DISCOURSE
Chronicle - July 2023

The Promise and Limitations of Using Large Language Models to Understand Language in Psychosis

Gina Kuperberg MD PhD Dennett Stibel Professor of Cognitive Science, Department of Psychology, Tufts University, and Psychiatrist Mass. General Hospital

Large Language Models (LLMs), such as OpenAI's pioneering GPT series, have taken the world by storm. Many of us now routinely use them in our everyday activities, from drafting e-mails to automating administrative tasks.
In this blog, I will argue that these models have tremendous potential for understanding the nature of language and communication impairments in psychosis. However, I also caution that they have important limitations, and for a deeper understanding of language in psychosis, we will need to use them in conjunction with other types of computational models that are more biologically plausible.

The Promise of LLMs

LLMs are essentially prediction machines. They are exposed to vast volumes of text and trained to predict each upcoming word based on the full context that preceded it. By being "rewarded" for predicting accurately, they implicitly learn complex statistical patterns within the data, enabling them to produce language that closely resembles human speech.

The human brain can also be viewed as a prediction machine. We've known for many years that the predictability of each word in a sentence is one of the most robust predictors of behavior and neural activity. Therefore, it's perhaps not so surprising that LLMs are capable of generating such remarkably human-like language.
Understanding the role of prediction in language is critically important for researchers studying "positive thought disorder" in psychosis. By definition, positive thought disorder is characterized by disorganized language that lacks coherence. We have long suspected that it may stem from the difficulties patients have in rapidly using prior context (what they previously said) to anticipate what they should say next. However, to effectively test this hypothesis, we needed a method of quantifying the predictability of every single word in hundreds of speech samples. Until recently, this was not feasible. This all changed with the advent of GPT and other LLMs.

In recent work, we provided GPT-3 with each word in speech samples produced by a large sample of people with first-episode schizophrenia. As we theorized, the words produced by the patients were indeed less predictable (i.e., more surprising) than those produced by the control participants. Most notably, by manipulating the amount of context available to the model, we demonstrated that this effect originated specifically from an impairment in using global versus local context. We also showed that this global-local deficit selectively predicted clinical ratings of positive thought disorder.

This serves as just one illustration of how we can employ LLMs to test cognitively-informed hypotheses to deepen our understanding of language use in psychosis. And it's just the beginning; there are numerous other avenues to explore, from examining the predictability of words in natural conversation to using advanced methods such as topic modeling to better understand the structure of discourse produced by individuals with schizophrenia. In addition, LLMs also offer a powerful instrument for quantifying how patients use context to predict upcoming words during naturalistic language comprehension — a significantly understudied area of research.

The Limitations

Despite the potential of LLMs to yield insights into language in psychosis, it is also crucial to appreciate their limitations, especially with regard to their biological plausibility.

There are some parallels between the architecture of GPT and the human brain. For example, like the human cortex, LLMs are comprised of multiple "layers." Some of these layers can even learn complex representations that allow them to generate predictions based on lengthy prior contexts. In GPT, this is facilitated by a so-called "attention" mechanism that enables the model to assign varying degrees of importance to specific portions of the input sequence as it generates these predictions.

However, in models like GPT, the way these layers are connected together is quite different from how layers of the human cortex are connected. In GPT, the connections between layers are strictly "feedforward." In other words, there is no mechanism for a prediction generated at a higher layer to influence activity at a lower layer before the user inputs the next word. This contrast with the human brain, where feedback connections are omnipresent, playing a vital role in enabling higher levels of the cortex to actually pre-activate information at lower layers ahead of new information reaching these layers. In healthy individuals, this feedback allows lower levels of the cortex to gain a "head-start" in processing, enabling the brain to keep up with the rapid pace of real-time communication.

In schizophrenia, predictive processing is most likely to falter under time pressure, leading researchers to theorize that impairments in predictive language processing may stem from a disruption of feedback connectivity. GPT, however, cannot be used to directly test this hypothesis. Instead, we need a more biologically plausible model that incorporates feedback connections.

One such model is predictive coding, in which the generation of top-down predictions through feedback connections is a key step in the computational algorithm. We have shown that, in healthy adults, predictive coding can explain the brain's sensitivity to contextual predictability, its neural dynamics, and its sensitivity to various lexical variables, priming, and their interactions. The next step in our research is to conduct simulations in which we disrupt these feedback connections to determine whether this can explain the predictive deficits we see in schizophrenia.

Predicting the Future

Of course, we don't have to choose between using LLMs or predictive coding models to study language in psychosis. I believe that the best way forward is to leverage the strengths of each approach while acknowledging their respective limitations. LLMs enable us to systematically understand the nature of predictive language abnormalities in psychosis, to quantify these abnormalities, and to ask how they impact patients' struggles with day-to-day language and communication. However, their biological implausibility, and their black box nature mean that they cannot be used to understand the underlying causes of these deficits. Models like predictive coding models provide greater biological plausibility, transparency, and interpretability. Therefore, we can use them to test specific hypotheses about the underlying causes of these predictive abnormalities. By combining these two approaches, we should be able to attain a deeper and more comprehensive understanding of thought disorder and social communication in schizophrenia.

Email: Gina.Kuperberg@tufts.edu or GKuperberg@mgh.harvard.edu NeuroCognition Lab website: https://projects.iq.harvard.edu/kuperberglab

DISCOURSE IN PSYCHOSIS