On the limits of LLM Surprisal as a functional explanation of ERPs

Benedict Krieger, Harm Brouwer, Christoph Aurnhammer, Matthew W. Crocker

The impressive comprehension-like behavior of attention-based neural language models (LLMs) – trained on the task of next-word prediction – has led researchers to explore the extent to which LLMs are an accurate model of (aspects of) human comprehension mechanisms [1,2]. In particular, the amplitude of the N400, an event-related potential (ERP) component sensitive to the expectancy of a word in context, has been statistically approximated by LLM-derived operationalizations of information-theoretic surprisal, i.e., the negative log-probability of a word given its preceding context (see [3,4] for two recent examples).

Importantly however, the N400 component is sensitive not only to expectancy but also to semantic association, defined as the extent to which the meaning of a word is primed by its prior context (see [5]). While LLMs have been also been shown to be sensitive to association [6], 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 elicit increases in N400 amplitude (e.g., [7-9], see Fig. 1 as an example). While these words were clearly surprising to humans, as reflected in increased P600 amplitude, it is unclear how LLMs perform in these cases. Indeed, the P600 has in fact been found to be graded for plausibility while remaining insensitive to association [9,10], thereby supporting its role as a potential index of a *comprehension-centric* notion of surprisal [11]. However, as of yet the P600 has received little attention in studies investigating the relation between LLM-derived surprisal and human language comprehension.

The present work examines the ability of LLM surprisal to model four German ERP-studies that specifically sought to disentangle the influence of expectancy, plausibility, and association on both the N400 and the P600 [8-10,12]. Using two German-trained unidirectional transformer models of different size - one smaller model (GPT-2) and one larger state-of-the-art model (LeoLM) - we replicated previous findings demonstrating the sensitivity of LLMs to both expectancy and association. However, results from an rERP analysis [13] investigating the ability of LLM-derived surprisal to re-estimate ERPs led to mixed results for both ERP components: The estimated N400 effect patterns did not match the observed data qualitatively, as surprisal predicted an N400 effect that was not observed (in [8], see Fig. 2), or was observed with reversed directionality between two of the conditions (in [12]). Furthermore, the magnitude of N400 effects was often 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 [9], as well as P600s elicited by more subtle script knowledge violations (Fig. 2).

These findings suggest that LLM surprisal may not offer an accurate characterisation of the underlying functional generators of either the N400 or P600, when evaluated against studies that isolated the differential effects of association, expectancy and plausibility. While future LLMs may offer a better account of either the N400 or the P600, the extent to which LLMs approximate the mechanisms of human comprehension depends on their ability to account for both components. Hence, we argue such data points are crucial going forward, and motivate exploring alternative LLM-derived linking hypotheses to the N400 and P600 informed by mechanistic accounts of the processes associated with these components [14-17].

References

[1] Schrimpf et al., (2021), PNAS; **[2]** Goldstein et al., (2022), Nat. Neurosci.; **[3]** De Varda et al., (2023), Behav. Res. Methods; **[4]** Michaelov et al., (2023), Neurobiol. Lang.; **[5]** Kutas & Federmeier (2011), Annu. Rev. Psychol.; **[6]** Michaelov & Bergen (2022), CoNLL.; **[7]** Nieuwland & Van Berkum, (2005), Cogn. Brain Res.; **[8]** Delogu et al., (2019), Brain Cogn.; **[9]** Aurnhammer et al., (2023), Psychophysiology; **[10]** Aurnhammer et al., (2021), PLOS ONE; **[11]** Brouwer et al., (2021), Front. Psychol.; **[12]** Delogu et al., (2021), Brain Res.; **[13]** Smith & Kutas, (2015), Psychophysiol.; **[14]** Brouwer et al., (2017), Cogn. Sci.; **[15]** Fitz & Chang, (2019), Cogn. Psychol.; **[16]** Li & Futrell, (2023), Proc. Annu. Meet. Cogn. Sci. Soc.; **[17]** Li & Ettinger, (2023), Cognition

Figures

Cond.	Assoc.	Plaus.	Cloze	Stimulus
A	6.32	6.28	0.38	John entered the restaurant. Before long, he opened the *menu*...
B	6.32	2.42	0.13	John left the restaurant. Before long, he opened the *menu*...
C	1.56	1.93	0.008	John entered the apartment. Before long, he opened the *menu*...

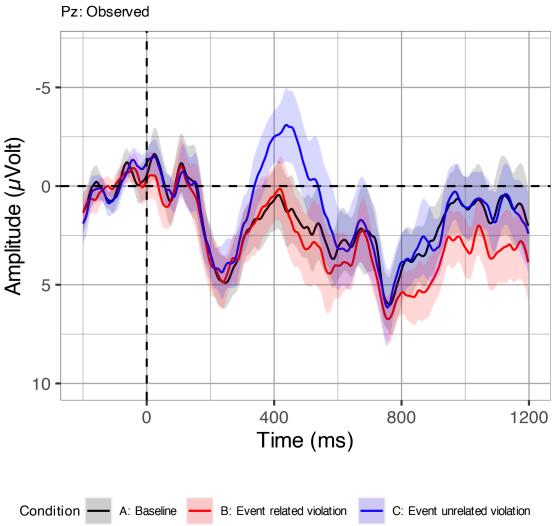

Figure 1: Experimental conditions from [8], including mean association, plausibility and cloze judgements alongside an example item (left), as well as the observed ERPs (right).

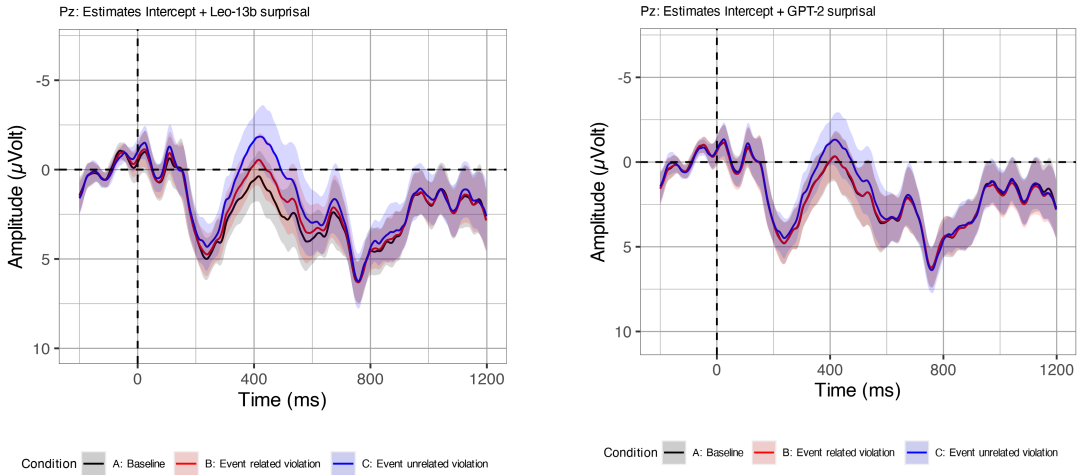

Figure 2: rERP forward estimates using LeoLM (left) and GPT-2 (right) surprisal as predictor.