On the Limits of LLM Surprisal as a functional Explanation of ERPs

Benedict Krieger (Saarland University), Harm Brouwer (Tilburg University),
Christoph Aurnhammer (Saarland University), Matthew W. Crocker (Saarland University)
bkrieger@lst.uni-saarland.de

The impressive comprehension-like behavior of LLMs trained on next word prediction has led researchers to suggest that these models are to some extent accurate models of human comprehension (e.g., Goldstein et al., 2022; Schrimpf et al., 2021). Studies correlating LLM-derived surprisal and neural correlates have focused predominantly on the N400 - an event-related potential (ERP) component sensitive to the expectancy of a word in context - in naturalistic comprehension (De Varda et al., 2023; Michaelov et al., 2024). Experiments, however, show that beyond expectancy, the N400 is also sensitive to semantic association, defined as the extent to which word meaning is primed by its prior context (see Kutas & Federmeier, 2011). While LLMs are also sensitive to association (Michaelov & Bergen, 2022), the influence of expectancy on the N400 can be overridden entirely when target word meaning is contextually primed, such that semantically unexpected words do not increase N400 amplitude (e.g., Nieuwland & Van Berkum, 2005 and Delogu et al., 2019., shown in Fig. 1). While these words were clearly surprising to humans, as reflected in increased P600 amplitude, it is unclear how LLMs perform in these cases. Moreover, the P600 - which has received little attention in this line of research - has been found to be graded for plausibility and insensitive to association (Aurnhammer et al., 2023; Aurnhammer et al., 2021; Brouwer et al., 2021). We examine the ability of LLM surprisal to model three German ERP-studies that specifically sought to disentangle the influence of expectancy, plausibility, and association on both the N400 and P600 (Aurnhammer et al., 2023; Aurnhammer et al., 2021; Delogu et al., 2019). Using two transformer models, a smaller model (GPT-2) and a larger state-of-the-art model (LeoLM), we replicated the sensitivity of LLMs to both expectancy and association. However, results from an rERP analysis (Smith & Kutas, 2015) using LLM-derived surprisal to re-estimate ERPs led to mixed results: Surprisal collected with the larger LLM predicted an N400 difference that was unobserved (in Delogu et al. 2019, see Fig. 1, right panel), while surprisal collected with the smaller LLM did not predict such a difference – in line with the observed ERP profile, but revealing sensitivity of surprisal towards association. Furthermore, the magnitude of effects was underestimated. For the P600, LLMs were able to capture violations of selectional restrictions, but failed to account for the graded sensitivity of the P600 to plausibility (Aurnhammer et al., 2023). If LLMs are indeed an accurate characterisation of (aspects) of human comprehension mechanisms, they should account for N400 and P600 effects and their differential sensitivity to association, expectancy and plausibility. Our findings suggest that LLM surprisal may not offer an accurate characterisation of the underlying functional generators of either the N400 or P600, and motivate exploring alternative LLM-derived linking hypotheses to the N400 and P600 informed by mechanistic accounts of the processes associated with these components (Brouwer et al., 2017; Fitz & Chang, 2019; Li & Futrell, 2023; Li & Ettinger, 2023). We argue that until LLMs are shown to account for critical data points through such linking hypotheses, strong conclusions about their validity as models of the human comprehension system (e.g., Goldstein et al., 2022; Schrimpf et al, 2021) are too premature.

References

Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. W. (2023). The P600 as a continuous index of integration effort. Psychophysiology, 60(9), e14302.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., & Crocker, M. W. (2021). Retrieval (N400) and integration (P600) in expectation-based comprehension. PLOS ONE,16(9), e025743.

Brouwer, H., Crocker, M., Venhuizen, N., & Hoeks, J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. Cognitive Science, 41(S6),1318-1352.

Brouwer, H., Delogu, F., Venhuizen, N., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. Frontiers in Psychology, 12, 615538.

Delogu, F., Brouwer, H., & Crocker, M. W. (2019). Eventrelated potentials index lexical retrieval (N400) and integration (P600) during language comprehension. Brain and Cognition, 135, 103569.

De Varda, A. G., Marelli, M., & Amenta, S. (2023). Cloze probability, predictability ratings, and computational estimates for 205 English sentences, aligned with existing EEG and reading time data. Behavior Research Methods.

Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. Cognitive Psychology, 111, 15–52.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., . . . Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. Nature Neuroscience, 25(3), 369–380.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annual review of psychology, 62, 621-47.

Li, J., & Ettinger, A. (2023). Heuristic interpretation as rational inference: A computational model of the N400 and P600 in language processing. Cognition, 233, 105359.

Li, J., & Futrell, R. (2023). A decomposition of surprisal tracks the N400 and P600 brain potentials. In Proceedings of the 45th Annual Meeting of the Cognitive Science Society (p. 587-594).

Michaelov, J., Bardolph, M., Van Petten, C., Bergen, B., & Coulson, S. (2024). Strong prediction: Language model surprisal explains multiple N400 effects. Neurobiology of Language, 1-29.

Michaelov, J., & Bergen, B. (2022). Collateral facilitation in humans and language models. In Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL) (pp. 13–26).

Nieuwland, M. S., & Van Berkum, J. J. (2005). Testing the limits of the semantic illusion phenomenon: ERPs reveal temporary semantic change deafness in discourse comprehension. Cognitive Brain Research, 24(3), 691-701.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., . . . Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proceedings of the National Academy of Sciences, 118(45), e2105646118.

Smith, N., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. Psychophysiology, 52(2), 157-168.

Figures

Condition	Assoc.	Plaus.	Cloze
A Baseline	6.32	6.28	0.38
B Event-related	6.32	2.42	0.13
C Event-unrelated	1.56	1.93	0.008

Mean association ratings between prime noun and target, mean plausibility ratings for stimuli up until including target, cloze probabilities for target.

Example Item
John entered the restaurant. Before long, he opened the menu...
John left the restaurant. Before long, he opened the menu...
John entered the apartment Before long, he opened the menu...

English translation.

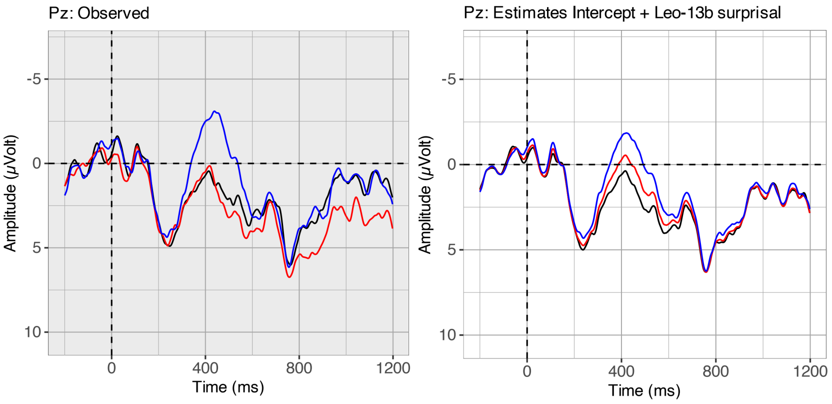

Figure 1: Left: experimental conditions from Delogu et al. (2019), middle: observed ERP profile, right: rERP forward estimates with LeoLM surprisal