Finding Modularity in Large Language Models: Insights from Aphasia Simulations

Chengcheng Wang (chengcheng.wang@my.cityu.edu.hk)
Department of Linguistics and Translation, City University of Hong Kong
83 Tat Chee Avenue, Hong Kong

Jixing Li (jixingli@cityu.edu.hk)
Department of Linguistics and Translation, City University of Hong Kong
83 Tat Chee Avenue, Hong Kong

Abstract

Recent large language models (LLMs) excel at complex linguistic tasks and share computational principles with human language processing. However, it remains unclear whether their internal components specialize in distinct functions, such as semantic and syntactic processing, as seen in humans. To explore this, we selectively disrupted the components of LLM to replicate the behavioral patterns of aphasia—a disorder characterized by specific language deficits resulting from brain injury. Our experiments revealed that simulating semantic deficits akin to Wernicke's aphasia was relatively straightforward, whereas reproducing syntactic deficits characteristic of Broca's aphasia proved more challenging. These results highlight both parallels and divergences between the emergent modularity of LLMs and the human language system, offering new insights into information representation and processing in artificial and biological intelligence.

Keywords: aphasia; large language model; modularity of language processing

Introduction

Prior research on aphasia has revealed that language processing in the brain follows a modular organization (Dronkers & Ivanova, 2023). Aphasia, an acquired language disorder caused by brain damage, disrupts abilities such as language production, comprehension, or repetition (Goodglass, 1993). Different subtypes of aphasia have been identified, each associated with specific brain regions and unique linguistic impairments. Two well-known subtypes are Broca's aphasia (Broca, 1861) and Wernicke's aphasia (Wernicke, 1874). Broca's aphasia is marked by significant difficulties with syntax, particularly in constructing and understanding complex sentences. In contrast, Wernicke's aphasia is characterized by fluent but non-sensical speech, resulting from deficits in semantic processing. In addition to these two subtypes, there exists a more severe subtype known as Global aphasia. This condition is characterized by widespread impairments at the lexical, semantic, and syntactic levels (Kemmerer, 2022, ch.2).

Although extensive neuropsychological research supports the modular structure of the human language system, state-of-the-art large language models (LLMs), such as GPT-4 (OpenAI et al., 2023), DeepSeek-V3 (DeepSeek-AI et al., 2024) and LLaMA-3 (Grattafiori, Dubey, Jauhri, et al., 2024) are often treated as monolithic systems (Qiu, Huang, & Fu, 2024). In this work, we employ the open-source multimodal LLM Visual-Chinese-LLaMA-Alpaca (VisualCLA; Cui, Yang, & Yao, 2024; Yang, Pan, & Cui, 2023) to conduct a picture description task, a common diagnostic tool for assessing aphasia (Goodglass & Kaplan, 1983). We systematically disrupted individual layers, self-attention heads, and key parameters within VisualCLA's text model to replicate language deficits similar to those seen in human aphasia (see Figure 1 for an overview of our analysis pipeline).

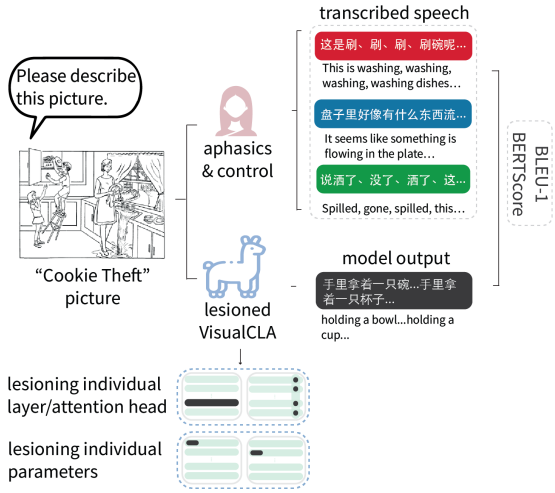

Figure 1: Overview of the analysis pipeline. Transcribed speech from the "Cookie Theft" picture description task was collected from aphasics and healthy control and compared to outputs from the lesioned VisualCLA model. The model was lesioned at the individual layer, self-attention head, and parameter levels. Model outputs were evaluated using BLEU-1 and BERTScore to quantify their similarity to aphasic speech.

Methods

Aphasia dataset

We used an existing aphasia dataset (Bi et al., 2015; Han et al., 2013) comprising a total of 51 aphasics (15 females, mean age=48.08±12.15 years, mean education levels=12.8±3.7 years) and 43 healthy controls (21 females, mean age=49.3±10.7 years, mean education level=13.7±3.8 years). All participants were right-handed native Mandarin speakers. Based on the Aphasia Battery of Chinese (Gao

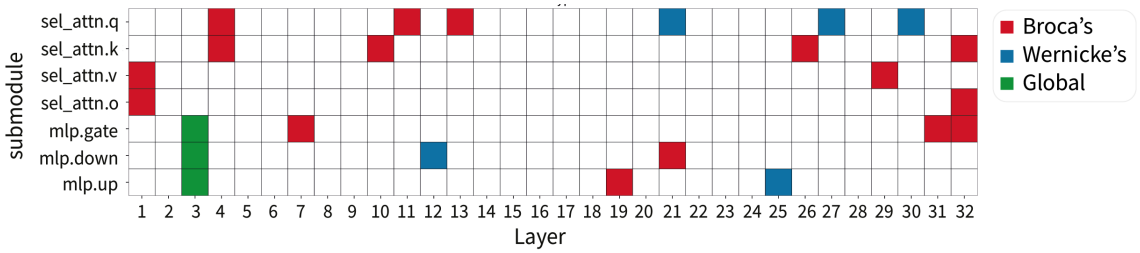

Figure 2: Distribution of parameters identified as critical for simulating behaviors associated with different aphasia subtypes. Each square represents the top 1% of parameters within a submodule that exhibited the greatest gradient change during fine-tuning.

et al., 1993), the 51 patients were further categorized into 16 cases of Broca's aphasia, 11 cases of Wernicke's aphasia, 24 cases of Global aphasia. We analyzed the participants' behavioral outputs from a picture description task, in which participants viewed the black-and-white "Cookie Theft" image (see Figure 1) from the Boston Diagnostic Aphasia Examination (BDAE; Goodglass & Kaplan, 1983) and provided a verbal description of its contents.

Simulating aphasic behavior by lesioning model components

To simulate aphasic behaviors, we systematically disrupted specific components of the text model in VisualCLA, targeting individual layers, attention heads, or parameters within specific submodules. The text model of VisualCLA comprises 32 layers (excluding the embedding layer), each containing 32 self-attention heads. We systematically disrupted individual layers or attention heads and analyzed their effects on the model's performance during the "Cookie Theft" picture description task.

In addition to lesioning individual model layers and self-attention heads, we also disrupted individual parameters within each submodule of the text model of VisualCLA. Specifically, we fine-tuned the model using outputs from the control group and evaluated the relative impact of each parameter by analyzing the magnitude of their gradient changes, following the methodology described by Z. Zhang, Zhao, Zhang, Gui, and Huang (2024). We identified the top 1% of parameters for each of the 224 submodules. Each of these top-performing parameters was lesioned, and the model's outputs were collected for the "Cookie Theft" picture description task. We evaluated whether these lesioned models exhibited language deficits resembling recognized aphasia subtypes using BLEU-1 (Papineni, Roukos, Ward, & Zhu, 2002) and BERTScore (T. Zhang, Kishore, Wu, Weinberger, & Artzi, 2020).

Results

Lesioning individual layers and self-attention heads

Lesioning either a single layer or a self-attention head resulted in output more similar to Wernicke's aphasia. The mean BLUE-1 (0.09 ± 0.05) and BERTScore (0.63 ± 0.03) comparing the lesioned models and Wernicke's outputs averaged over all layers were significantly higher than those for Broca's (BLEU-1: 0.06 ± 0.03; BERTScore: 0.60 ± 0.02) and Global aphasia (BLEU-1: 0.08 ± 0.02; BERTScore: 0.57 ± 0.01).

Lesioning individual parameters

Out of the 224 submodules, disrupting the top 1% of high-impact parameters in 16 submodules generated deficits akin to Broca's aphasia, while disrupting 5 submodules aligned with Wernicke's aphasia and disrupting 3 submodules generated outputs similar to Global aphasia (see Figure 2). A larger number of submodules required for a given aphasia subtype suggests greater difficulty in reproducing that specific deficit, as more parameters in the model needed to be disrupted.

In contrast to the large lesion size typically associated with Global aphasia in human cases, we found that it was the easiest to simulate in lesioned models, requiring disruption of only the top 1% of parameters from three submodules. A closer analysis of the model's output suggests that these parameters may play a critical role in encoding Chinese characters, as their removal led to the generation of random symbols such as 't-000}'. On the other hand, Broca's aphasia, the most prevalent aphasia subtype in human cases, proved the most difficult to replicate in lesioned models, requiring the disruption of parameters across 16 submodules. Wernicke's aphasia was also easier to simulate, requiring lesioning parameters from only 5 submodules.

Discussion and Conclusion

In this study, we systematically disrupted components of a LLM and compared the resulting behavioral deficits to those observed in Broca's, Wernicke's and Global aphasia. We found that lesioning specific components of LLMs could replicate behaviors characteristic of different aphasia subtypes. However, while semantic deficits as seen in Wernicke's aphasia were relatively straightforward to simulate, syntactic impairment characteristic of Broca's aphasia was more challenging to replicate. These results highlight differences in how information is represented and processed within LLMs, as well as the training objectives that guide their language task performance.

Acknowledgments

We thank Yanchao Bi and Zaizhu Han for providing the aphasia dataset, and we thank Zhiyu Fan for conducting an initial analyses of the human behavioral data.

References

Bi, Y., Han, Z., Zhong, S., Ma, Y., Gong, G., Huang, R., ... Caramazza, A. (2015, April). The white matter structural network underlying human tool use and tool understanding. *The Journal of Neuroscience*, *35*(17), 6822–6835. doi: 10.1523/JNEUROSCI.3709-14.2015

Broca, P. (1861). Remarques Sur le Siége de la Faculté Du Langage Articulé, Suivies D'une Observation D'aphémie (Perte de la Parole). *Bull Soc Anat*, *6*, 330–357.

Cui, Y., Yang, Z., & Yao, X. (2024, February). *Efficient and effective text encoding for Chinese LLaMA and Alpaca* (No. arXiv:2304.08177). arXiv. doi: 10.48550/arXiv.2304.08177

DeepSeek-AI, et al. (2024, December). *DeepSeek-V3 Technical Report* (No. arXiv:2412.19437). arXiv. doi: 10.48550/arXiv.2412.19437

Dronkers, N. F., & Ivanova, M. V. (2023). The neuroscience of language and aphasia. In *APA handbook of neuropsychology: Neuroscience and neuromethods, Vol. 2* (pp. 139–158). Washington, DC, US: American Psychological Association. doi: 10.1037/0000308-007

Gao, S. R., Wang, Y., Shi, S., Liu, J., Lin, G., & Rao, B. (1993). Aphasia. *Beijing Medicine University and China Xiehe Medicine University Joint Press, Beijing*.

Goodglass, H. (1993). *Understanding aphasia*. San Diego, CA, US: Academic Press.

Goodglass, H., & Kaplan, E. (1983). *Boston diagnostic aphasia examination booklet*. Lea & Febiger.

Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024, November). *The Llama 3 herd of models* (No. arXiv:2407.21783). arXiv. doi: 10.48550/arXiv.2407.21783

Han, Z., Ma, Y., Gong, G., He, Y., Caramazza, A., & Bi, Y. (2013, October). White matter structural connectivity underlying semantic processing: Evidence from brain damaged patients. *Brain*, *136*(10), 2952–2965. doi: 10.1093/brain/awt205

Kemmerer, D. (2022). *Cognitive Neuroscience of Language* (2nd ed.). New York: Routledge. doi: 10.4324/9781138318427

OpenAI, et al. (2023, December). *GPT-4 technical report* (No. arXiv:2303.08774). arXiv. doi: 10.48550/arXiv.2303.08774

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. doi: 10.3115/1073083.1073135

Qiu, Z., Huang, Z., & Fu, J. (2024, June). Unlocking Emergent Modularity in Large Language Models. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 2638–2660). Mexico City, Mexico: Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.144

Wernicke, C. (1874). *Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert.

Yang, Z., Pan, Y., & Cui, Y. (2023). *Visual-Chinese-LLaMA-Alpaca.* GitHub.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020, April). BERTScore: Evaluating Text Generation with BERT. In *Eighth International Conference on Learning Representations.*

Zhang, Z., Zhao, J., Zhang, Q., Gui, T., & Huang, X. (2024). Unveiling Linguistic Regions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6228–6247). Bangkok, Thailand: Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.338