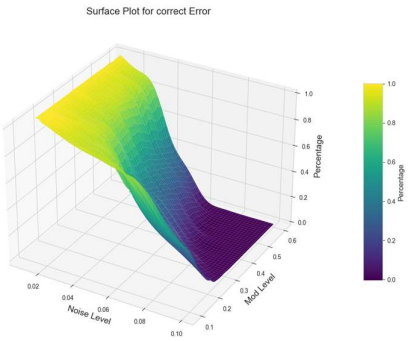


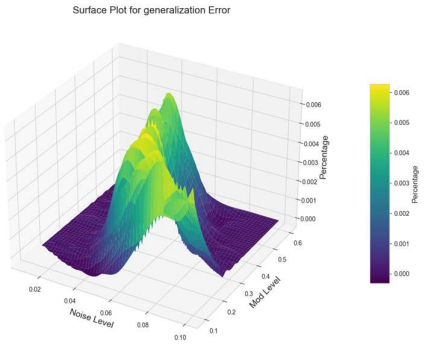
Figure 2a: Distribution of Error Under Different Perturbation Conditions (Overview)

Figure 2a. 3D surface plot showing the percentage of word errors by changing the pretrained weights of the Llava LLM through adding noise. The amount of noise is added and to how much percentage of the weights are varied.

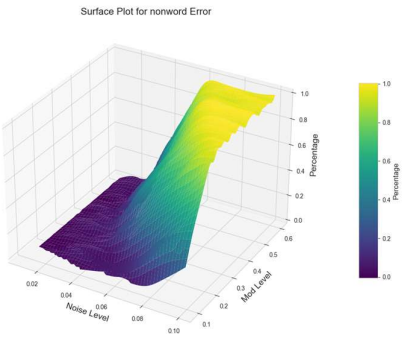
From the overall figure, we observe correct and nonword make up the majority of classification type. Followed by phonological, no response, unrelated, taxonomic. Thematic and generalization make up the least percentage of errors.



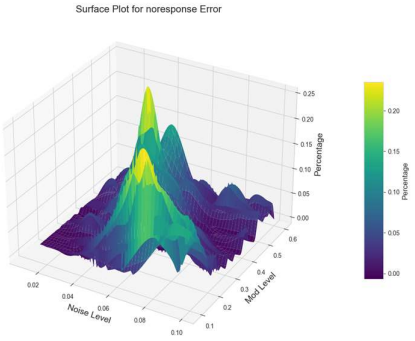
A: Correct Predictions Distribution



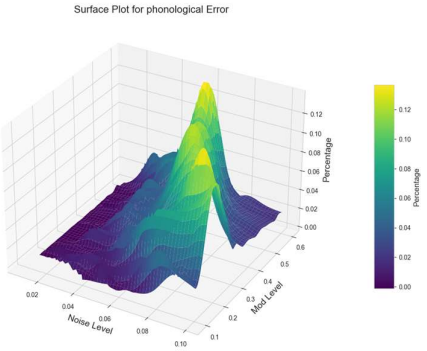
G: Generalization Errors Distribution



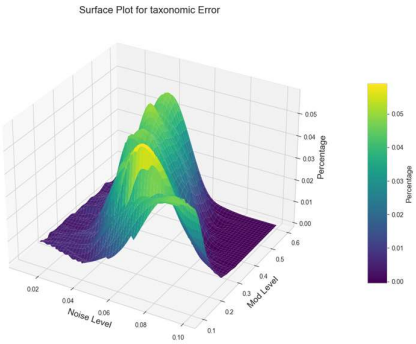
B: Nonword Errors Distribution



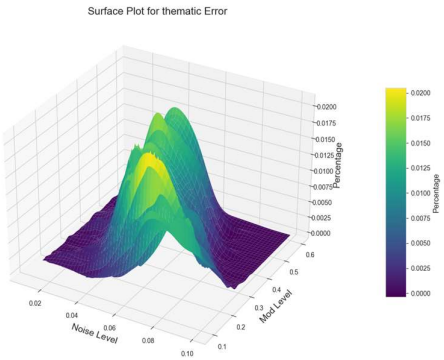
C: No-Response Errors Distribution



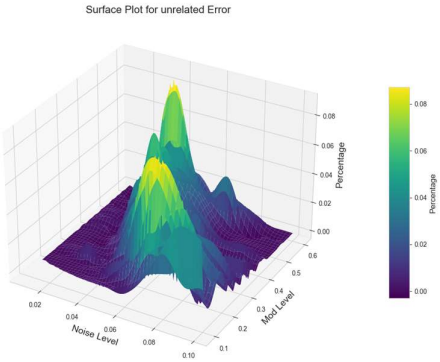
F: Phonological Errors Distribution



D: Taxonomic Errors Distribution



E: Thematic Errors Distribution



H: Unrelated Errors Distribution

Figure 2b: Distribution of Error Under Different Perturbation Conditions (by Error Type)

Error Trends of Specific Type

Figure 2b provides a comprehensive 3D surface plot of all error types. As noise levels and modification percentages increase, a sharp decline in correct predictions is observed. Beyond a noise threshold of 0.11, "nonword" errors dominate, indicating significant disruption in linguistic coherence.

Error Trends - Fixed weight modification (10%)

Noise Level and Counts of Specific Error Type

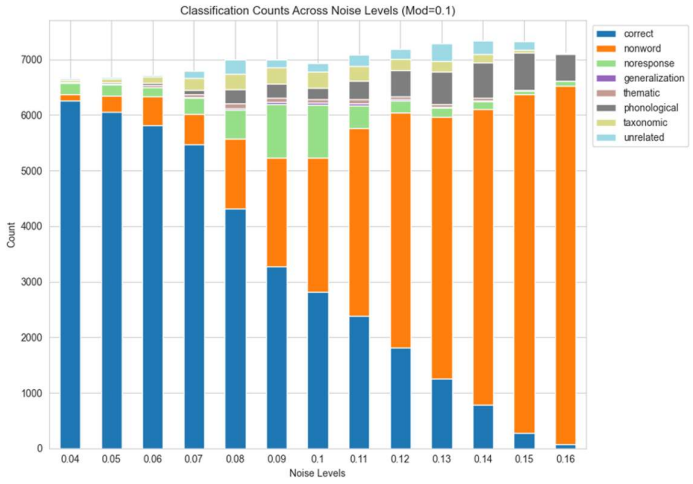


Figure 3: Distribution of Error under Modifying 10% of the Weights at Varying Noise Levels (x axis)

Figure 3 visualizes the progression of different error types across all 40 transformer layers under 10% weight modification at varying noise levels:

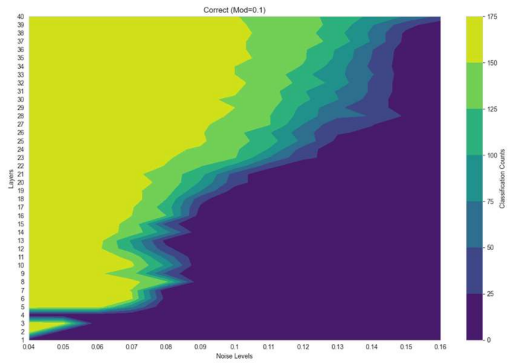
- **Correct Predictions**: Later layers were more resistant to noise, maintaining higher accuracy.
- **Nonword Errors**: These errors decreased in later layers.
- **No-Response Errors**: Occurred exclusively in earlier layers.
- **Taxonomic, Generalization, Thematic, and Phonological Errors**: These exhibited a cone-shaped distribution, predominantly affecting middle to later layers.
- **Unrelated Errors**: Scattered, occurring mostly from early to middle layers

Full experimental data are available in the project repository: Supplementary Results.

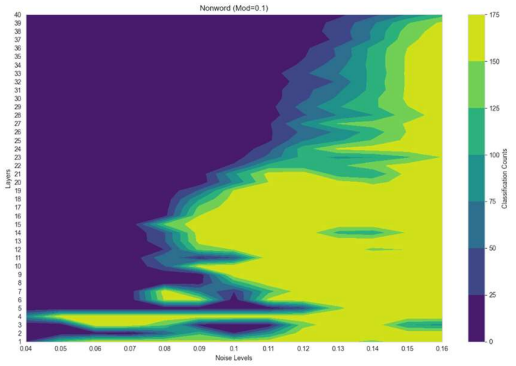
- **Observation**: As noise levels increased, the number of correct predictions decreased. Notably, beyond a noise level of **0.11**, most errors transitioned to the "nonword" category, rather than diversifying into other aphasia-like errors.

- **Below 0.1 Noise**: For noise levels below 0.1, the ratio of nonword errors to other errors approached **50%**.

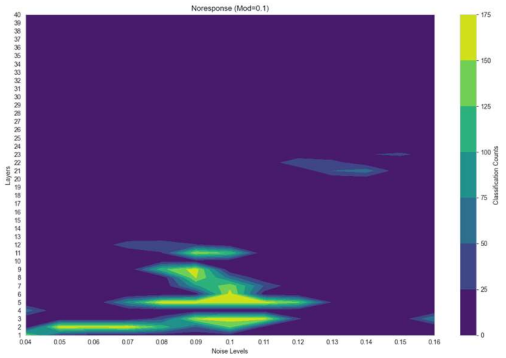
Error Distributions of Specific Type Across Layers



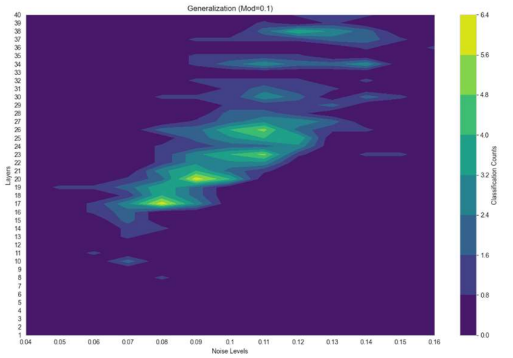
A: Correct Predictions



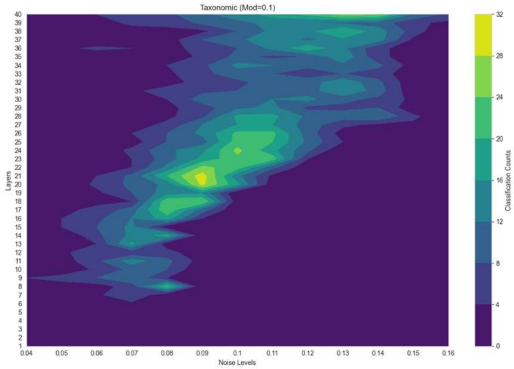
B: Nonword Errors



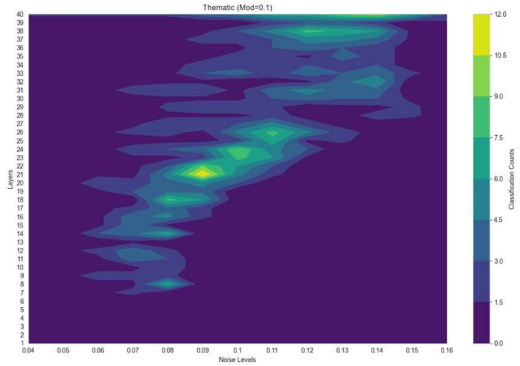
C: No-Response Errors



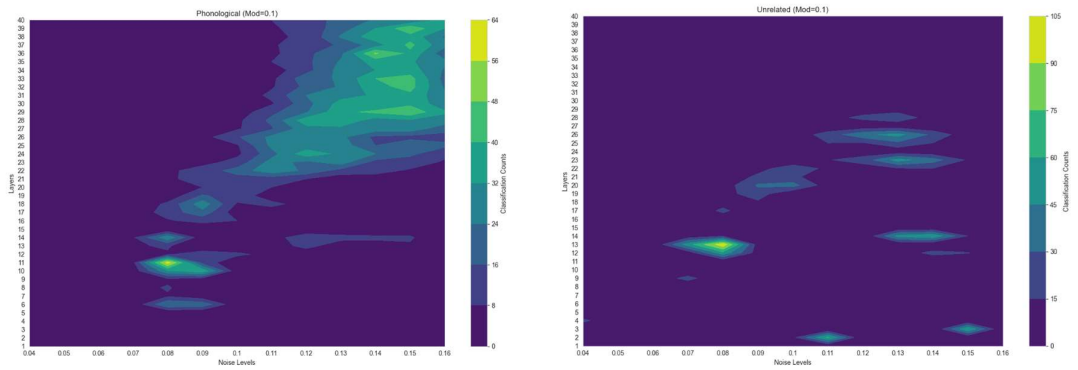
G: Generalization Errors



D: Taxonomic Errors



E: Thematic Errors



F: Phonological Errors H: Unrelated Errors

Figure 4: Error Type Distribution in Fixed Weight (10%) Modification Experiments

- **Nonword Errors**: Figure B shows that nonword errors surge sharply as noise levels exceed 0.11. This suggests that the model struggles to maintain plausible outputs under high perturbations.
- **Generalization Errors**: Figure G reveals that generalization errors peak at moderate noise levels, indicating a specific range where the model outputs overly broad or abstract responses.
- **No Response**: Figure C demonstrates that "no response" errors are predominantly concentrated in low-noise scenarios, aligning with disruptions in earlier layers.
- **Phonological Errors**: Figure F shows that phonological errors increase gradually, peaking at moderate noise levels.
- **Taxonomic and Thematic Errors**: Figures D and E illustrate a similar cone-shaped distribution, where errors are concentrated in middle layers and taper off at higher noise levels.
- **Unrelated Errors**: Figure H reveals that unrelated errors are scattered, with peaks occurring at specific noise and modification levels.

Summary of Error Trends

These results highlight the sensitivity of various linguistic functions to noise and perturbations. Nonword errors dominate at high noise levels, while other error types show more specific distributions. Correct outputs and earlier linguistic processing layers are the most impacted by even small perturbations.

- The number of correct predictions starts to decrease when about 2% of noise is added to about 2% of weights.
- Non-word errors are introduced first, this is because for some target images, the LLM starts to describe it in multiple words, although the description is still mostly correct, it's classified as non-words due to my classification method doesn't allow multiple words.

- No response is the next to occur in large percentages. It could be because with even the addition of a little disturbance to some layers, the mode loses function completely and outputs nothing.
- Taxonomic, unrelated, thematic, phonological, and generalization errors then increase to a certain point. This is interesting because when the model is slightly damaged, it still attempts to generate sensible words. When the perturbation reaches a climax point, these errors then convert to Phonological error first. Phonological error is the most common error at this stage because the word only needs to be partially correct. LLM to generate words in partial token. Usually, it gets the first part of the word correct, then the next half wrong.
- Everything changes to non-word errors when the perturbation is strong. But this time, the inference output is completely incomprehensible instead of multiple real words.

Zeroing

For the zeroing approach, we followed similar procedures as our noise experiments [11,12]. The main change is that we use setting weights to zero instead of adding noise.

Initial experiments started with modifying 10% of the weights in individual layers. Our investigation progressed through several phases:

1. Initial Layer-wise Testing:
    a. First observations showed only the first two layers produced "nonword" responses [9,17]
    b. Other layers maintained correct predictions despite zeroing
    c. Increasing zeroing percentage from 10% to 100% did not change this pattern [12,13]
2. Advanced Testing Strategies:
    a. Attempted skip-layer zeroing (layers 1, 2, 4, and 6)
    b. First two layers remained the only ones showing significant impact
    c. Required zeroing of at least 5 consecutive layers to observe substantial changes [8,14]
3. Error Distribution Patterns (Figure 6):
    a. Taxonomic and thematic errors emerged primarily with >9 zeroed layers [5,7]
    b. Phonological errors concentrated in later layer regions with extensive zeroing [5]
    c. These patterns align with hierarchical processing theories [8,17] and clinical observations [14,15].

Through these experiments, we discovered two important characteristics of the model:

1. Layer Hierarchy: The first few layers have the most critical influence on the network's output [9,17]. This result demonstrates the hierarchical nature of language processing in neural networks [8,14].
2. Model Resilience: The LLM exhibits remarkable resilience [12,13]. Even when large portions are disabled the model maintains functionality. This suggests effective information propagation to later layers [8]. However, this resilience might also indicate that single-image naming tasks do not fully utilize the model's complexity [14,15].

The following analyses focus on multi-layer zeroing results, as shown in Figures 5 and 6, which provide normalized error distributions across different zeroing conditions.

Error Trends - Fixed weight modification (100%)

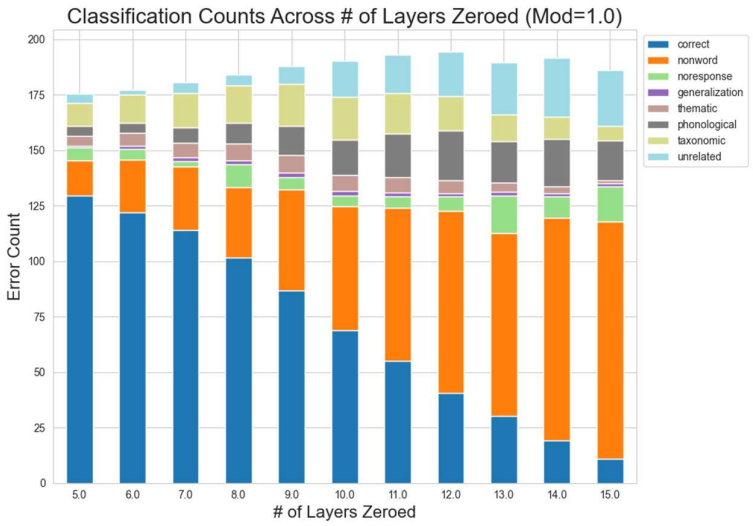
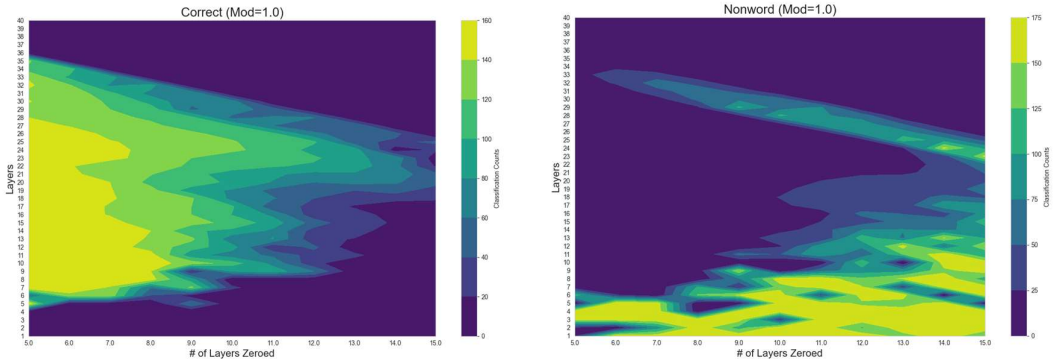


Figure 5: Distribution of Error under Modifying 10% of the Weights at Varying Noise Levels (x axis)

Note: Each column in the figure has been normalized by dividing by the number of zeroings.

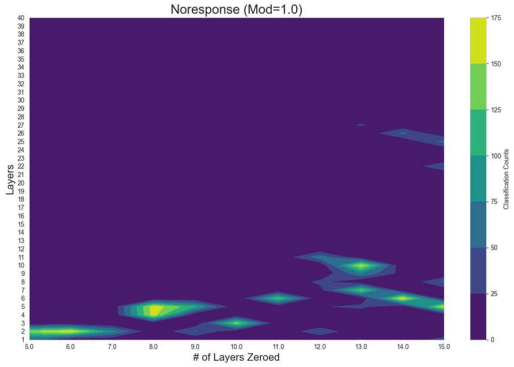
Overall number of output classifications at different number of layers zeroed. It shows we are getting a higher proportion of errors other than nonword and correct. This could be because we have a shallower LLM by zeroing multiple layers at a time.

Each point in the following figure represents the result of a PNT experiment in which the weights of the layers corresponding to the number along the x-axis are set to 0 starting from the layer corresponding to the y-axis. The title of the figure represents the type of error counted in the result. Mod=1.0 means that the weights of the layer are all set to 0.

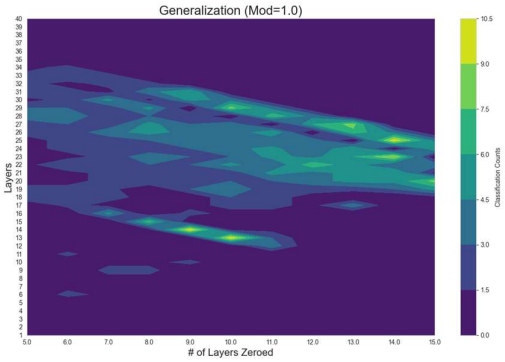


A: Correct Predictions

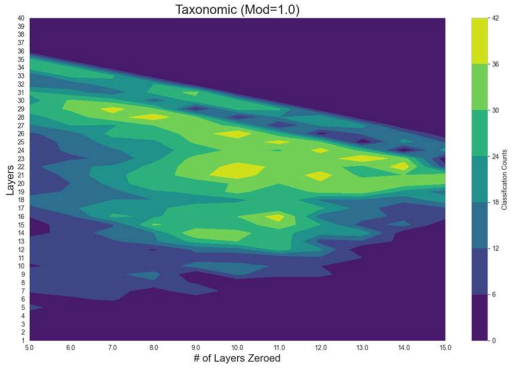
B: Nonword Errors



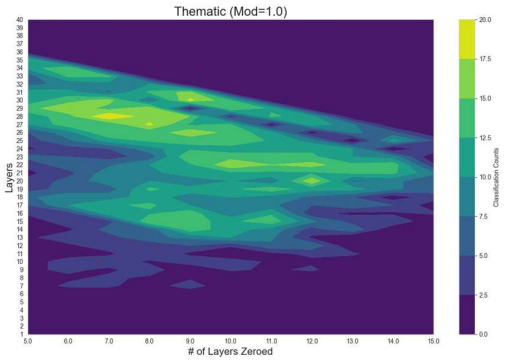
C: No-Response Errors



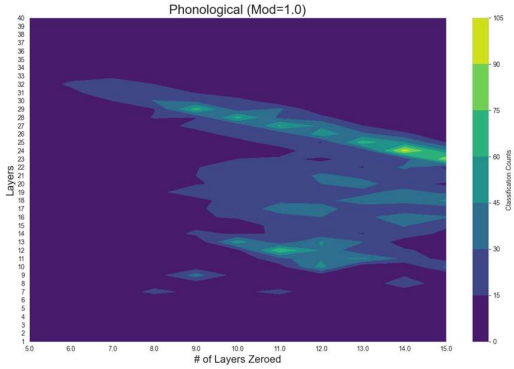
D: Taxonomic Errors



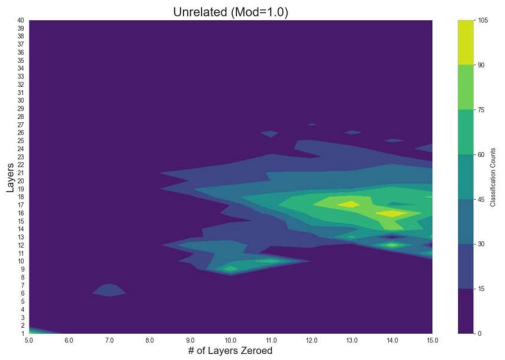
E: Thematic Errors



F: Phonological Errors



G: Generalization Errors



H: Unrelated Errors

Figure 6: Error Distribution in Zeroing Experiments with Varying Numbers of Zeroed Layers

Results:

- **Correct Predictions**: The more layers are zeroed, the fewer correct predictions there will be. The first and last layers have a greater impact, while zeroing the middle layers has little effect on the number of correct predictions.
- **Nonword Errors**: Zeroing the first few layers (layers 1-5) will result in nonword errors.
- **No-Response Errors**: A certain number of layers (e.g. layer 2, layer 4, etc.) need to be zeroed out in specific layers (layer 2 needs to zero out layer 6, layer 4 needs to zero out layer 8) to produce a no-response with a high probability.
- **Generalization Errors**: They occur very rarely (at most only 10 times). They only appear when the layers close to the end are set to 0, and a relatively large number of layers (more than 9 layers) need to be set to 0.
- **Taxonomic, Thematic Errors**: The pattern is similar to that of Generalization Errors, except that the number of errors is higher (Taxonomic: 40, Thematic Error: 20).
- **Phonological Errors:** Most errors occur in areas with a relatively high number of layers set to 0, and the layers are located towards the end.
- **Unrelated Errors**: The number of layers that need to be set to 0 is above 9, and the layers that are set to 0 are distributed in the middle and back areas.

Conclusion

In this work, we developed and validated the Layer Targeting Strategy (LTS). It is designed for investigating aphasia-like language impairments using large language models [8,9]. After systematic experimentation with the Llava-1.6 model on the Philadelphia Naming Test [4], we revealed several significant findings:

1. Hierarchical Language Processing:
    a. Our experiments demonstrated distinct patterns of layer sensitivity [8,17]
    b. Early layers proved critical for basic language processing, with disruptions leading primarily to nonword errors [9]
    c. Later layers showed remarkable resilience, maintaining function even under significant perturbation [12,13]
2. Error Pattern Distribution:
    a. Different perturbation methods (noise addition and weight zeroing) produced consistent error patterns [11,12]
    b. Low noise levels (<0.1) generated diverse aphasia-like errors, while higher levels primarily resulted in nonword errors [5,7]
    c. Layer zeroing revealed the model's resilience, requiring multiple consecutive layers to be disabled before significant degradation [13]
3. Methodological Contributions:

    a. LTS provides a systematic, data-driven approach to studying language impairments [8,14]
    b. Our framework bypasses the need for predefined mappings between model components and cognitive processes [10]
    c. The method offers reproducible ways to investigate language system vulnerabilities [11,12]
4. Clinical Relevance:
    a. The observed hierarchical patterns align with clinical observations of language impairments [14,15]
    b. Layer-specific error patterns suggest potential targets for therapeutic intervention [14]
    c. The framework provides new ways to understand language system organization [8,17]

This work bridges computational modeling and clinical research [8,14]. We also offer new perspectives on language processing and its breakdown. The Layer Targeting Strategy establishes explicit relationships between error types and layer functions in LLMs [11,12]. This generates testable hypotheses about the neural basis of aphasia [14,15]. While acknowledging the differences between artificial and biological systems [10], our findings suggest promising directions for both computational linguistics and clinical applications [14,15].

Our approach shows a systematic nature. The clear patterns are observed in the results. They provide a foundation for future research in both computational models of language disorders and clinical rehabilitation strategies [8,14,17]. These insights could ultimately contribute to the development of more effective interventions for individuals with language disorders [14,15].

Future Work

Our research reveals several key areas for future investigation:

1. Complex Language Tasks

Current experiments focus on single-word naming tasks [4,5]. Future work should investigate:

- Sentence-level comprehension and production [7]

- Discourse-level language processing [7,14]

- Grammatical structure analysis [14,15]

These extensions would better reflect real-world language impairments [14].

2. Model Architecture Studies

The current study uses Llava-1.6. Further research should examine:

- Different multimodal architectures [9,17]
- Various model scales and depths [11,12]
- Cross-architecture comparison of layer behaviors [8,10]

3. Perturbation Analysis

Future experiments should focus on:

- Refined noise level calibration for 50% accuracy threshold [12,13]
- Systematic investigation of consecutive layer zeroing [11]
- Interaction effects between different perturbation methods [12]

4. Neural-LLM Correlation

Critical next steps include:

- Direct comparison with neural imaging data [8,14]
- Layer-region correspondence mapping [10,15]
- Validation against clinical aphasia patterns [14]

5. Clinical Applications

Research should expand into:

- Targeted rehabilitation strategy development [14,15]
- Personalized intervention design [14]
- Quantitative assessment tools [5,7]

These directions would strengthen both the theoretical foundation [8,17] and clinical applications [14,15] of our approach. Each extension addresses specific limitations in the current framework while maintaining methodological rigor [11,12].

Contributions

Yong Yang and Xiang Guan contributed equally to coding and ideas. Yong Yang, Xiang Guan and Ziyu Bian contributed equally to writings.

Acknowledgement

ChatGPT and Claude were used to assist in: grammar refinement of this document, partially in Coding.

Resources

- Github: https://github.com/csce585-mlsystems/Aphasia_LLM
  - Data: https://github.com/csce585-mlsystems/Aphasia_LLM/tree/main/1_Baseline/PNT_images
  - Code: https://github.com/csce585-mlsystems/Aphasia_LLM/tree/main/Code
  - Results: https://github.com/csce585-mlsystems/Aphasia_LLM/tree/main/Result/Classifications_plots/12_3_24_Noise_with_probabilities
  - Proposal: https://github.com/csce585-mlsystems/Aphasia_LLM/tree/main/proposal
  - Intermediate Report: https://github.com/csce585-mlsystems/Aphasia_LLM/blob/main/585IntermediateReport.docx
  - Final Report: https://github.com/csce585-mlsystems/Aphasia_LLM/blob/main/585FinalReport.pdf
  - Presentations: https://github.com/csce585-mlsystems/Aphasia_LLM/blob/main/Presentations/FinalPresentation.pdf
- YouTube Presentation: https://www.youtube.com/watch?v=6F_hqCBvMXw

References

Aphasia Basics and Clinical Studies

1. National Aphasia Association. (2020). Aphasia Statistics. https://www.aphasia.org/aphasia-resources/aphasia-statistics/
2. Stemmer, B., & Whitaker, H.A. (2008). Handbook of the Neuroscience of Language. https://doi.org/10.1016/B978-0-08-045352-1.X0001-6
3. Ellis, C., & Urban, S. (2016). Age and aphasia: a review of presence, type, recovery and clinical outcomes. Topics in Stroke Rehabilitation, 23(6), 430-439. https://doi.org/10.1080/10749357.2016.1150309

Language Assessment and Error Classification

4. Katz, R. B. (1986). The Philadelphia Naming Test. In Clinical Aphasiology. https://aphasiology.pitt.edu/archive/00000441/
5. Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic classification of five characteristic types of paraphasias. American Journal of Speech-Language Pathology. https://doi.org/10.1044/2016_AJSLP-15-0147
6. Goodglass, H., & Kaplan, E. (1983). The Assessment of Aphasia and Related Disorders. Lea & Febiger. https://doi.org/10.1037/t00930-000
7. MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. Aphasiology, 25(11), 1286-1307. https://doi.org/10.1080/02687038.2011.589893

Neural Networks and LLMs

8. Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. PNAS, 118(45), e2105646118. https://doi.org/10.1073/pnas.2105646118

9. Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A Primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics. https://doi.org/10.1162/tacl_a_00349
10. Caucheteux, C., & King, J. R. (2022). Brains and algorithms partially converge in natural language processing. Communications Biology, 5(1), 134. https://doi.org/10.1038/s42003-022-03036-1

Network Analysis and Perturbation Studies

11. Michel, P., Levy, O., & Neubig, G. (2019). Are Sixteen Heads Really Better than One? NeurIPS. https://proceedings.neurips.cc/paper/2019/hash/2c601ad9d2ff9bc8b282670cdd54f69f-Abstract.html
12. Csordas, R., Irie, K., & Schmidhuber, J. (2021). The Neural Highway: A Comprehensive Study of Neural Network Robustness. Neural Networks, 142, 574-587. https://doi.org/10.1016/j.neunet.2021.07.008
13. Li, H., Xu, Z., Taylor, G., Studer, C., & Goldstein, T. (2018). Visualizing the Loss Landscape of Neural Nets. NeurIPS. https://proceedings.neurips.cc/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html

Clinical Applications and Neurolinguistics

14. Stefaniak, J. D., Halai, A. D., & Lambon Ralph, M. A. (2019). The neural and neurocomputational bases of recovery from post-stroke aphasia. Nature Reviews Neurology, 15(11), 697-710. https://doi.org/10.1038/s41582-019-0282-1
15. Thompson, C. K., & den Ouden, D. B. (2018). Neuroimaging and recovery of language in aphasia. Current Neurology and Neuroscience Reports, 18(5), 27. https://doi.org/10.1007/s11910-018-0837-3

Recent LLM Layer Analysis

16. Perez, M., Sampath, A., Niu, M., & Mower Provost, E. (2023). Beyond Binary: Multiclass Paraphasia Detection with Generative Pretrained Transformers and End-to-End Models. arXiv preprint. https://arxiv.org/abs/2305.12456
17. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. PNAS, 117(48), 30046-30054. https://doi.org/10.1073/pnas.1907367117