SIMULATING APHASIA THROUGH WEIGHT MODIFICATIONS IN PRETRAINED LARGE LANGUAGE MODELS

Authors: Yong Yang, Xiang Guan, Ziyu Bian (Ph.D. Student)

CONTENT

- Problem Statement
- Technical Challenges
- Related Work
- Approach and Results
- Broader Impact

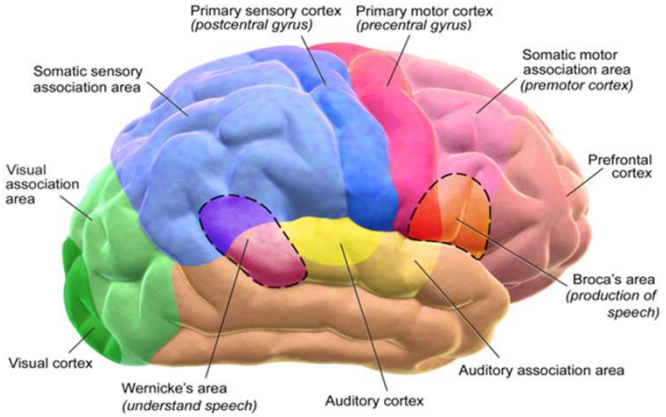
UNIVERSITY OF
South Carolina

PROBLEM STATEMENT

Problem:
Although Large Language Models (LLMs) have achieved remarkable success in natural language processing tasks, their application in simulating human language disorders has been minimally explored.

Aphasia, a language disorder caused by brain damage, provides a unique perspective on how language is processed in the human brain.

However, due to the lack of suitable computational models, progress in understanding the mechanisms and rehabilitation strategies for aphasia has been slow..



UNIVERSITY OF South Carolina

PROBLEM STATEMENT

The input to our system consists of:

- PNT Test
  - The input dataset used in this research is the Philadelphia Naming Test (PNT), which is widely used to evaluate language functions in individuals with aphasia. The PNT consists of 185 images, including 10 warmup images and 175 test images. Each image depicts a single object, and the participant is instructed to describe the image using one word.

- LLM Model
  - Llava-1.6 Vicuna-13B version, which achieved 90% accuracy in the benchmark testing

- The goal of this project is to simulate aphasia in LLMs. We modify specific layers in the model to study their impact on language. This approach helps us understand how damage affects language systems. It also allows us to explore similarities between LLMs and the brain. The findings from this study may lead to:

  1. New insights into how language breaks down under damage.
  2. Better understanding of which parts of a language system are most vulnerable.
  3. Ideas for improving rehabilitation methods using computational models.

- This study focuses on LLMs as tools to explore human cognition. While LLMs and brains are different, these models provide a controlled way to study language disruptions.

UNIVERSITY OF South Carolina

CHALLENGES

- Designing an effective and meaningful perturbation strategy for LLM layers.
- Interpreting the relationship between perturbed layers and simulated language deficits.
- Validating the simulated deficits against clinical data from aphasia patients.
- Scaling the simulations to more complex and naturalistic language tasks.
- Translating insights from LLM simulations to clinical applications for aphasia diagnosis and treatment.

RELATED WORK :

- Previous studies have explored the relationship between LLMs and brain-like language processing to simulate or understand linguistic impairments. Schrimpf et al. (2021) 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, Feghhi et al. (2024) questioned over-reliance on brain scores by deconstructing the mappings between LLM representations and neural data, emphasizing the importance of understanding the features captured by different layers.

- In terms of characterizing language errors, Fergadiotis et al. (2016) contributed significantly to understanding word-level errors by categorizing paraphasias into phonological, semantic, and unrelated categories, laying the groundwork for modern error classification approaches.

- Perez et al. (2024) further demonstrated the utility of generative pretrained transformers for multiclass paraphasia detection, aligning with the use of LLMs for aphasia-like impairment simulations in this study.

- While these approaches provide insights into functional alignment and error characterization, they rely on predefined mappings between LLMs and neural or behavioral data. In contrast, the Layer Targeting Strategy (LTS) proposed in this paper bypasses this reliance by focusing on stochastic perturbations and output-driven analysis to identify critical layers for specific language functions.

UNIVERSITY OF
South Carolina

METHODOLOGY: IDEA

- The core idea is to systematically perturb specific LLM layers and analyze the resulting errors to identify critical layers for different language functions.

- Rather than relying on predefined mappings or assumptions, we employ a data-driven, exploratory approach. By systematically perturbing the LLM layers and rigorously analyzing the resulting errors, we aim to uncover empirical relationships between the model's structure and the language deficits it exhibits under different perturbation conditions.

- This search-based methodology allows us to identify critical layers for various language functions in a more objective and unbiased manner. By letting the data guide our understanding of the model's functional organization, we minimize the risks associated with making a priori assumptions about how the model's components map onto brain regions or cognitive processes.

- Ultimately, our goal is to leverage the LLM as a controlled experimental platform to generate new insights into the nature of language breakdown, which can then inform hypotheses about the neural basis of aphasia.

METHODOLOGY

- Step 1: Model Selection and Initialization
- Step 2: Randomized Layer Perturbation
- Step 3: Task-Specific Output Evaluation
- Step 4: Statistical Analysis of Perturbation Effects
- Step 5: Iterative Refinement

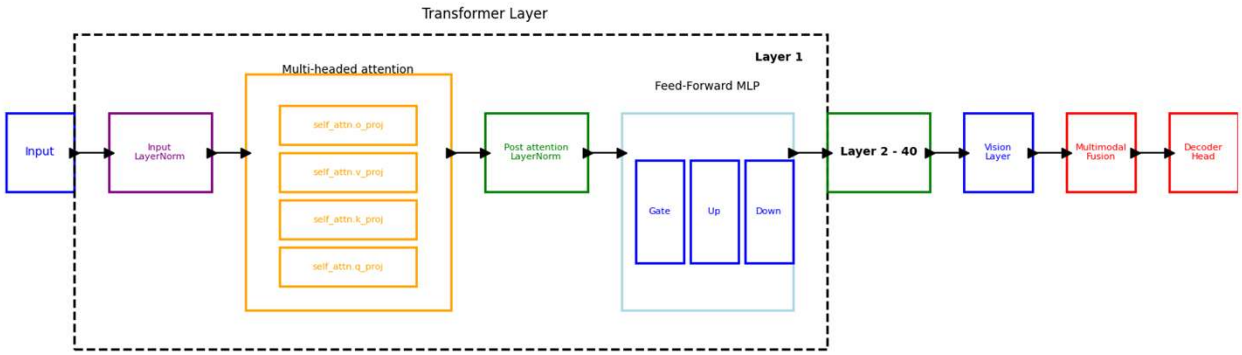
CLASSIFICATION TYPE

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.

UNIVERSITY OF South Carolina

LLM ARCHITECTURE

- **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.

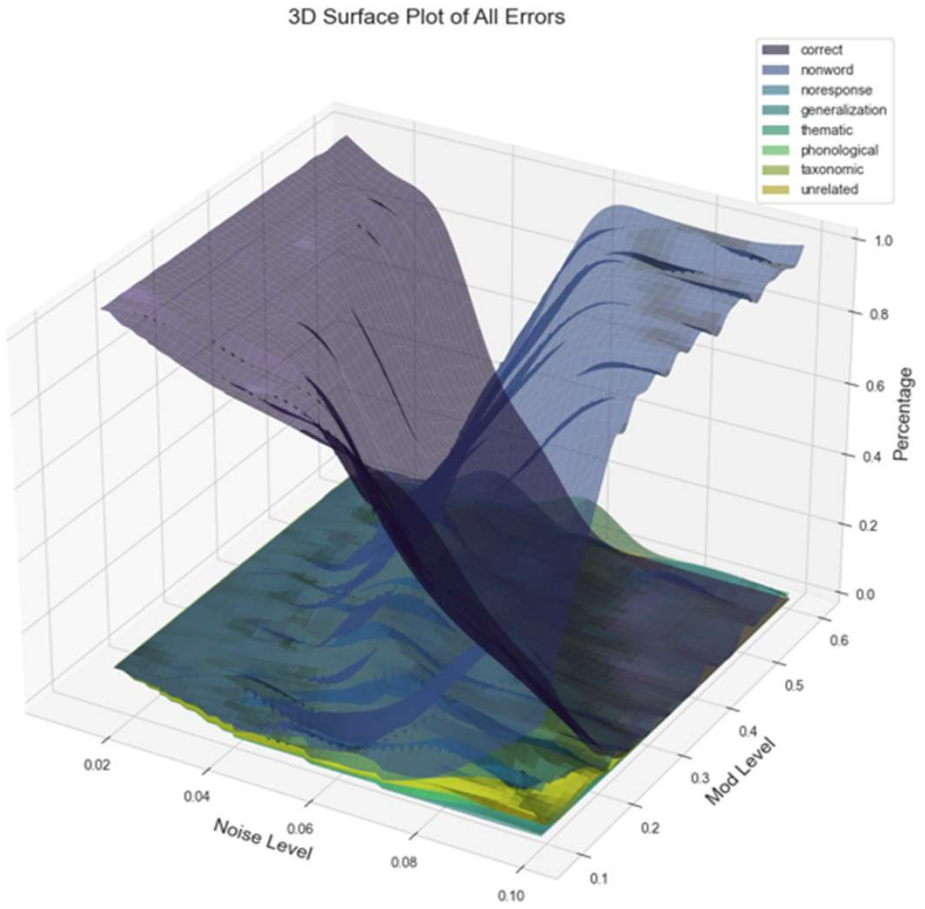


UNIVERSITY OF South Carolina

EXPERIMENTS

- Adding Noise
  - Two experimental variables were tested: (1) the **percentage of weights modified** and (2) the **percentage of noise added**.
- Zeroing
  - (1) the percentage of weights modified
- 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.

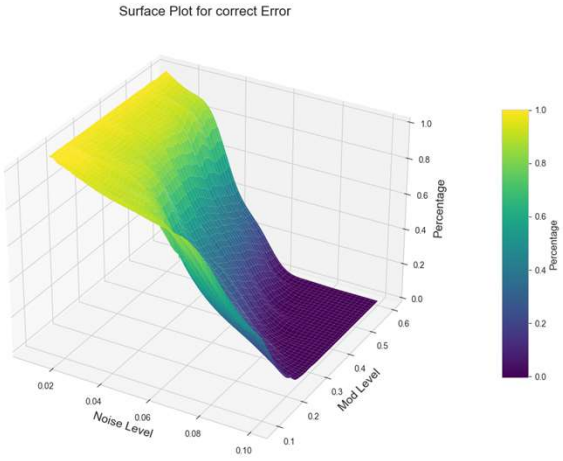
RESULTS - ERROR TRENDS - VARIABLE WEIGHT MODIFICATION (10% TO 60%)



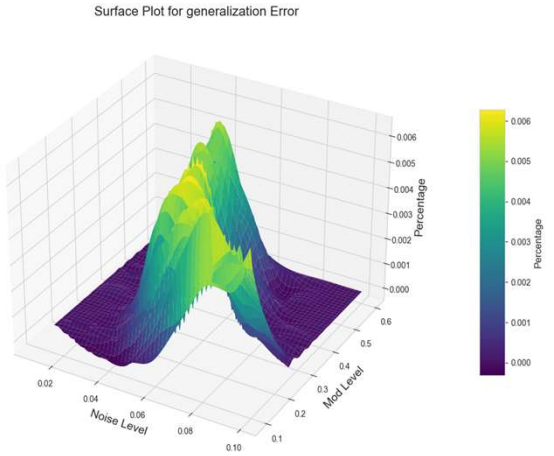
UNIVERSITY OF South Carolina

RESULTS

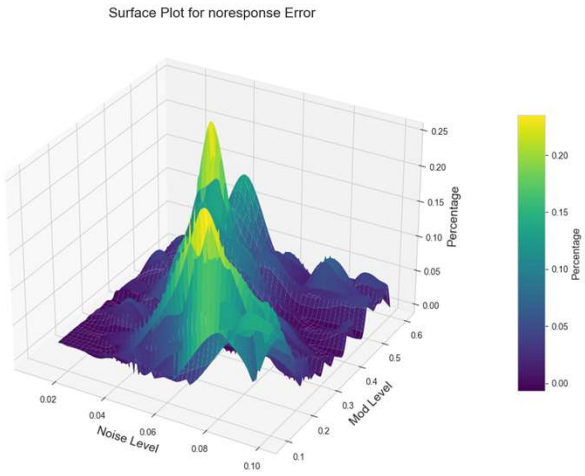
Figure 2b: Distribution of Error Under Different Perturbation Conditions (by Error Type)



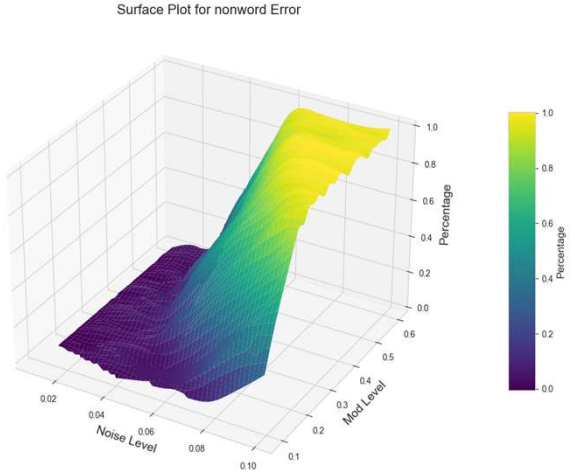
A: Correct Predictions Distribution



G: Generalization Errors Distribution



C: No-Response Errors Distribution



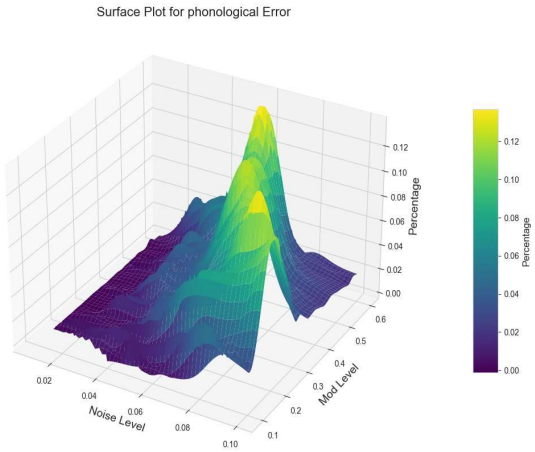
B: Nonword Errors Distribution

RESULTS

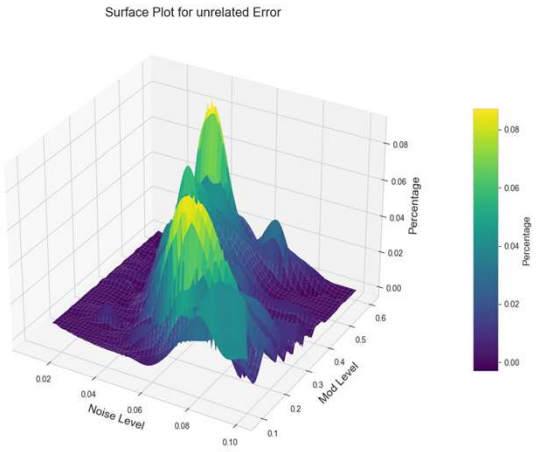
Figure 2b provides a comprehensive 3D surface plot of all error types.

As noise levels and modification percentages increase, a sharp decline in correct predictions is observed.

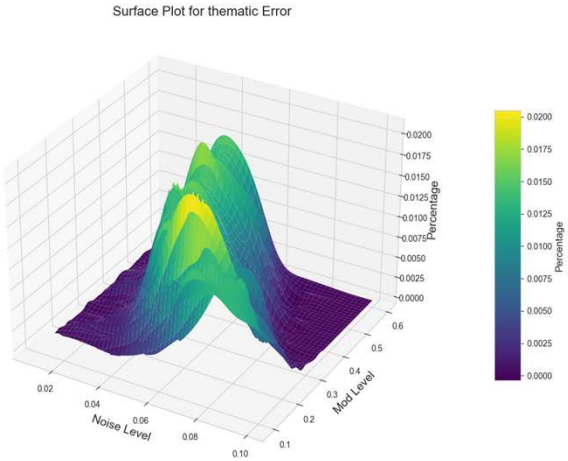
Beyond a noise threshold of 0.11, "nonword" errors dominate, indicating significant disruption in linguistic coherence.



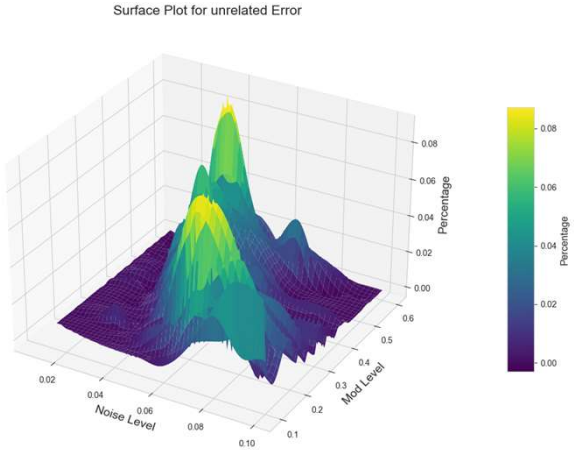
F: Phonological Errors Distribution



D: Taxonomic Errors Distribution



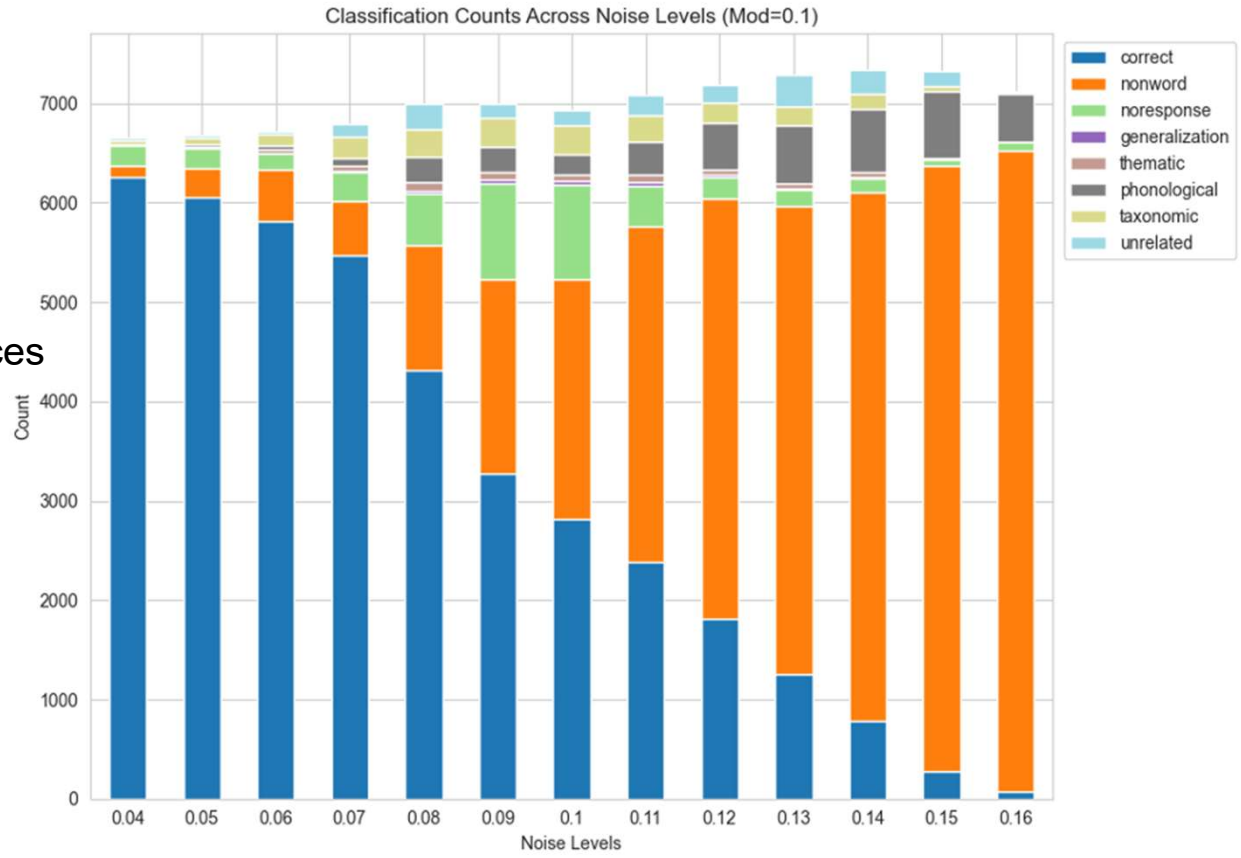
E: Thematic Errors Distribution



H: Unrelated Errors Distribution

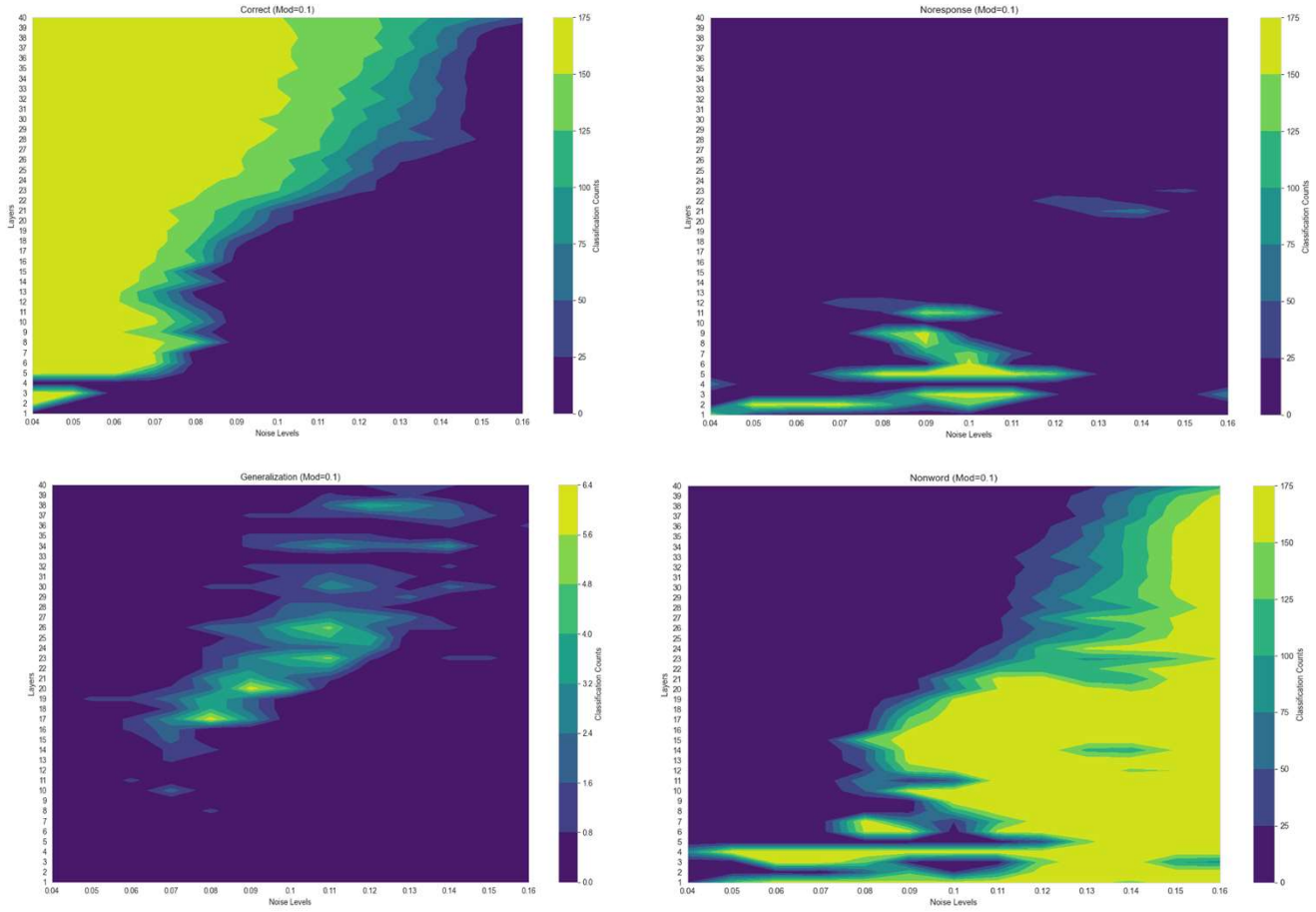
RESULTS - ERROR TRENDS - FIXED WEIGHT MODIFICATION (10%)

40 layers x
166 images
≈ 6500 Inferences



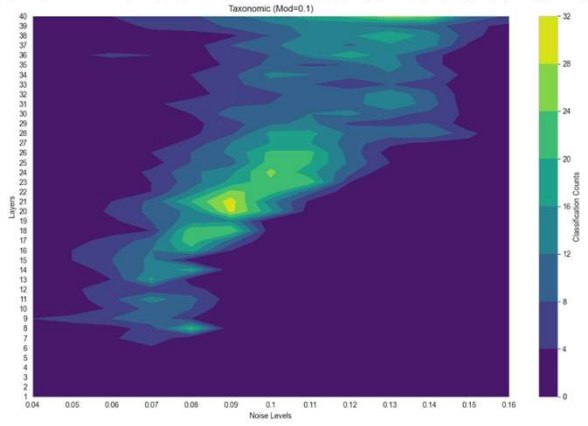
UNIVERSITY OF South Carolina

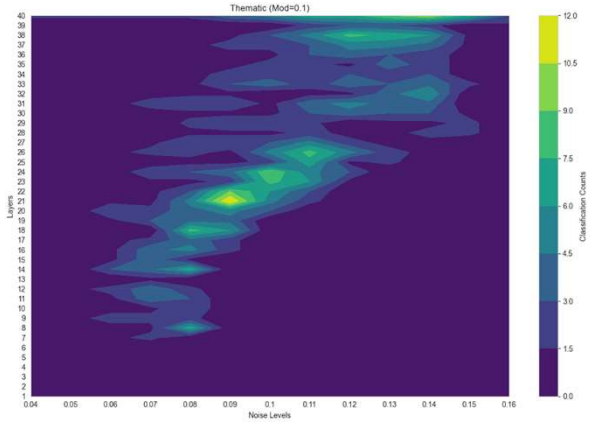
RESULTS – ADDING NOISE

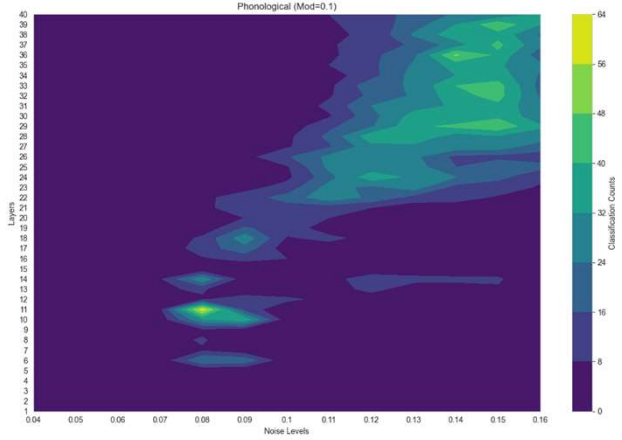


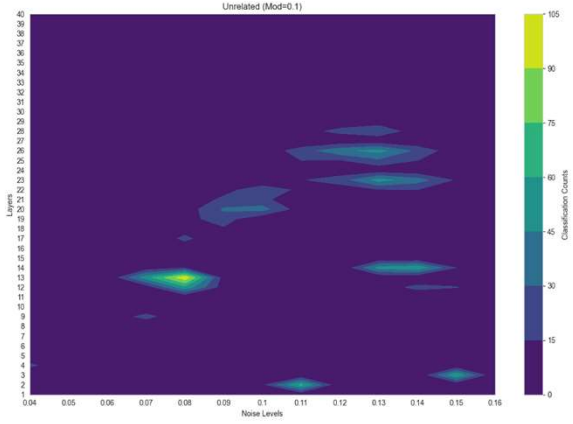
IVERSITY OF
outh Carolina

RESULTS – ADDING NOISE









ERSITY OF
uth Carolina

EXPERIMENTS-ZEROING

- With the weights set as zero, we also systematically modify one transformer layer at a time.

- Two experimental variables were tested: (1) Incremental zeroing, increasing the percentage of weights zeroed from 10% to 100% (2) skip-layer zeroing, randomly selecting layers 1, 2, 4, and 6 for zeroing

- We discovered that the first few layers have the most critical influence on the network's output, and the network's performance only significantly deteriorates when a large number of consecutive layers have their parameters zeroed out. This provided valuable insights into the internal mechanisms of neural networks.

RESULTS - ERROR TRENDS – ZEROING DIFFERENT NUMBER OF LAYERS

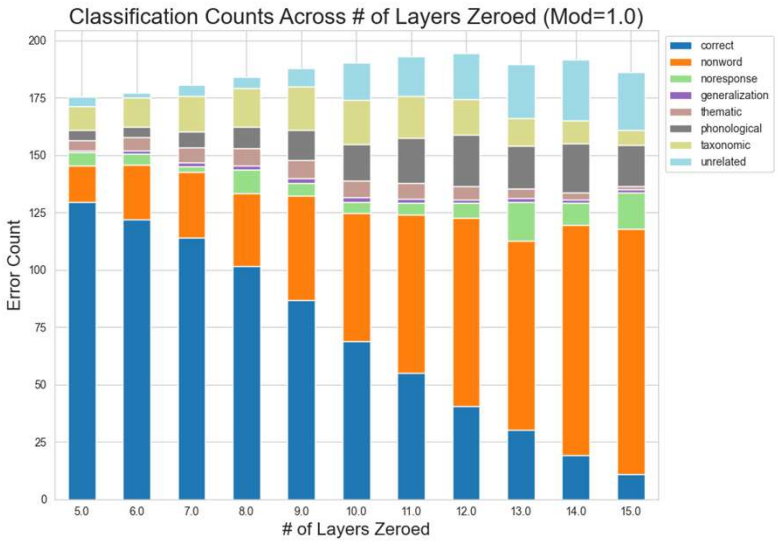
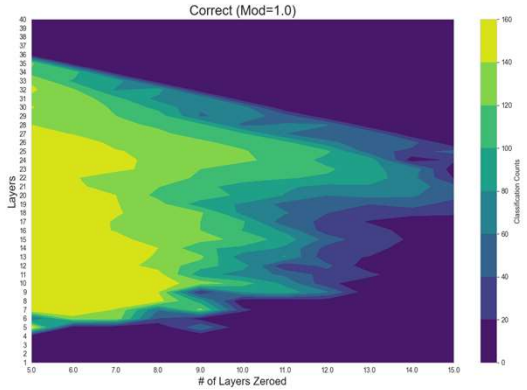


Figure 5: Distribution of Error under Modifying 10% of the Weights at Varying Noise Levels (x axis)
Note: Each column in the figure has been normalized by dividing by the number of zeroings.

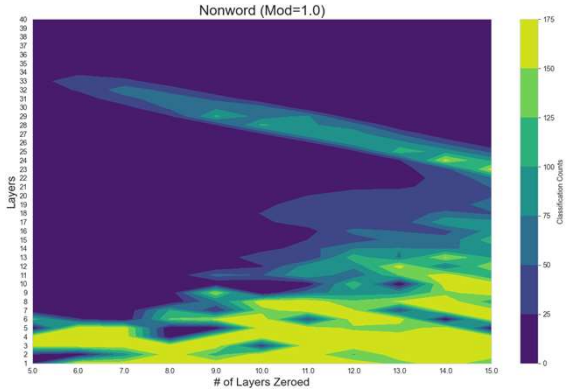
Overall number of output classifications at different number of layers zeroed. It shows we are getting a higher proportion of errors other than nonword and correct. This could be because we have a shallower LLM by zeroing multiple layers at a time.

UNIVERSITY OF South Carolina

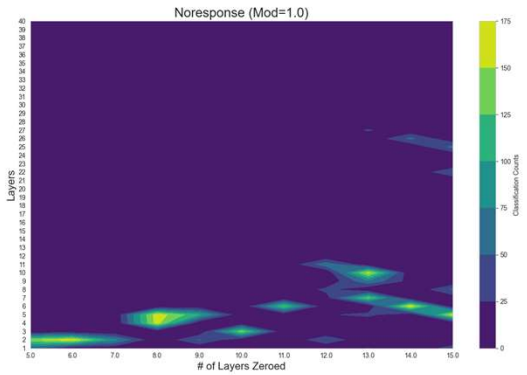
RESULTS - ZEROING



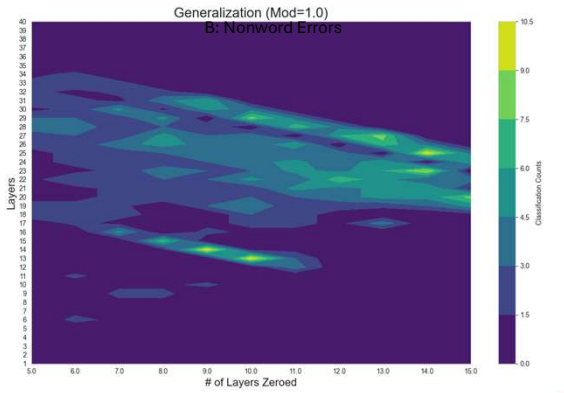
A: Correct Predictions



B: Nonword Errors



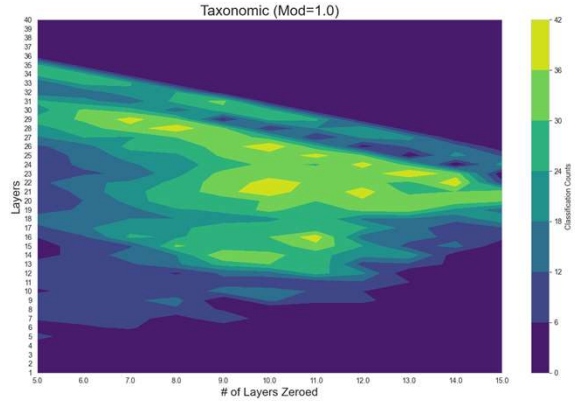
C: No-Response Errors



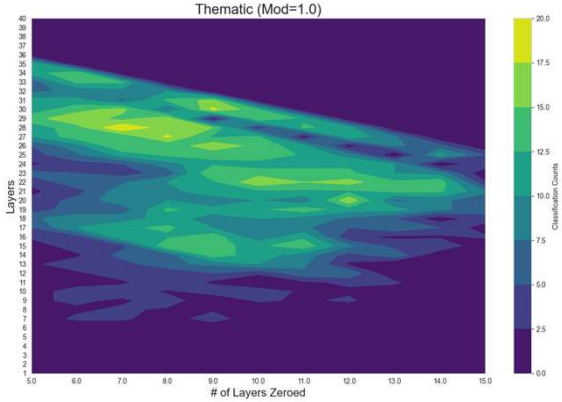
D: Taxonomic Errors

UNIVERSITY OF South Carolina

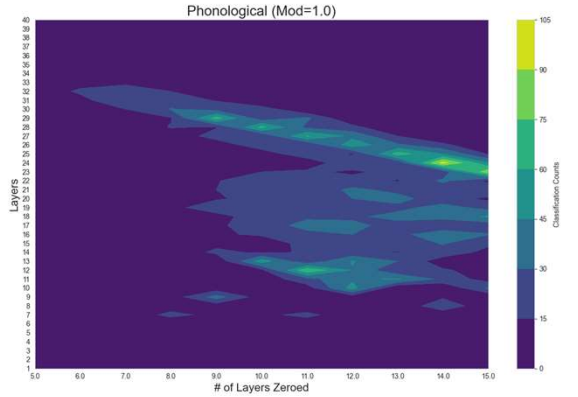
RESULTS - ZEROING



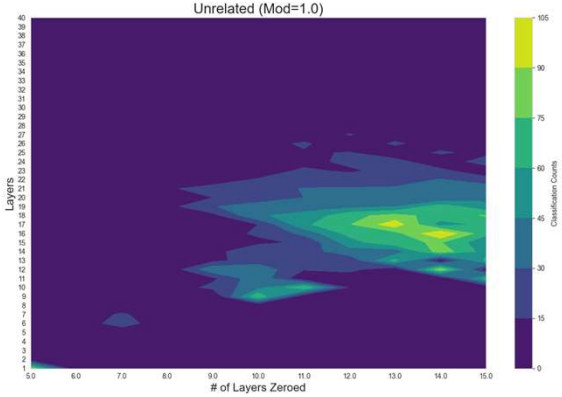
E: Thematic Errors



F: Phonological Errors



G: Generalization Errors



H: Unrelated Errors

UNIVERSITY OF South Carolina

BROADER IMPACT - LIMITATIONS

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.

Future Possible Improvement:

Complex Linguistic Tasks
Model Comparisons
Layer Sensitivity Thresholds
Neural Data Integration

REFERENCE

- 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
- Goodglass, H., & Kaplan, E. (1983). The Assessment of Aphasia and Related Disorders. Lea & Febiger.
- Katz, R. B. (1986). The Philadelphia Naming Test. In Clinical Aphasiology.
- MacWhinney, B., Fromm, D., Forbes, M., & Holland, A. (2011). AphasiaBank: Methods for studying discourse. Aphasiology, 25(11), 1286-1307.
- Perez, M., Sampath, A., Niu, M., & Mower Provost, E. (2024). Beyond Binary: Multiclass Paraphasia Detection with Generative Pretrained Transformers and End-to-End Models. *Proceedings of Interspeech 2024*. https://doi.org/10.21437/interspeech.2024-1281

UNIVERSITY OF South Carolina

SUPPLEMENTARY

- 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

UNIVERSITY OF South Carolina

THANKS!

Name
Title
Email
Social

UNIVERSITY OF
South Carolina