Simulating Aphasia through Weight Modifications in Pretrained Large Language Models

Comments:

1. Is there evidence that damaging large language models can produce aphasia-like responses?

   This is the key question that we aim to investigate. There is no existing research on whether this is possible or not, making this project novel and compelling.

2. How does simulating aphasia in large language model relate to the actual brain? There is no direct physical model to link between the two. Both have been used as tools to study the other.
3. What are the main contributions and deliverables for this project?
   This project aims to explore a research question. Our goal is to follow scientific methods, even if our findings show no direct correlation, negative results can still provide valuable insights.

Introduction:

Aphasia is a communication disability that affects around 2 million people in the United States. Aphasia typically occurs following brain injuries such as stroke, when it damages brain areas related to language production or understanding. The goal of this project is to examine whether it's feasible to simulate aphasia on a computer representation of the brain, by manipulating pretrained large language models (LLM). If successful, it can lead to better understanding of aphasia, LLM, and their relations. Which could offer new perspective on developing better rehabilitation strategies and treatments for aphasia.

Methodology:

To begin, we will select an LLM to represent the 'healthy' mind. We will input some images and verify that the LLM is able to correctly caption the images, to serve a baseline for normal language. Given aphasia is generally caused through external damage to the brain, we want to simulate this by also 'damaging' the LLM. We conceptualize the layers of the LLM as analogous to different regions of the brain, and we will simulate the damage by modifying the model's internal weights. We hypothesize that these disruptions could create aphasia-like outputs.

For input data, we can use images from standard clinical test banks such as the "Philadelphia Naming Test", which is frequently used to assess language function in aphasia patients. For LLM, we found serval open models: Flamingo, MiniGPT-4, BLIP-2,

and LLaVA. We can start out with simpler models to speed up the prototyping. Then we will select one based on the performance in the image captioning test.

For the inference output, we will develop a classifier for classifying into non-aphasia, non-fluent aphasia, and fluent aphasia. We can use AphasiaBank and C-STAR databases as training data. We will search for classification methods in existing literature.

Evaluation:

The success of this project is based on how well we can adjust the weights of the LLM to simulate aphasia-like behavior. Initially we will do some random search on which layers of LLM to change, and the magnitude of the changes. Our methods consist of adding noise and zeroing out weights to simulate a disruption in neural communication. Ideally, if we can find a gradient, then we use it to adjust; otherwise, changes would be done randomly.

Literature Review:

To the best of our knowledge, we did not find similar work that used specific approach of 'damaging' pretrained LLM to simulate aphasia. This could be because open LLM models are relatively new. Prior works have focused primarily on using LLM to detect aphasic speech and text.

Reference:

Li, Y., Zeng, C., Zhong, J., Zhang, R., Zhang, M., & Zou, L. (2024). Leveraging large language model as simulated patients for clinical education. *Wangxuan Institute of Computer Technology, Peking University; School of Computer Science, Wuhan University; CureFun Co.*

Fergadiotis, G., Gorman, K., & Bedrick, S. (2016). Algorithmic classification of five characteristic types of paraphasias. *American Journal of Speech-Language Pathology*. DOI: 10.1044/2016_AJSLP-15-0147

M. Perez, A. Sampath, M. Niu, and Emily Mower Provost, "Beyond Binary: Multiclass Paraphasia Detection with Generative Pretrained Transformers and End-to-End Models," pp. 4119–4123, Sep. 2024, .