GPT Reveals Selective Impairments in Global *vs.* Local Context Use in Speech among Treatment-Naïve Patients with Positive Thought Disorder

Gina Kuperberg1,4, Tori Sharpe1, Sabrina Ford2, Samer Nour-Eddine1, Lin Wang,1,4, Lena Palaniyappan2,3

1Tufts University, Department of Psychology, Medford, Massachusetts, United States

2Robarts Research Institute, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada

3Douglas Mental Health University Institute, Department of Psychiatry, McGill University, Montreal, Quebec, Canada

4 Department of Psychiatry and the Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, United States

Background: The advent of large language models is revolutionizing psychosis research, offering excellent discriminatory capabilities to distinguish between patients and control participants. However, they have not yet been used to quantify or understand the most prominent language atypicality in schizophrenia—disorganized language production, i.e., positive thought disorder. Psychopathologists have long characterized positive thought disorder as relying on relationships between individual words (local context) at the expense of broader discourse coherence (global context). Until now, however, there has been no way to objectively and automatically characterize the relationship between words and their prior context in natural speech. Instead, to assess thought disorder, clinicians and researchers are reliant on subjective, time-consuming rating scales. Here, we use GPT-3 to quantify the influence of global vs. local context on each word in speech samples produced by a large sample of treatment-naïve, first-episode psychosis patients. This allowed us to determine (a) the extent to which each word's surprisal relies on global vs. local context in patients vs. controls, and (b) whether the degree of selective dependence on local vs. global context specifically predicts the severity of positive thought disorder.

Methods: Seventy first-episode psychosis patients (all unmedicated and treatment-naïve) and 36 demographically-matched control participants described three pictures, each for approximately one minute. After speech transcription, we used GPT-3 to extract the probability of each word while manipulating the amount of context the model had access to. The surprisal (i.e., negative log probability) of each word, based on these different context lengths, served as the dependent measure in a series of linear mixed-effects models with which we tested our primary hypotheses. Thought disorder was assessed using the Thought and Language Index (TLI). Symptoms more generally were assessed using the PANSS. Finally, domain-general cognitive function was assessed using Semantic Fluency, Digit-Symbol Substitution, and the Trail-Making Test.

Results: We began by comparing the surprisal of each word using all available context to its surprisal with no context (estimated by replacing prior words with random words from an unrelated speech sample). This revealed a significant interaction between Context and Group

(Est. = .322, p = .001), indicating that patients were less able than controls to use the available prior context to reduce the surprisal of each word they produced. We then asked the key question of whether this was driven by a selective impairment in using global vs. local context by titrating the window of context GPT had access to (from 1 to 50 prior words). We found a significant interaction between Log Context Length and Group (Est. = .086, p < .001), such that, as context length increased, the differences in surprisal between patients and controls became increasingly larger, indicating a disproportionate deficit in the use of global context in schizophrenia. This effect could not be explained by atypical domain-general cognitive function in patients. Most importantly, within the patient group, graded insensitivity to global-vs.-local context predicted the severity of positive thought disorder (assessed via the Disorganization subscore of the TLI). This effect was specific: There was no evidence of a relationship with overall symptom severity (PANSS total) or negative thought disorder (TLI Impoverishment subscore).

Conclusion: We show, for the first time, that global-vs.-local surprisal selectively predicts positive thought disorder in first-episode schizophrenia. This has several important implications. First, from a clinical perspective, we provide a measure that could be developed into a sensitive linguistic biomarker for fast, automated, and objective quantification of language disorganization. Such a biomarker could facilitate early detection of illness, symptom monitoring, prediction of outcome, and possibly the trajectory of thought disorder over time. It could also potentially provide a sensitive measure for detecting more subtle, subclinical atypicalities in communication that might impair psychosocial functioning in schizophrenia. Second, from a neurocognitive perspective, these findings directly bridge clinical characterizations of thought disorder in natural speech with neurocognitive evidence for selective deficits in the processing of global vs. local information in language comprehension, as well as in other perceptual and cognitive domains. Third, from a neurocomputational perspective, these findings are consistent with hierarchical generative models of psychosis. These theories posit that uncertainty over global representations, represented over longer time-scales at the highest levels of the cortical hierarchy, results in weaker predictions being propagated down to lower cortical levels, representing individual words, leading to reduced suppression of word-level surprisal (prediction error). To directly test this hypothesis, we need to move beyond GPT, which lacks the feedback connections that drive healthy language processing in the brain. We are therefore developing a model that is based on a more biologically plausible *predictive coding* architecture, which will allow us to explicitly simulate the effects of perturbed feedback on global vs. local surprisal.