LLM-Based Surprisal Metrics as Potential Predictors of Cognitive and Mental Health

Nadra Salman, Sarah C. Wilson, Sarah Newman-Norlund, Nicholas Riccardi, Amit Almor, Leonardo Bonilha, Dirk Den Ouden

University of South Carolina, Columbia, SC, USA

Introduction. Lexical surprisal, a measure of how surprising a word is within its linguistic context, offers valuable insight into cognitive effort during language processing. Surprisal theory posits that less predictable words demand greater cognitive effort, but high surprisal rates have also been linked to reduced semantic coherence, as evidenced by studies linking high unpredictability in speech to psychosis and schizophrenia [1, 2]. While surprisal has been extensively studied in language comprehension using methods like EEG [3], recent advances in large language models (LLMs) have enabled scalable, objective analyses of language production. Surprisal can be efficiently calculated using LLMs like LLaMA-2 [2, 4]. LLMs provide critical insight into discourse patterns, aiding in the classification of conditions such as aphasia, and predicting syndromes in at-risk populations [5]. This study investigates whether LLM-based lexical surprisal in spoken discourse is associated with cognitive performance and mental health (from self-ratings of anxiety and depression) in healthy aging adults.

Methods. This study analyzed data from 181 healthy native English-speaking adults (ages 20-80) who completed the Cat Rescue picture description task as part of the Aging Brain Cohort's protocol [6]. Using LLaMA-2, token-level log probabilities were computed based on preceding context, with multi-token words aggregated using conditional probability rules. Lexical surprisal metrics were derived for each sample by negating the log-probabilities and computing mean, variance, and slope. Mean surprisal reflects overall word predictability; variance captures fluctuations in predictability; and slope indicates changes in predictability over the course of the sample. These metrics were then correlated with demographic data, total and index sub scores on the Montreal Cognitive Assessment (MoCA) [7], and self-reported depression and anxiety scores from the PROMIS surveys. Linear regression analyses included age and story length to isolate the effects of surprisal metrics on these outcome measures.

Results. We ran multiple linear regression analyses and found that surprisal was a significant negative predictor of both MOCA Total Score ($\beta = -0.155$, t = -2.949, p = 0.003) and MOCA Language Index ($\beta = -0.159$, t = -2.767, p = 0.006), indicating that higher surprisal is associated with lower cognitive performance. Surprisal slope was a significant negative predictor of PROMIS Anxiety ($\beta = -0.118$, t = -2.276, p = 0.023), suggesting that reduced variation in surprisal across linguistic input is linked to higher anxiety symptoms (see Figure 1). Surprisal variance was not a significant predictor of any outcome.

Conclusion. Our findings show that surprisal metrics derived from discourse can serve as markers of cognitive performance and mental health in healthy aging adults. Higher overall surprisal (mean) was associated with lower MoCA scores, supporting the notion that lower cognitive performance is reflected in reduced semantic coherence. Variance in surprisal was not significantly associated with cognitive scores or mental health factors. However, surprisal slope, which captures the direction and rate of change in surprisal over discourse, was negatively associated with anxiety. This suggests that individuals whose discourse showed consistent increase or decrease in surprisal over time tended to report lower anxiety levels. These findings highlight the utility of LLMs in providing scalable, non-invasive tools for assessing cognitive and mental health, with implications for psycholinguistic research and clinical applications.

References

[1] Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., Bearden, C. E., & Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry, 17(1), 67 75. https://doi.org/10.1002/wps.20491

[2] Sharpe, V. P. (2023). *Taking it with a grain of salt and pepper: Spared and impaired use of contextual information in the language production of people with schizophrenia* [Doctoral dissertation, Tufts University]. ProQuest Dissertations & Theses Global. https://www.proquest.com/dissertations-theses/taking-with-grain-salt-pepper-spared-impaired-use/docview/2864782092/se-2

[3] Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language, 115*(3), 149-161. https://doi.org/10.1016/j.bandl.2010.07.006

[4] Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences, 121*(10). https://doi.org/10.1073/pnas.2307876121

[5] Bedi, G., Carrillo, F., Cecchi, G. A., Slezak, D. F., Sigman, M., Mota, N. B., Ribeiro, S., Javitt, D. C., Copelli, M. & Corcoran, C. M. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia, 1*(1), 1-7. https://doi.org/10.1038/npjschz.2015.30

[6] Newman-Norlund, R. D., Newman-Norlund, S. E., Sayers, S., Nemati, S., Riccardi, N., Rorden, C., & Fridriksson, J. (2021). The Aging Brain Cohort (ABC) repository: The University of South Carolina's multimodal lifespan database for studying the relationship between the brain, cognition, genetics and behavior in healthy aging. *Neuroimage: Reports, 1*(1), 100008. https://doi.org/10.1016/j.ynirp.2021.100008

[7] Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*(4), 695-699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

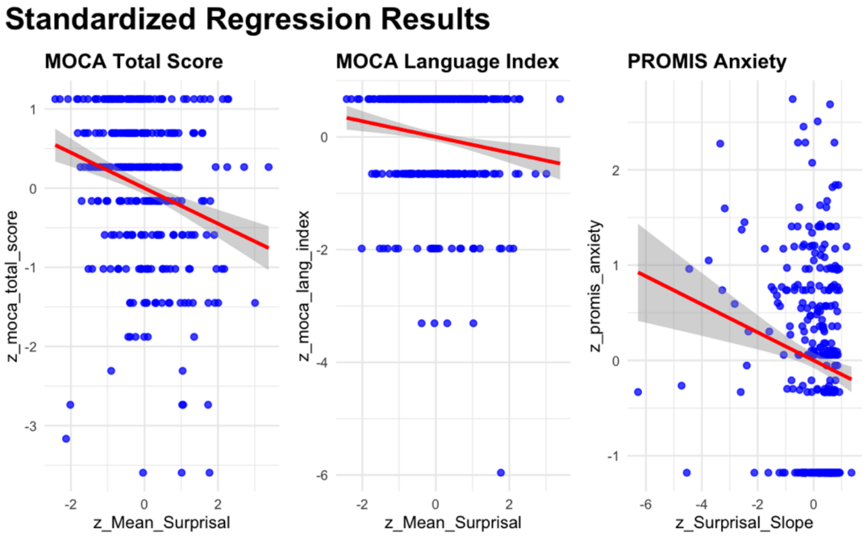

Figure 1: Standardized Regression Results. The left panel shows the association between mean surprisal and MOCA Total Score, the middle panel shows mean surprisal and MOCA Language Index, and the right panel shows surprisal slope and PROMIS Anxiety. Red lines indicate regression fits with 95% confidence intervals (gray shading).