On the limits of LLM Surprisal as functional explanation of ERPs

Benedict Krieger,[1] **Harm Brouwer,**[2] **Christoph Aurnhammer,**[1] **Matthew W. Crocker**[1]
bkrieger@lst.uni-saarland.de, h.brouwer@tilburguniversity.edu,
aurnhammer@coli.uni-saarland.de, crocker@coli.uni-sb.de
[1]Department of Language Science and Technology, Saarland University
[2]Department of Cognitive Science and Artificial Intelligence, Tilburg Univeristy

Abstract

Surprisal values from large language models (LLMs) have been used to model the amplitude of the N400. This ERP component is sensitive not only to contextual word expectancy but also to semantic association, such that unexpected but associated words do not always induce an N400 increase. While LLMs are also sensitive to association, it remains unclear how they behave in these cases. Moreover, another ERP component, the P600, has shown graded sensitivity to plausibility-driven expectancy, while remaining insensitive to association; however, its relationship to LLM surprisal is not well researched yet. In an rERP analysis, we evaluate surprisal values of two unidirectional transformers on their ability to model N400 and P600 effects observed in three German ERP studies isolating the effects of association, plausibility, and expectancy. We find that surprisal predicts an N400 increase for associated but implausible words, even when no such increase was observed in humans. Furthermore, LLM surprisal accounts for P600 effects elicited by violations of selectional restrictions, but captures neither P600 effects from more subtle script knowledge violations nor graded P600 modulations. The results of our investigation call into question the extent to which LLM surprisal offers an accurate characterisation of the functional generators of either the N400 or P600.

Keywords: large language models; N400; P600; event-related potentials; human language comprehension; psycholinguistics

Introduction

Human utterance comprehension is driven by incremental expectations about upcoming words. In order to formalize this notion, the concept of *surprisal* (1), originating from information theory (Shannon, 1948), has been introduced. Surprisal theory (Hale, 2001; Levy, 2008) posits that the cognitive effort required to process a word in an utterance is proportional to its surprisal, defined as the negative log-probability of this word, given the context:

$$Surprisal(w_{t+1}) = -log_2 P(w_{t+1}|w_{1...t}) \quad (1)$$

Indeed, language models, trained on the task of next-word prediction, generate probability estimates for words in context. In recent years, transformer-based models (Vaswani et al., 2017) have become prevalent and their increased scale in terms of model complexity and training data size has led to the term large language models (LLMs). Although not designed for this purpose, surprisal values computed from LLMs have been found to be predictive of not only behavioral indices of human language processing effort, such as reading times and eye movements, but also neural indices, such as event-related potentials, which offer a direct, multidimensional window into language comprehension in the brain (Frank, Otten, Galli, & Vigliocco, 2015; Michaelov, Bardolph, Van Petten, Bergen, & Coulson, 2023).

In particular, previous research has established a strong link between LLM derived estimations of expectancy and the N400, a negative going ERP component peaking 400 ms post-stimulus onset. 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 Kutas & Federmeier, 2011). While LLMs have been also been shown to be 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 elicit increases in N400 amplitude (e.g., Aurnhammer, Delogu, Brouwer, & Crocker, 2023; Nieuwland & Van Berkum, 2005; Delogu, Brouwer, & Crocker, 2019). While these words were clearly surprising to humans, as reflected in increased amplitude of the P600 – a later, positive going ERP component – it is unclear how LLMs perform in these cases. Indeed, the P600 has recently been found to be graded for plausibility while remaining insensitive to association (Aurnhammer, Delogu, Schulz, Brouwer, & Crocker, 2021; Aurnhammer et al., 2023), thereby supporting its role as a potential index of a *comprehension-centric* notion of surprisal (Brouwer, Delogu, Venhuizen, & Crocker, 2021), which is sensitive to both the unfolding utterance meaning and our knowledge about the world. However, as of yet, the P600 has received little attention in studies investigating the relation between LLM-derived surprisal and human language comprehension.

ERPs elicited by surprisal To elucidate the interplay of expectancy and association in surprisal values computed from LLMs, we examine three German ERP studies. Importantly, while experiments often operationalize expectancy as cloze probability, this measure is poor at distinguishing possible vs. implausible words, which may both have zero cloze but differing surprisal. Therefore, we focus on studies that specifically sought to disentangle the influence of association, expectancy, and plausibility on both the N400 and the P600. These three ERP studies revealed: (1) additive influences of

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

association and expectancy on the N400 (Aurnhammer et al., 2021); (2) that the influence of expectancy on the N400 can be overridden by strong semantic association (Aurnhammer et al., 2021; Delogu et al., 2019); (3) that P600 increases were elicited not only by strong violations of selectional restriction (Aurnhammer et al., 2021), but also by violations of script knowledge (Delogu et al., 2019), and (4) a graded sensitivity of the P600 to plausibility (Aurnhammer et al., 2023).

The predictive power of LLMs for ERPs Although LLM surprisal values are generally interpreted as reflecting expectancy, it has been demonstrated that these surprisal values are also sensitive to the semantic association of an implausible word with the context (Michaelov & Bergen, 2022 in the stimuli of Metusalem et al., 2012; see also Michaelov et al., 2023). Importantly, it remains unclear how well LLM surprisal can explain cases where association to the context entirely overrides the influence of expectancy on the N400 (e.g. Aurnhammer et al., 2023; Delogu et al., 2019). Furthermore, although the P600 was shown to reflect expectancy and to index plausibility in a graded manner while remaining insensitive to association, to our knowledge only one study has explored the link between the P600 and LLM surprisal values (De Varda, Marelli, & Amenta, 2023), which found that only surprisal from larger models may capture P600 effects.[1]

The aforementioned ERP studies thus provide an ideal test to examine the extent to which LLM-derived surprisal values account for the differential sensitivities of the N400 and the P600, thereby evaluating whether LLMs accurately reflect comprehension-centric surprisal (Brouwer, Delogu, Venhuizen, & Crocker, 2021).

Method

LLM surprisal

Transformer-based models with more parameters provided a better model-fit to both N400 and P600 amplitude in De Varda et al. (2023). However, when evaluated on reading time measures, models that were architecturally more complex or trained on more data were shown to provide a poorer fit (Oh & Schuler, 2022). Since a clear relationship between model complexity, traning data size and ERP components is not established yet, we therefore focused on the evaluation of two pre-trained transformer models that vary in terms of model complexity and number of parameters: a smaller GPT-2 model[2] with 124 million parameters and a larger model with 13 billion parameters based on the Llama-2 architecture, LeoLM.[3] Both models implement a decoder-only transformer architecture, featuring a masked self-attention mechanism that allows them to selectively weigh the influence of all preceding tokens in generating a probability distribution over the vocabulary to predict the next token. Importantly, the models differ also with respect to their German training data: The GPT-2 model was trained on a 16 GB corpus, comprising several smaller sub-corpora, including texts from Wikipedia, NewsCrawl, ParaCrawl, EU bookshop corpus, and Open Subtitles. In contrast, LeoLM was trained on the 595 GB OSCAR 23.01 corpus, as well as on approximately 10 GB of Wikipedia texts and a small 65 MB set of German news articles. Lastly, the two models used different tokenization techniques: While the GPT-2 model uses Byte-Pair-Encoding (BPE; Sennrich, Haddow, & Birch, 2016), LeoLM uses a Llama-tokenizer that is based on SentencePiece (Kudo & Richardson, 2018; see Nair & Resnik, 2023, for a discussion of potential problems of these methods for psycholinguistic research).

To obtain surprisal values we presented the LLMs with the items of the evaluated studies up to, but not including the target word. Probability estimates for the target words were then converted to surprisal by computing their negative logarithm (see Equation 1). When target words were tokenized into sub-words, surprisal values of their sub-words were added (see also Oh & Schuler, 2022; De Varda et al., 2023).

ERP Analysis

To assess the capabilities of surprisal values to account for the ERPs, we apply the rERP method (Smith & Kutas, 2015), a technique in which a separate linear model is fitted for each subject, electrode, and time sample. The individual linear regressions optimally combine the specified predictors to explain the variability in the observed signal and in sum allow us to analyse ERPs at full temporal and spatial resolution. In our approach, we aim to re-estimate the N400 and P600 effects observed between conditions in the aforementioned studies by using the surprisal values from the two language models as a single predictor (apart from the intercept), thus leading to the following regression model specification.

$$Y = \beta_0 + \beta_1 \text{surprisal} + \varepsilon \quad (2)$$

Using the fitted regression models, forward estimates for the entire ERP datasets were computed and averaged across subjects and conditions, analogous to the traditional ERP-averaging procedure. Importantly, the estimates of the models were only informed by the surprisal values for the target words. That is, the linear models did not have access to condition-coded predictors and the estimates were only averaged per condition retrospectively (see Aurnhammer et al., 2021, 2023; Brouwer, Delogu, & Crocker, 2021 for the same approach applied to the three studies at hand).

Due to space limitations, we refrain from reporting the full results of the rERP analysis, consisting of coefficients, residuals, as well as t- and p-values and instead focus on the regression estimates only. To evaluate whether the re-estimated ERPs match the observed ERPs, we thus inspect the estimated ERPs in the N400 and the P600 time window. Crucially, this approach goes beyond assessing the significance of the surprisal predictor and examines whether surprisal can

[1] See also (Frank et al., 2015; Michaelov & Bergen, 2020) for RNN-based investigations.

[2] https://huggingface.co/stefan-it/secret-gpt2/tree/main

[3] https://huggingface.co/LeoLM/leo-hessianai-13b

adequately re-estimate the observed N400 and P600 effects.

Experiments & Results

We present our results per study, respectively introducing the original design and findings first. Example items, alongside cloze probabilities as well as association and plausibility judgements are shown in Table 1. The experimentally elicited ERP profile is presented in the top row of Figure 2.

In order to assess the extent to which the surprisal values pattern with association and plausibility ratings as well as cloze probabilities – and therefore which effect patterns they may be able to capture – we present the distribution of the surprisal values obtained from the two LLMs for the three ERP studies, grouped by condition (Figure 1; note differences in axis scales). While we report on the statistical significance of surprisal (see Aurnhammer et al., 2023, for inferential statistics from rERPs), we focus on the forward estimates generated by the linear models using surprisal from the two LLMs. We restrict our report to electrode Pz where the N400 and P600 effects were peaking in the studies examined here.

Aurnhammer et al. (2021)

The study conducted by Aurnhammer et al. (2021) differentially manipulated target word expectancy through the selectional restrictions of the main verb ("sharpened ... the *axe*" vs. "ate ... the *axe*") and lexical association to an intervening adverbial clause ("before he the wood stacked, the *axe*" vs. "before he the movie watched, the *axe*"; cf. Table 1, for transliterations of the stimuli, mean noun-target association ratings, and cloze probabilities). The stimuli elicited additive N400 modulations (300-500 ms) from both expectancy and association, whereas the P600 (600-1000 ms) was increased only for unexpected relative to expected targets, while remaining insensitive to association (Figure 2.1).

Both LLMs produce the lowest surprisal values for condition A (strong association and high cloze) and the highest surprisal for condition D (weak association and low cloze). Moreover, less expected and less associated conditions result in higher mean surprisal, when keeping the other factor constant, respectively (Figure 1, first column). In the rERP analysis, LLM surprisal values allow us to approximate the influence of both expectancy and association on the N400 component (Figure 2.2 & 2.3, 300-500 ms): Both LLMs predict an N400 difference in the unexpected relative to the expected conditions (Conditions C/D vs. A/B). While GPT-2 surprisal values estimate an N400 difference of both associated conditions relative to the unassociated conditions (B/D vs. A/C), LeoLM surprisal values appear to predict an N400 difference from association only between the unexpected conditions. Turning to the P600, we find that the LeoLM regression estimates exhibit a clear P600 increase for selectional restriction violations (Conditions C/D vs. A/B; Figure 2.2, 600-1000ms). The forward estimates computed from GPT-2 surprisal appear to exhibit the same trend, albeit with a smaller magnitude (Figure 2.3). For both LLMs, the surprisal values that are sensitive to association also lead to an erroneous prediction of small association effects in the P600. LLM surprisal was significant in both time windows for both LLMs.

Delogu et al. (2019)

The influence of association was also examined by Delogu et al. (2019) who observed that for two conditions in which semantic association was equally strong (between "restaurant" and "menu"), a difference in plausibility – induced by a violation of script knowledge (opening the menu after entering/leaving the restaurant) – did not lead to an N400 effect (see the *event-related violation*, condition B, compared to the baseline, A, in Figure 2.4). Rather, the difference in plausibility led to a P600 effect (800 - 1000 ms). While no P600 effect was observed between the *event-unrelated violation* (C) condition relative to baseline, subsequent studies revealed that a large N400 increase elicited by the target words – which were not only implausible but also unassociated to the context – obscured the spatio-temporally overlapping P600 (Brouwer, Delogu, & Crocker, 2021; Delogu, Brouwer, & Crocker, 2021).

Here, where implausible words did not lead to an increase in N400 amplitude due to strong association, we observe a different pattern of surprisal values from the two LLMs (Figure 1). While LeoLM yields noticeably higher mean surprisal values for condition B relative to A, in line with mean plausibility and cloze, GPT-2 appears to produce very similar surprisal values for both conditions, in line with mean association. Hence, although no N400 effect was observed between condition B relative to baseline in the original study, LeoLM surprisal predicts an N400 difference between these conditions in the rERP analysis (Figure 2.5). In contrast, entering GPT-2 surprisal into the rERP analysis does not predict any difference between the conditions – in line with the observed ERPs (Figure 2.6). In the rERP analysis of the P600 window, neither LLM captures the originally observed effect of condition B relative to A (800-1000 ms). Note that even though LeoLM is sensitive to the script-violation, its sensitivity to association would prevent it to explain the data even when taking spatiotemporal component overlap into account; see Brouwer, Delogu, & Crocker, 2021).[4] LLM surprisal was significant only in the N400 time window for both LLMs.

Aurnhammer et al. (2023)

Aurnhammer et al. (2023) demonstrated a graded link of plausibility to the P600 across a plausible, less plausible and implausible condition – a relation that was modelled statistically by continuous plausibility ratings (cf. Table 1c, Figure 2.7). Further, repetition priming of the target word in a preceding context paragraph led to the absence of any N400 effects between conditions.[5]

The surprisal values generated by both LLMs are higher

[4] Results qualitatively similar to those for Delogu et al. (2019) were also obtained for the data by Delogu et al. (2021) and are omitted here.

[5] See Aurnhammer et al. (2023), for discussion of a mismatch negativity elicited by the less plausible condition, which additionally created semantic attraction towards a distractor word (250-400 ms).

Table 1: Example items from the evaluated studies.

Condition	Assoc.	Plaus.	Cloze	Stimulus
A Assoc+Exp+	6.29	-	0.67	Yesterday sharpened the lumberjack, before he the wood stacked, the axe ...
B Assoc-Exp+	2.09	-	0.64	Yesterday sharpened the lumberjack, before he the movie watched, the axe ...
C Assoc+Exp-	6.29	-	0.008	Yesterday ate the lumberjack, before he the wood stacked, the axe ...
D Assoc-Exp-	2.09	-	0.008	Yesterday ate the lumberjack, before he the movie watched, the axe ...
				Aurnhammer et al. (2021)
A Baseline	6.32	6.28	0.38	John entered the restaurant. Before long, he opened the menu ...
B Event-related	6.32	2.42	0.13	John left the restaurant. Before long, he opened the menu ...
C Event-unrelated	1.56	1.93	0.008	John entered the apartment. Before long, he opened the menu ...
				Delogu et al. (2019)
[Context:] *A tourist wanted to take his huge suitcase onto the airplane. The suitcase was however so heavy that the woman at the check-in decided to charge the tourist an extra fee. After that, the tourist opened his suitcase and threw several things out. Now, the suitcase of the ingenious tourist weighed less than the maximum of 30 kilograms.*				
A Baseline	-	5.84	0.8	Then dismissed the lady the tourist ...
B Less plausible	-	3.69	0.09	Then weighed the lady the tourist ...
C Implausible	-	1.42	0.02	Then signed the lady the tourist ...
				Aurnhammer et al. (2023)

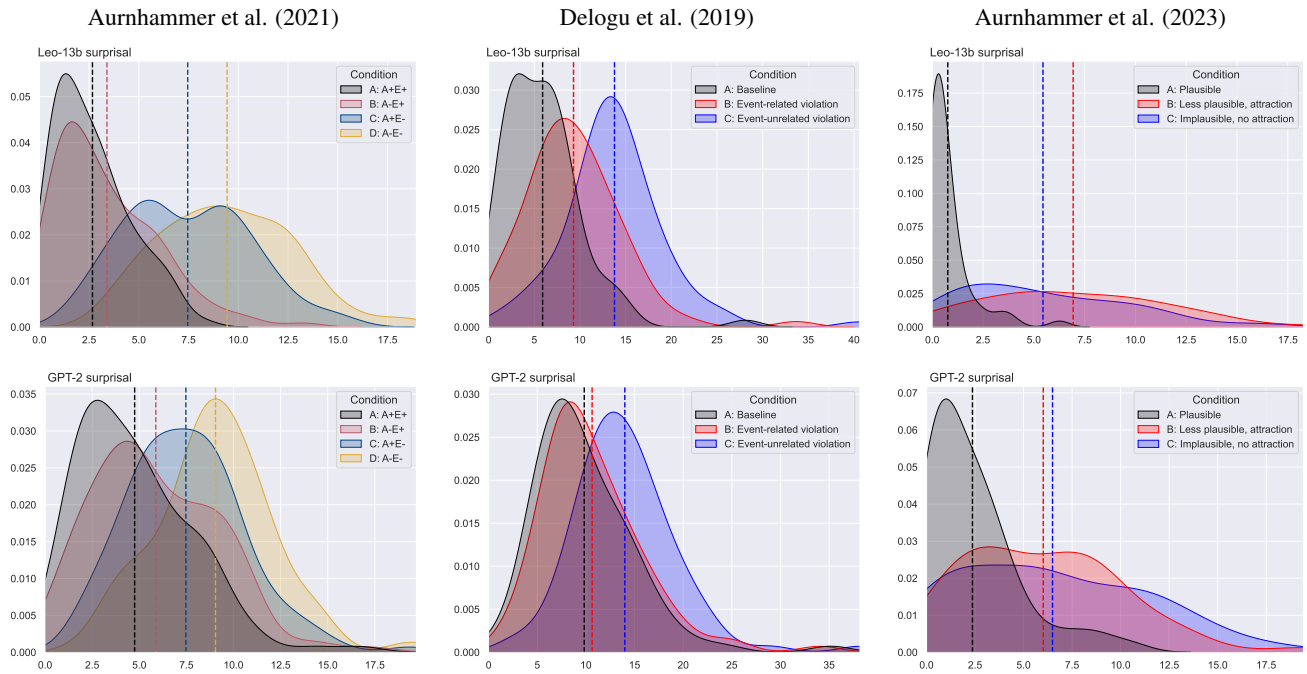

Figure 1: Densities of LLM surprisal values for target words, split by condition. Dashed lines indicate average values.

and exhibit a wider spread in the less plausible and implausible condition (B and C) compared to the plausible baseline (A), with LeoLM producing the highest surprisal for B. rERPs reveal that both sets of LLM surprisal values predict a small negativity for both the less plausible condition B and the implausible condition C relative to baseline (Figure 2.8, 2.9, 300-500 ms). In the P600 window, the linear models for both LLMs predict the less plausible condition B and the implausible condition C to be more positive than the plausible baseline. Crucially however, neither model predicts the graded P600 response to plausibility, i.e., the observed graded pattern between A, B *and* C (Figure 2.8 & 2.9, 600-1000 ms). LLM surprisal was significant in both time windows for LeoLM, but only significant in the P600 window for GPT-2.

Discussion

Large language models have demonstrated impressive performance in approximating brain activity during language comprehension, leading researchers to explore the hypothesis that both humans and LLMs might share similar processing mechanisms (Goldstein et al., 2022; Schrimpf et al., 2021). Building on recent studies showing that LLM-derived surprisal

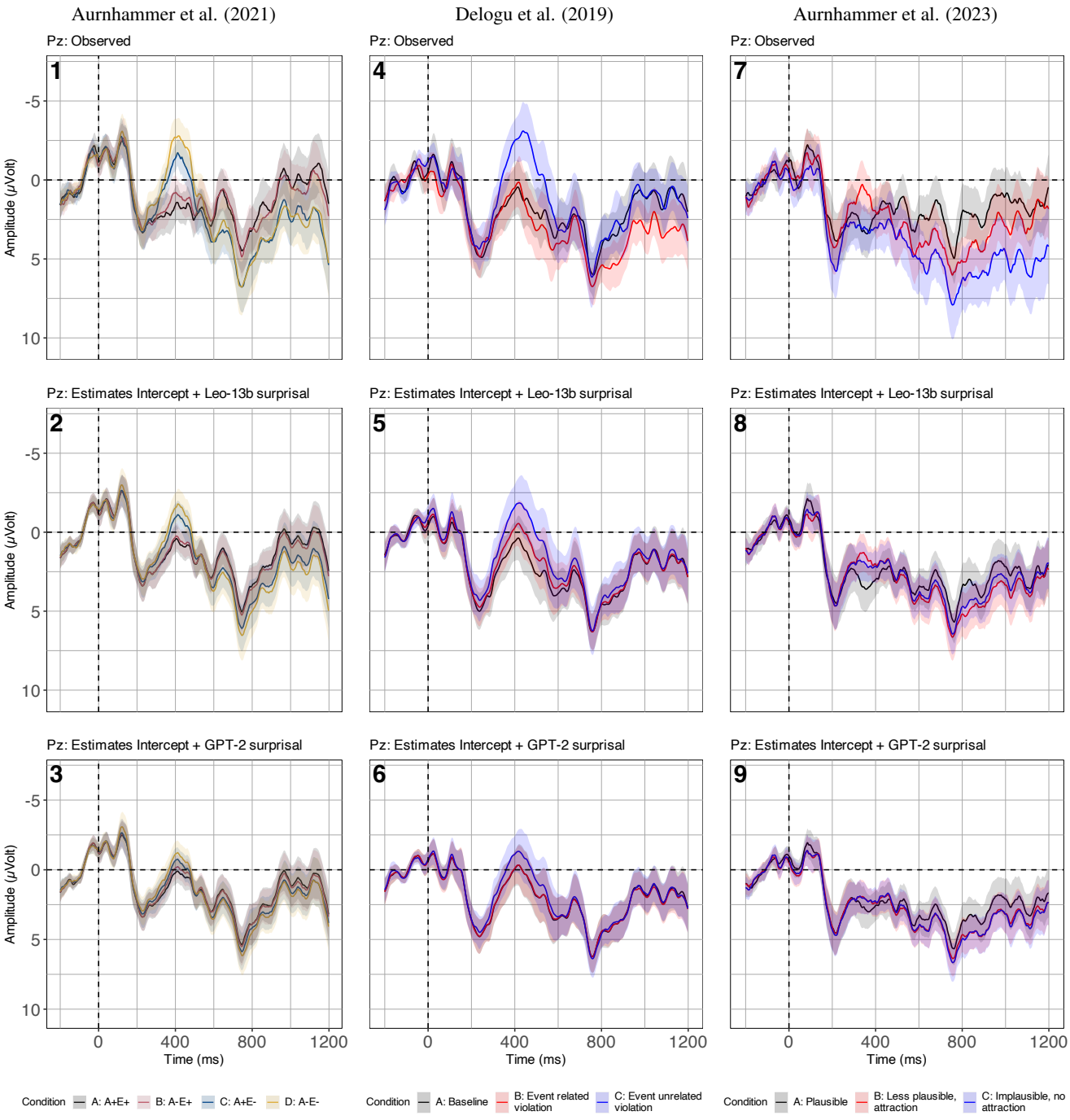

Figure 2: Observed ERPs (row 1), rERP estimates using LeoLM (row 2) and GPT-2 (row 3) surprisal values.

provides a good fit on a range of N400 findings, we evaluated their performance on a set of German studies demonstrating that the P600 may provide a more direct index of human comprehension-centric surprisal, as association can attenuate or eliminate expectation effects in the N400.

Examining Aurnhammer et al. (2021), we observe that both LLMs produce surprisal values that capture the observed N400 pattern in the rERP analysis. That is, the unexpected and/or unassociated conditions yield successively higher mean surprisal relative to the baseline condition. Therefore, both LLMs' surprisal values can successfully predict the additive effects of association and expectancy in the N400 window and the effect of expectancy in the P600 window. However, the sensitivity to association in both LLMs' surprisal values leads to an erroneous prediction of a (numerical) P600 difference induced by association in the rERPs.

For Delogu et al. (2019), the two LLMs produce different average surprisal values: LeoLM additively produces higher mean surprisal for stimuli with decreasing plausibility and association. This pattern leads to the estimation of an N400 increase in an associated but implausible condition that exhibited only an increase in P600 amplitude, but not N400 amplitude. GPT-2, in turn, produces an increase in surprisal only for the unassociated condition, whereas the surprisal of an associated but implausible condition is close to baseline. Thus, in the rERP procedure, GPT-2 correctly predicts only an N400 effect for the unassociated but implausible condition. Interestingly, it is precisely the inability of GPT-2 to account for the script-knowledge violation that allows it to correctly model the N400 modulation pattern and prohibits it from accurately modeling the P600 pattern. While LeoLM produces higher surprisal for script-knowledge violations, its additional sensitivity to association means surprisal values do not accurately capture the P600 modulations.

Finally, for Aurnhammer et al. (2023), both LLMs generate increased surprisal values for less plausible and implausible conditions; however surprisal from neither LLM accurately reflects the differences between these two conditions. Consequently, both LLMs failed to account for the gradedness of the P600 in the rERP analysis of this study.

These observations are enabled by rERP analyses that focus on examining whether surprisal values can reproduce the observed ERP effects, which requires going beyond assessing only significance and raw model fit (as measured by, e.g., AIC). When evaluated against studies that isolated the differential effects of association, expectancy and plausibility, LLM surprisal does not offer a complete characterisation of the underlying functional generators of either the N400 or P600. Although the differential N400 and P600 effect pattern observed in the evaluated studies is impossible to capture with a single predictor, a viable model of human surprisal should be expected to account for plausibility and expectancy-driven effects to a greater extent than for association-driven effects. While our results do suggest that larger language models may move to such a notion of surprisal, as LeoLM fares better in accounting for the implausible stimuli of Delogu et al. (2019), even this larger model remains sensitive to association (with highest surprisal for the unassociated *event-unrelated violation*). Thus, the question remains to what extent pure association impacts the probability distributions generated by LLMs with regard to how accurately those distributions model human expectancy.

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 (e.g., Brouwer, Crocker, Venhuizen, & Hoeks, 2017; Fitz & Chang, 2019; Li & Ettinger, 2023; Li & Futrell, 2023).

For instance, on the Retrieval-Integration account, the N400 indexes the retrieval of word meaning from long-term memory, while the P600 reflects the integration of this meaning into an unfolding utterance representation (Brouwer et al., 2017; Brouwer, Fitz, & Hoeks, 2012). In a neurocomputational instantiation of this account, processing a sentence word-by-word proceeds in repeated steps of retrieval and integration. Estimates for both N400 and P600 amplitude emerge as the degree of change in layer activation in the respective retrieval and integration modules from w_t to w_{t+1}. In contrast, LLMs remain opaque regarding the functional interpretation of their internal mechanisms. Recent research has, however, begun to elucidate these mechanisms (Geva, Caciularu, Wang, & Goldberg, 2022; Oh & Schuler, 2023), which may allow us to assess whether LLMs instantiate subprocesses functionally similar to those of retrieval and integration within human comprehenders.

Conclusion

In this work we examined to what extent LLM surprisal can model experimentally observed N400 and P600 effects, applying the rERP framework to re-estimate the ERPs elicited in three German studies that independently manipulated the influences of contextual association, plausibility, and expectancy on both ERP components.

The nature of the carefully controlled stimulus materials – as opposed to more naturalistic language – led to the observation of surprisal values that appear to reflect different sensitivities of the two LLMs towards association and plausibility, leading to inconsistent estimations of N400 and P600 differences in the rERPs. Although it is impossible for LLM surprisal as a single predictor to account for all observed ERP differences due to the orthogonality of the underlying manipulations, surprisal theory (Hale, 2001; Levy, 2008) predicts that the surprisal of a word should be unaffected by its contextual association and only be driven by its syntactic and semantic expectedness. Such a tendency could be observed more clearly in the larger model's surprisal values. Importantly however, surprisal values from both models underestimated the observed differences in all three studies and were also unable to fully capture the influence of script knowledge violations and graded plausibility on the P600.

Therefore, the question remains how strongly the probability distributions of LLMs derived from next-word prediction match the (plausibility-driven) expectations of human comprehenders and thus to what extent LLMs can be viewed as models of online human sentence comprehension. Our results motivate exploring alternative LLM-derived linking hypotheses – informed by mechanistic accounts of the underlying processes of the N400 and P600 – and we argue that until LLMs are shown to account for critical data points through such precise linking hypotheses, strong conclusions about their validity as models of the human comprehension system (e.g. Goldstein et al., 2022) are too premature.

Acknowledgments

This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 232722074—SFB 1102.

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., & Crocker, M. W. (2021). Splitting event-related potentials: Modeling latent components using regression-based waveform estimation. *European Journal of Neuroscience*, *53*, 974-995.

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.

Brouwer, H., Fitz, H., & Hoeks, J. C. J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143.

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

Delogu, F., Brouwer, H., & Crocker, M. W. (2021). When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. *Brain Research*, *1766*, 147514.

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.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1-11.

Geva, M., Caciularu, A., Wang, K. R., & Goldberg, Y. (2022). Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. doi: 10.48550/ARXIV.2203.14680

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.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)* (Vol. 2, p. 1-8).

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66–71).

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.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.

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).

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567.

Michaelov, J., Bardolph, M., Van Petten, C., Bergen, B., & Coulson, S. (2023). Strong prediction: Language model surprisal explains multiple N400 effects. *Neurobiology of Language*, 1-29.

Michaelov, J., & Bergen, B. (2020). How well does surprisal explain N400 amplitude under different experimental conditions? In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 652–663).

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).

Nair, S., & Resnik, P. (2023). *Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship?* arXiv. (arXiv:2310.17774 [cs])

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.

Oh, B.-D., & Schuler, W. (2022). *Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times?* (arXiv: 2212.12131 [cs.CL])

Oh, B.-D., & Schuler, W. (2023). Token-wise Decomposition of Autoregressive Language Model Hidden States for Analyzing Model Predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.562

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.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., … Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (p. 5998–6008).