Table of Contents

Research Week 2023

Large Language Model Prompting Strategies for Automated Scoring of Aphasia Assessments: An Empirical Study

Steven Bedrick, Alexandra Salem, Robert Gale, Gerasimos Fergadiotis
Department of Medical Informatics and Clinical Epidemiology

Keywords

artificial intelligence, informatics, aphasia, stroke, natural language processing

Abstract

Anomia is the inability to access and retrieve the intended words during language production, and is a cardinal feature of the acquired neurogenic language disorder known as aphasia. Aphasia affects 2.5-4 million people in the US and, given the aging trend in the population, the incidence of aphasia will increase in the coming decades. Communication difficulties have a significant impact on the health-related quality of life of people with aphasia (PWA), and are associated with substantial healthcare costs. As such, diagnosis and assessment of word retrieval is essential. Fortunately, there exist robust and validated instruments for characterizing word retrieval ability; unfortunately, many of these assessments require significant manual labor to administer, and are difficult to score in an objective. Our previous work has focused on automating the scoring of certain aspects of confrontation naming tests using computational methods, a task that we were largely able to accomplish but with some important limitations.

Recent advances in machine learning have led to a new class of large language model (LLM) and a new approach to natural language processing. Instead of a classifier trained using supervised learning, modern LLMs incorporate textual descriptions of their desired task into their input, and produce textual output in a generative manner. This provides researchers with a dramatic *increase* flexibility, and a dramatic *decrease* in the amount of training data required for many applications. These benefits come at the cost of complexity and ambiguity, as much depends on specific details of how input is formulated (i.e., how the model is "prompted"), as well as risks in terms of inconsistent output. In this work, we applied two modern LLMs (one commercial, one open) to the classical task of scoring a confrontation naming test, and empirically validate several different prompting strategies in terms of their resulting classification accuracy and consistency of output across multiple repeated runs. We found significant differences between models, both in terms of their performance as well as in terms of which prompting strategy proved most effective. Our findings may be generalizable to other neurospsychological and linguistic assessments that contain a semantic aspect.