Brain Research 1865 (2025) 149841

Contents lists available at ScienceDirect

Brain Research

journal homepage: www.elsevier.com/locate/brainres

Research paper

On the limits of LLM surprisal as a functional explanation of the N400 and P600

Benedict Krieger [a,*], Harm Brouwer [b], Christoph Aurnhammer [a], Matthew W. Crocker [a]

[a] Department of Language Science and Technology, Saarland University, Germany
[b] Department of Cognitive Science and Artificial Intelligence, Tilburg University, The Netherlands

ARTICLE INFO

Edited by Marian Berryhill

Dataset link: https://github.com/benedict-krieger/llm-surprisal-rerps

Keywords:
Large language models
N400
P600
Event-related potentials
Human language comprehension
Psycholinguistics

ABSTRACT

Expectations about upcoming words play a central role in language comprehension, with expected words being processed more easily than less expected ones. Surprisal theory formalizes this relationship by positing that cognitive effort is proportional to a word's negative log-probability in context, as determined by distributional, linguistic, and world knowledge constraints. The emergence of large language models (LLMs) demonstrating the capacity to compute richly contextualized surprisal estimates, has motivated their consideration as models of comprehension. We assess here the relationship of LLM surprisal with two key neural correlates of comprehension – the N400 and the P600 – which differ in sensitivity to semantic association and contextual expectancy. While prior work has focused on the N400, we propose that the P600 may offer a better index of surprisal, as it is unaffected by association while still patterning continuously with expectancy. Using regression-based ERPs (rERPs), we examine data from three German factorial studies to evaluate the extent to which LLM surprisal can account for ERP differences. Our results show that LLM surprisal captures neither component consistently. We find that it is contaminated by simple association, particularly in smaller LLMs. As a result, LLM surprisal can partially account for association-driven N400 effects, but not for the full attenuation of N400 effects. Correspondingly, this property of LLMs compromises their ability to model the P600, which is sensitive to expectancy but not to association.

1. Introduction

Expectations regarding the next word play a central role in language comprehension, as they reflect how linguistic and world knowledge interact with context to constrain how the linguistic signal is likely to unfold. As a consequence, listeners process expected words with greater ease than less expected ones. Empirical evidence for expectation-based processing dates back several decades. For instance, expected words were found to be read more quickly (Ehrlich and Rayner, 1981) or to elicit an attenuated N400 amplitude (Kutas and Hillyard, 1984) during reading. A general formalization of this relationship between expectedness and processing effort was introduced with surprisal theory (Hale, 2001; Levy, 2008), which posits that the cognitive effort required to process a word is proportional to its negative log-probability in context:

$$\text{difficulty} \propto \text{surprisal}(w_{t+1}) = -log_2 P(w_{t+1}|w_{1...t}) \tag{1}$$

The *true* expectancy of a word should in principle reflect all relevant determinants of what word can appear next – including distributional, linguistic, and world knowledge-based plausibility constraints – while negative log expectancy (true surprisal) should be proportional to cognitive effort (Levy, 2008; Venhuizen et al., 2019). It follows from this formalization that words that are less expected will result in higher surprisal and will be more difficult to integrate into the mental representation of the utterance, while expected words will require less effort. Importantly, the link between expectancy and cognitive effort in Eq. (1) can inform our understanding about (a) which empirical measures best index true surprisal, and (b) which models best approximate both true surprisal and – if divergent – cognitive indices of surprisal. The latter may be particularly relevant in determining the extent to which models use mechanisms and representations similar to those underlying human comprehension. The present study examines critical evidence from the two most salient neural correlates of comprehension – the N400 and the P600 components of the EEG signal – to assess how well they index surprisal as operationalized by current large language models (LLMs).

The empirical support for surprisal theory is considerable. Since its introduction, numerous studies have found word predictability to be correlated with various indices of cognitive processing effort. This includes not only evidence from behavioral metrics such as self-paced reading and eye-tracking data (e.g., Brouwer et al., 2010; Demberg and

[*] Corresponding author.
E-mail address: bkrieger@lst.uni-saarland.de (B. Krieger).

https://doi.org/10.1016/j.brainres.2025.149841
Received 30 January 2025; Received in revised form 26 June 2025; Accepted 15 July 2025
Available online 7 August 2025

Keller, 2008; Mitchell et al., 2010, Fernandez Monsalve et al., 2012; Oh and Schuler, 2023a; Smith and Levy, 2008; Wilcox et al., 2020), but also measures from brain activity, such as EEG and fMRI (e.g., Frank et al., 2015; Frank and Willems, 2017; Michaelov et al., 2024; Shain et al., 2024). Moreover, the predictions of surprisal theory have been shown to robustly hold across multiple languages (Wilcox et al., 2023b) and to transfer to multiple linguistic levels (Ettinger et al., 2014; Hu et al., 2023; Malisz et al., 2018). In sum, there exists broad evidence in support of surprisal and consequently of the notion that language processing in the human brain is guided by probabilistic expectations.

Importantly, the link between various neurobehavioral indices of processing effort and word predictability as posited by surprisal theory is formulated at the computational level in Eq. (1), which Marr (1982) specifies as *what* problem the system seeks to solve. This leaves open the algorithmic level, which Marr defines as *how* the computational problem is solved. Critically, any generative stochastic process that is able to estimate contextual word probabilities – also known as a *language model* – can be considered an algorithmic level implementation of the computational theory that is able to estimate surprisal. Hence, surprisal acts as a “causal bottleneck” between different algorithmic language model implementations and observable processing phenomena (Levy, 2008). When introduced by Hale (2001), surprisal was computed with an Earley parser (Earley, 1970; Stolcke, 1995) on a probabilistic context-free phrase-structure grammar (PCFG). Since then, numerous studies have operationalized surprisal using a variety of computational models, including PCFGs (e.g., Demberg and Keller, 2008), n-grams (e.g., Smith and Levy, 2008), recurrent neural networks (RNNs; e.g., Aurnhammer and Frank, 2019) and, more recently, LLMs (e.g., Oh and Schuler, 2023a).

1.1. Large language models as models of human comprehension

Large language models (LLMs) are deep neural network models that predict the next word in an input sequence by generating a probability distribution over all possible candidate tokens in their vocabulary. Throughout training, their parameters are adjusted in order to minimize prediction error, which is often evaluated by computing perplexity, the exponentiated average negative log-likelihood per token (Meister and Cotterell, 2021). Thus, LLMs directly compute richly contextualized surprisal estimates. Their grounding in predictive processing, as well as their ability to generate coherent and deceptively human-like text, has led to considerable interest in exploring the status of LLMs as cognitive models at the computational level (Contreras Kallens et al., 2023). Piantadosi (2023), for example, views the finding that they to some extent encode semantic and syntactic representations (e.g., Manning et al., 2020) as strong counter-evidence to traditional generative linguistic approaches (Chomsky, 1965), and proposes to treat LLMs as serious models of human cognition which allow to “develop compelling theories of the interplay of structure and statistics” (p. 383).

Moreover, the internal representations of LLMs were successfully mapped to brain responses during natural language comprehension in a number of studies (e.g., Caucheteux and King, 2022; Caucheteux et al., 2023; Goldstein et al., 2022; Schrimpf et al., 2021) across several neuroimaging response measures, such as electrocorticography (ECoG), functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG). The overall strong correlation between model representations and brain responses was interpreted as evidence that human language processing is based on predictive coding, to the extent that “predictive processing fundamentally shapes the language comprehension mechanisms in the human brain” (Schrimpf et al., 2021). Going even further, Goldstein et al. (2022) argue that both the brain and autoregressive transformer LLMs share certain mechanisms, specifically pre-onset word prediction, post-onset tracing of prediction error, and reliance on contextual embeddings. From a broader methodological perspective, Caucheteux et al. (2023) view their results as illustrating “how the synergy between neuroscience and artificial intelligence can unravel the computational bases of human cognition”.

Together, these studies highlight the recent appeal of LLMs as models of human comprehension due to their impressive performance and correlation to human brain responses during language comprehension. Even though such strong claims are debated (see for example Katzir, 2023, for a reply to Piantadosi, 2023), what is remarkable is that the performance of LLMs is solely grounded in the task of next-word prediction.[1] Indeed, not only do LLMs – compared to previous computational models – generate surprisal estimates which provide a closer fit to neurobehavioral indices of expectancy-related processing effort (Merkx and Frank, 2021; Michaelov et al., 2021), they also yield hidden states that correlate well with brain activity during language comprehension (Schrimpf et al., 2021). These observations have motivated the investigation of whether similarities between LLMs and humans are not limited to the computational, but extend to the algorithmic level, for instance by instantiating predictive coding, in which an error signal arising from pre-onset predictions is traced back in order to update an internal probability distribution (Goldstein et al., 2022; Michaelov et al., 2024). In other words, LLMs and humans may not only be similar in *what* they predict, but also in *how* they predict. That is, if the representations and mechanisms underlying the prediction process in LLMs are sufficiently similar to those involved in prediction in the brain during language comprehension, then reducing perplexity on an appropriate training corpus may lead to surprisal estimates that better approximate predictive behaviors in humans.

Identifying the degree to which LLM surprisal approximates human responses and/or where it diverges qualitatively and quantitatively can offer insights about the extent to which they may implement a function – such as next word prediction – in an algorithmically similar manner. It is important to note, however, that many of the above studies are based on evidence from naturalistic language such as podcasts, novels or newspapers, i.e., language that has not been modified with any particular hypothesis in mind. While naturalistic data offers the advantage of observing responses to language in a natural setting, and potentially increases the generalizability of results (Frank et al., 2015), such datasets may not reveal how distinct properties of language affect processing differentially. Indeed, this observation underlies the rich history of controlled factorial experiments, in which participants are exposed to items in different conditions. As these conditions only vary with respect to specific factor levels, systematic differences in responses across participants and items can be attributed to the experimental manipulations. Such studies have been crucial to identifying how distinct neural correlates of processing effort, as measured using event-related potentials, are differentially sensitive to properties of a word other than just its overall contextual expectancy.

1.2. Neural correlates of surprisal

Event-related potentials (ERPs) offer a multidimensional window into language comprehension at a high temporal resolution, allowing for the investigation of the time course of its unfolding sub-processes. Since its discovery by Kutas and Hillyard (1980), the N400 has been shown to be one of the most robust neural markers indicating processing effort related to how predictable a word is in a given context – the less predictable the word, the more negative the N400 response. This relationship was identified long before the introduction of surprisal theory and, in the context of ERP studies, the predictability of a word has often been operationalized as cloze probability – the proportion of participants who offered this word as a completion in a separate norming study (Taylor, 1953; see Kutas and Hillyard, 1984 for its first application in ERP research). While cloze probability offers a good

[1] We are excluding models that incorporate reinforcement learning by human feedback here.

estimate of predictable words, it poorly captures the lower end of the probability spectrum, such that both implausible and plausible but relatively rare words typically yield a cloze probability of zero. This constitutes one reason why language model surprisal – as another operationalization of expectancy that may cover the full probability distribution more adequately – has become popular in recent years (for more operationalizations of expectancy and how well they fit to neurobehavioral data, see de Varda et al., 2023).

Frank et al. (2015) were among the first to use surprisal values from three different language models (PCFG, n-gram, RNN) to predict the amplitude of six different ERP components, including the N400. Naturalistic sentences of written English from the UCL corpus (Frank et al., 2013) were used as materials, and the authors found a strong relationship between surprisal and the N400. The same dataset was used in a number of further studies to test the fit of surprisal values collected from different types of language model architectures, furthermore focusing predominantly on the amplitude of the N400 (e.g., Aurnhammer and Frank, 2019; de Varda et al., 2023; Merkx and Frank, 2021; Frank and Willems, 2017). More recently, N400 responses from controlled factorial experimental studies have also been modeled with LLM surprisal (Michaelov and Bergen, 2020; Michaelov et al., 2021; Michaelov et al., 2023; Michaelov et al., 2024).

However, the N400 is also sensitive to a number of other linguistic and non-linguistic stimulus properties beyond the contextually-determined expectancy of a word (Kutas and Federmeier, 2011). One such property is the semantic association of a word to the preceding context, i.e., the degree of its semantic relatedness. In naturalistic stimuli, association and expectancy are often confounded in that the words that are likely to come next, will often also be associated with the context. Critically, however, association and expectancy are distinct; that is, a word that is highly unexpected to immediately follow, may nonetheless be strongly associated to the context. This contrast was directly investigated by Delogu et al. (2019), in an ERP study containing the following experimental conditions:

(Ex. 1) **Assoc+Exp+** John entered the restaurant. Before long, he opened the menu...
Assoc+Exp– John left the restaurant. Before long, he opened the menu...
Assoc–Exp– John entered the apartment. Before long, he opened the menu...

The target word *menu* is expected in condition Assoc+Exp+, but unexpected in condition Assoc+Exp–, as verified by plausibility ratings and cloze norming. Crucially, however, *menu* is equally associated to the context in both conditions. No N400 difference was observed between these conditions, indicating expectancy does not necessarily modulate N400 response, a phenomenon which has been observed in multiple studies, perhaps most notably role reversal anomalies (e.g. Hoeks et al., 2004; Kim and Osterhout, 2005; Kuperberg et al., 2007; see Brouwer et al., 2012 for a review). Indeed, an N400 effect of expectancy was only observed when the unexpected target was also unassociated, as is the case in condition Assoc–Exp–.

Cases such as these, in which contextual association modulates the N400 to a greater extent than expectancy are generally challenging for surprisal – and thus language models – to explain: In order to capture the observed absence of an N400 effect, the model would need to assign a similar probability to the target word in conditions Assoc+Exp+ and Assoc+Exp–. Indeed, due to its sensitivity to association – as well as the observation that the N400 is generally insensitive to words that are syntactically unexpected (Gouvea et al., 2010) – it could therefore be debated whether the N400 should be considered a reliable index of true surprisal, as surprisal is formally defined as a measure of the likelihood of a word that can immediately follow a given context. Conversely, any operationalization of surprisal that captures this absence of an expectancy effect, must either have not learned the role of world/event-knowledge constraints on expectations (see e.g., Kauf et al., 2023), or be influenced by association in a manner which is inconsistent with the goal of minimizing perplexity (see e.g., Cong et al., 2023; Michaelov and Bergen, 2022).

Another salient ERP component that is sensitive to the context-driven expectancy is the P600. This component is elicited when words are unexpected due to syntactic (see Gouvea et al., 2010 for a review), semantic (see Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012; Kuperberg et al., 2007 for reviews) or pragmatic constraints (see Hoeks and Brouwer, 2014 for a review). Indeed, in the Delogu et al. (2019) study above, a P600 effect was observed in both unexpected conditions compared to the expected condition, when P600 amplitude was corrected for its overlap with the N400 (see Brouwer et al., 2021a; Delogu et al., 2021, 2025). Moreover, Aurnhammer et al. (2023) have recently shown that the P600 is sensitive to graded plausibility-driven expectancy, consistent with varying degrees of surprisal. Combined with the fact that the P600 is insensitive to association (Aurnhammer et al., 2021), these findings suggest that the P600 may in fact be a better index of expectancy than the N400 (Brouwer et al., 2021b). Evidence supporting this hypothesis is, however, rather limited. Frank et al. (2015) did not find any effects of n-gram, RNN or PCFG surprisal on P600 amplitude (in a relatively early time window), but speculated that "more sophisticated systems are likely to be better capable at simulating cognitive processes" (p.9). Indeed, de Varda et al. (2023) did find LLM surprisal from different GPT models to be predictive of the P600 amplitude in the same dataset, while Xu et al. (2024) found LLM surprisal to be predictive of both the N400 and P600 in the context of joke comprehension.

Taken together, the above findings motivate the investigation of which ERP component – the N400 or the P600 – is best indexed by LLM surprisal. Importantly, the experimental manipulations of the studies we evaluate elicited a partially orthogonal pattern of N400 and P600 effects, such that if LLM surprisal is able to adequately model the N400, it cannot at the same also adequately capture the P600, and vice versa. More specifically, we pursue two objectives with this work. First, we aim to investigate N400 findings that appear challenging for LLM surprisal, that is, cases where association was shown to override the influence of expectancy, such that less expected targets did not elicit a stronger negativity. Second, we aim to test how well LLM surprisal predicts P600 modulations elicited by plausibility manipulations. We evaluate three German ERP studies that were specifically designed to disentangle the influences of association, plausibility and expectancy on the N400 and P600:

- **Study 1** (Aurnhammer et al., 2021) crossed association with expectancy, revealing additive effects of both factors in the N400, but only an effect of expectancy in the P600. To the extent that LLM surprisal is unaffected by association, we hypothesize that it captures the expectancy effects in both time windows, but not the association effect in the N400.
- **Study 2** (Delogu et al., 2019), discussed above, found that association can override expectancy in the N400, while both unexpected conditions elicited a P600 effect. Depending on whether LLM surprisal is sensitive to association, it may or may not predict N400 differences between the three conditions. Conversely, LLM surprisal should only be able to capture the P600 differences between the conditions if it is insensitive to association.
- **Study 3** (Aurnhammer et al., 2023) used repetition priming of the target word to achieve strong contextual association in all three conditions, such that no N400 effects were observed. By contrast, the graded implausibility of the conditions elicited an increasing P600 response. If LLM surprisal reflects graded plausibility, and is insensitive to association, we expect it to capture the graded P600 response to plausibility. Conversely, to predict the absence of any N400 effects, the LLM must assign a similar probability to the target word in all conditions, despite their graded plausibility.

2. Method

Surprisal values are collected for the target words of all studies using three German state-of-the-art transformer models of different training data size and model complexity. The primary motivation for this is simply to identify robustness across models, and in the discussion we also consider the extent to which model parameters affect fit. Following previous approaches (e.g., Michaelov et al., 2024), we assess how well LLM surprisal overall predicts mean amplitude across each time window in a linear mixed effects regression. While the original studies defined varying time intervals for the ERP components, we choose to uniformly operationalize the time windows, such that the N400 ranges from 300–500 ms, and the P600 ranges from 600–1000 ms. Turning then to a more detailed analysis, we apply regression-based ERPs (rERPs; Smith and Kutas, 2015), in which we fit one simple linear regression model per subject, electrode and time sample, predicting the observed voltages with LLM surprisal. This approach allows us to assess both the quantitative fit of LLM surprisal to the original data across time and electrodes, and the qualitative fit with the effect structure of the conditions in the original studies.

2.1. LLM selection and surprisal computation

Large language models exist in many different variations with regard to the specific details and complexity of their architecture, as well as to the amount and composition of textual data they have been trained on. We focus here on surprisal values computed with transformer-based models, which better predict ERP components compared to other architectures such as recurrent neural networks (de Varda et al., 2023; Merkx and Frank, 2021; Michaelov et al., 2021). A key feature of the transformer architecture is its *attention* mechanism, enabling the model to selectively weigh the influence of tokens from the context when predicting the next token (Vaswani et al., 2017). Attention can be applied in both directions of the target word position, allowing for language models that make use of both the preceding and subsequent context during training (e.g., Devlin et al., 2019). As bidirectional attention appears psychologically implausible for the purpose of modeling incremental language processing, we only consider strictly unidirectional transformer models that deploy a masked variant of attention, allowing them to attend only to the preceding context.

It has been previously argued, that model perplexity is inversely correlated with the goodness of fit of surprisal values to human data, such that models with lower perplexity generate surprisal values that provide a better fit (Goodkind and Bicknell, 2018; Wilcox et al., 2020). This hypothesis theoretically puts more advanced transformer models at an advantage, as larger models typically achieve lower perplexity. Oh et al. (2024), by contrast, found surprisal values from larger models to underestimate reading times for rare words due to frequency effects, suggesting that LLMs become overly accurate in predicting rare words. How model complexity, training data size and composition interact in influencing the fit of surprisal values to ERP components is not yet clearly established. We therefore consider three LLMs that differ with respect to their number of trainable parameters and amount of training data. For the purpose of replicability, we use LLMs that are publicly available via the Hugging Face platform (Wolf et al., 2020) and publish our code.[2]

Concretely, we use LeoLM, a Llama-2 model which was initialized with weights resulting from pre-training on English and which was then continued to be trained on a large German web corpus (Plüster, 2023).[3] Moreover, we use two GPT-2 models, GerPT-2 large and GerPT-2, that were also initialized with their respective English weights and were then trained on a different, smaller web corpus (Minixhofer, 2020).[4] That is, the GPT-2 models share the same training data and only differ with respect to their model complexity. For an overview of the specifications of the LLMs that were used, see Table 1. We note that a transfer from LLMs, which were pre-trained in English, to different languages is common practice due to economic and ecological considerations (see, e.g. Minixhofer et al., 2022).

Table 1
Overview of features of used LLMs.

	LeoLM	GerPT-2 large	GerPT-2
Parameters	13B	876M	176M
Vocabulary size	32,000	50,257	50,257
Context size	8192	1024	1024
Hidden layers	40	36	12
Hidden dimension	5120	1280	768
Attention heads	40	20	12
Training data size	595G	18G	18G
Training corpus	OSCAR-2301[a]	CC-100[b]	CC-100[b]

[a] https://huggingface.co/datasets/oscar-corpus/OSCAR-2301
[b] https://data.statmt.org/cc-100/

The stimulus materials are presented to the LLMs up until the target word. The target word itself is not part of the surprisal computation; its probability is collected from the output layer at the preceding word, to which a negative logarithm is applied. The LLMs we use here rely on tokenization schemes and vocabulary representations based on sub-words rather than words. Following previous work, when target words are tokenized into sub-words we sum the sub-word surprisal values to obtain a single surprisal value (see for example de Varda et al., 2023;Oh and Schuler, 2023b).[5]

2.2. LME analysis: Assessing overall fit of LLM surprisal to ERP amplitude

In previous studies, linear mixed effects models (LMEs) have been used to quantify the fit of surprisal values to ERP amplitude (e.g., Frank et al., 2015; Merkx and Frank, 2021; Michaelov et al., 2024). Usually, a null model is fitted, containing fixed effects that are known to have an overall influence on processing effort – such as word frequency, length or position within the sentence – and also random effects, accounting for variability specific to items, subjects and electrodes. Then, a model which additionally contains LLM surprisal as predictor is fitted and compared to the null model, for instance by computing Akaike's Information Criterion (AIC; Akaike, 1998) or conducting likelihood-ratio tests.

Concretely, we follow the approach of Michaelov et al. (2024). The authors used logarithmic word frequency and orthographic neighborhood size as fixed effects, and also included a random intercept for the target word in all models. Their approach is warranted, as their study implemented a target word manipulation design. That is, the target word varied within as well as across items. However, the studies we evaluate in this work feature a context manipulation design, under which the target words only vary between items. Therefore, we do not include target word as a random intercept, and we do not include orthographic neighborhood size as a fixed effect. For the purpose of baseline comparison, we still include logarithmic word frequency and also target word position within the target sentence as fixed effects. Word frequencies are obtained with the *WordFreq* package in Python (Speer, 2022), which is based on the Exquisite Corpus.[6] This corpus comprises different domains of text, which include Wikipedia, subtitles, news, books, web, and social media (Twitter and Reddit).

[2] https://github.com/benedict-krieger/llm-surprisal-rerps

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

[4] https://huggingface.co/benjamin/gerpt2

[5] We note that the commonly applied sub-word tokenization schemes may affect psycholinguistic modeling to a minor extent (see Oh et al., 2024; Nair and Resnik, 2023; Pimentel and Meister, 2024 for recent discussions).

[6] https://github.com/LuminosoInsight/exquisite-corpus

We include random intercepts for subject, item and electrode, but no random slopes (see Michaelov et al., 2024). For further comparison, we also fit a model with condition instead of LLM surprisal as fixed effect. In sum, each model contains the same random effects, fixed effects of word frequency and target word position, and then either condition or LLM surprisal from one of the three LLMs as an additional fixed effect. Except for condition, all fixed effects are standardized. For each of the time intervals we use the LMEs to predict mean N400 and P600 amplitude on the trial level, recorded from the set of 26 electrodes which is shared in all studies. We then compute AIC values for all fitted models and normalize them by the null model AIC. This allows us to compare the overall predictive power of surprisal values from different LLMs relative to the effect of condition per study and time window. In order to assess statistical significance of the surprisal predictors, we run likelihood-ratio tests, comparing each of the LMEs which include suprisal to the null regression LME.

2.3. rERPs: Assessing LLM surprisal across time and electrodes

While the methodological approach outlined above is well-suited to quantify the fit of LLM surprisal to naturalistic data, or to compare different predictors to each other, we aim to complement it with a more fine-grained analysis. That is, we wish to assess whether LLM surprisal can model each of the N400 and P600 effects observed in the original studies. In order to do so, we apply the regression-based ERP method (rERPs; Smith and Kutas, 2015). For every subject at every electrode, timestamp, and trial, the observed voltage is replaced by the estimate of a simple linear regression model. This estimated voltage is a linear combination of stimulus properties of the particular trial, which may for example be operationalized by human ratings. In our approach, we are interested in re-estimating the voltages based on the isolated effect of LLM surprisal. Therefore, the regression model as specified below only contains surprisal from one of the three language models as single predictor (apart from the intercept):

$$y = \beta_0 + \beta_1 \text{surprisal} + \varepsilon \tag{2}$$

Surprisal values are standardized (cf. Brouwer et al., 2021a). Both β_0, which denotes the intercept term, and the surprisal coefficient β_1 are determined by the least-squares principle. Each trial belongs to a certain condition of an item, and thus has a particular surprisal value associated with it. The regression will find coefficients for β_0 and β_1 which minimize the residual term ε across all trials for a given combination of subject, electrode, and timestamp. The fitted regression models are then used to compute trial-level voltage estimates, resulting in a new dataset of estimated voltages, which has the same dimensionality as the dataset of observed voltages. Analogous to the traditional ERP analysis procedure, these forward estimates are then grouped by condition, and first averaged within subjects, resulting in one estimate per subject, electrode, time sample and condition. Then, the estimates are averaged once more across subjects, to obtain one mean estimate per electrode and time sample in each condition. It is important to note, that in this way the linear models do not have access to condition-coded predictors and the estimates are only averaged per condition retrospectively. The grand-average estimates can then be plotted, allowing us to visually inspect how closely the forward estimates of the linear models incorporating surprisal approximate the observed voltages – in each condition, at each electrode and at each latency. Moreover, this fit is described by the residual error term ε of the trial-level models, which can be averaged and visualized in the same way as described above, allowing to evaluate how far off the forward estimates are at a given latency and whether they are too negative or positive, relative to the observed voltages. The more these average residuals per condition approximate zero, the better the fit of the rERP analysis.

Following Aurnhammer et al. (2023), we assess the significance of the surprisal predictor by computing the same models as specified in Eq. (2), but across subjects, instead of within subjects. This allows us to obtain a single t- and p-value per electrode and time sample. Due to the problem of multiple comparisons, we correct the p-values for the inflated false discovery rate by applying the method proposed by Benjamini and Hochberg (1995). The p-values are corrected within the N400 and P600 time windows defined in Section 3, and at the nine central electrodes (F3, Fz, F4, C3, Cz, C4, P3, Pz, P4), of which we report Pz. Importantly, however, while these p-values only reflect the overall fit of predictors across all trials, they are not indicative of the qualitative fit to the effect structure, and should therefore be interpreted with caution. That is, predictors may reach significance, but not adequately replicate the effect structure. Conversely, a predictor that does not reach significance, may still contribute to modeling the effect structure.

2.4. rERPs: Correcting for component overlap

While the design of **Study 2** will be described in more detail in Section 3, we note that the study became subject to a phenomenon for which we need to adjust our methodology: *component overlap*, that can occur when negative and positive components, that may temporally overlap to some degree, cancel each other out in the scalp-recorded signal (Brouwer and Crocker, 2017; Luck, 2005). In **Study 2**, the manipulation of two factors led to the observed effect structure: association modulated the N400, and expectancy – operationalized through plausibility – modulated the P600. Importantly, the decreased association in the Assoc– Exp– condition (see Ex. 1) elicited a negative response in the N400, which was so strong that it concealed a subsequent positivity elicited by the decreased plausibility in the observable *waveform-based* component structure (see Brouwer et al., 2021a for methodological and Delogu et al., 2021, 2025 for empirical evidence).

One advantage of the rERP method is, that it permits the direct modeling of the latent contribution of stimulus properties to the measured voltages, as described in Section 2.3. Brouwer et al. (2021a) showed how association and plausibility – which were inverted and standardized – linearly combine in re-estimating the originally observed voltages with the following model specification:

$$y = \beta_0 + \beta_1 \text{plausibility} + \beta_2 \text{association} + \varepsilon \tag{3}$$

Computing forward estimates with the fitted model results in a replication of the originally observed effect structure of **Study 2**. Critically, using this fitted model, one can neutralize the influence of a predictor on these estimates, by setting this predictor to its mean, thereby keeping its influence on the estimates for each trial constant. Setting association to its mean, and thus isolating the influence of plausibility, revealed that the P600 amplitude was indeed modulated by plausibility in the latent component structure (Brouwer et al., 2021a; Delogu et al., 2021), showing an increased P600 amplitude for the implausible conditions relative to the plausible baseline. Thus, in order to enable a fair comparison for LLM surprisal, we follow the same approach and re-estimate the observed data, setting association to its mean, hence, isolating the influence of plausibility. In the P600 window of **Study 2**, we evaluate LLM surprisal on the re-estimated data separately, both in the LME analysis, in which we predict overall fit in the time window, and in the rERP analysis, in which we predict differences between conditions across time and electrodes.

3. Results

We start by reporting the results from the linear mixed effects regression for all studies and both time windows. Then, we continue to present the results of our rERP analysis per study. First, we briefly introduce the original experimental design and findings – presenting the conditions, their mean ratings, an example item and the observed ERPs.

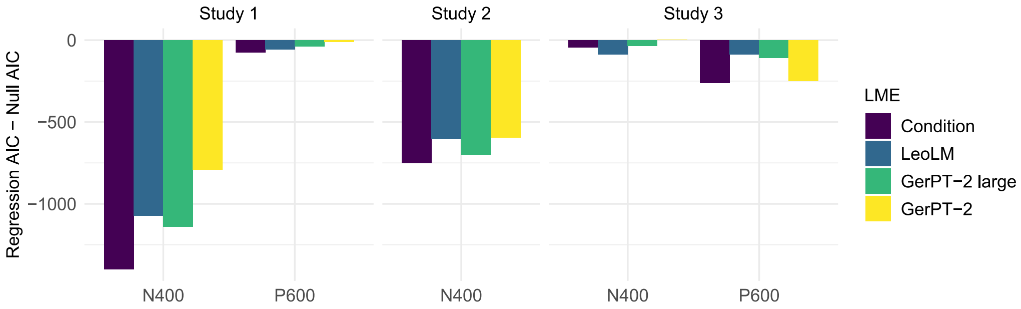


Fig. 1. AICs of linear regressions predicting N400 and P600 amplitude, normalized by the AIC of the null regression.

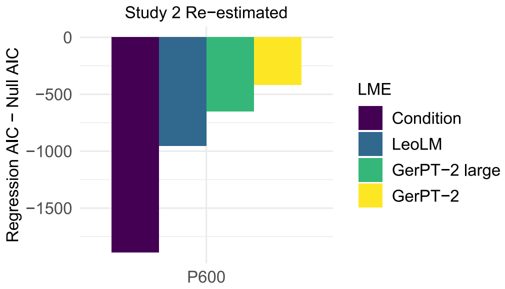


Fig. 2. AICs of linear regressions predicting P600 amplitude on the data of Study 2 which was corrected for component overlap. Note the difference in y-axis scale compared to Fig. 1.

The two factors, which were manipulated across studies, are contextual semantic association and expectancy. In **Study 1** and **Study 2**, association was operationalized via human ratings collected in a separate norming study. Participants rated the semantic relatedness between the target word and one or multiple context words on a Likert-scale ranging from 1 (weak) to 7 (strong). In **Study 3**, a strong association in all three conditions was achieved via repetition priming of the target word. **Study 1** operationalized expectancy via cloze probability. That is, in a separate study, participants completed the stimuli which they saw up until, but excluding the determiner of the target word. **Study 2** operationalized expectancy via cloze probability and plausibility. Cloze data was collected analogously to **Study 1**. Plausibility was operationalized by participants rating the plausibility of the stimuli up until, and including, the target word on a Likert-scale from 1 (weak) to 7 (strong). **Study 3** operationalized expectancy via cloze probability and plausibility, which were determined analogously to **Study 2**. We re-label the conditions across studies, such that they reflect the manipulation of association (Assoc+/−) and expectancy (Exp+/−/−−), the latter originally being operationalized through cloze or plausibility.

After presenting the original studies, we inspect the distribution of raw surprisal values grouped by condition, which allows us to reason a priori which types of ERP differences they may be able to capture. We then present the rERP forward estimates of the linear regression models using LLM surprisal as predictor, as specified in Eq. (2), and evaluate the qualitative fit of the re-estimated voltages to the observed voltages in the N400 and P600 window. Moreover, the average residual errors per condition of the forward estimates allow us to also assess this fit quantitatively. As a general observation, these residuals indicate that across studies and LLMs, ERP differences in both time windows are underestimated. For observed voltages, rERP forward estimates, and residuals, we present confidence intervals. We also report t- and p-values, which were computed and corrected as described in Section 2.3, but as noted earlier these may not reflect the quality of fit with the observed effect structure, which is the focus of the rERP analysis. We restrict our report to electrode Pz which was most responsive to the N400 and P600 effects in the studies examined here.

3.1. Assessing overall fit of LLM surprisal in both time windows

Fitting the LMEs per study and time window, as described in the Method section, leads to normalized AIC values, which are visualized in Fig. 1. Since the effect structure in the P600 window of **Study 2** was affected by component overlap, Fig. 1 only displays the AICs for the N400 window in this study. For the P600 window, we re-estimate the observed voltages (as described earlier) and fit the LMEs with the same model specifications to the re-estimated data. The AIC values for this separate set of LMEs are displayed in Fig. 2.

Following Michaelov et al. (2024), we also assess the significance of the fixed effects by conducting likelihood-ratio tests: we compare each of the models, which contain either condition or surprisal from one of the LLMs, to the null model which only contains word frequency and target word position as fixed effects. All predictors are significant in both time windows and in all studies (all $ps < 0.05$), except for GerPT-2 surprisal, which is not significant in the N400 time window of **Study 3**: $\chi^2(1) = 0.17$, $p = 0.68$.

Inspecting the AIC values, normalized by the null model, we observe that the LMEs including condition as fixed effect generally result in the lowest AICs, indicating the best fit to the data. An exception is the N400 window in **Study 3**, in which all AICs are close to zero. This result is unsurprising, since in this study, no N400 effects were elicited. In the P600 time window, surprisal of the smallest LLM, GerPT-2, produces the lowest AICs. For **Study 1** and **Study 2**, GerPT-2 large surprisal yields the lowest AICs in the N400 window and LeoLM surprisal yields the lowest AICs in the P600 window.

Crucially, although these results allow for an evaluation of which LLM produces surprisal values that best predict mean N400 and P600 amplitude in each of the studies, this analysis alone does not allow us to assess why this is the case. While these results reveal that LLM surprisal is a significant predictor of mean ERP amplitude in almost all time windows across all studies, we will now turn to a more fine-grained rERP analysis, which shows that LLM surprisal not only underestimates ERP differences, but also in multiple cases fails to model them qualitatively.

3.2. Additive effects of association and expectancy in the N400

Study 1 crossed association with expectancy, finding that both can additively modulate the N400 amplitude, whereas only expectancy modulated P600 amplitude. Fig. 3 shows an example item, mean association ratings and cloze probabilities across items, and the observed ERPs. Expectancy was manipulated through the selectional restrictions of the main verb: “sharpened ... the *axe*” in the high expectancy conditions Assoc+Exp+ and Assoc−Exp+ and “ate ... the *axe*” in the low expectancy conditions Assoc+Exp− and Assoc−Exp−. Association was manipulated through the lexical content of an intervening adverbial clause: “... before he the wood stacked, the *axe*” in the strongly associated conditions Assoc+Exp+ and Assoc+Exp− and “... before he the movie watched, the *axe*” in the weakly associated conditions Assoc−Exp+ and Assoc−Exp−.

Condition	Assoc.	Plaus.	Cloze	Example Item
Assoc+Exp+	6.29	-	0.67	Yesterday sharpened the lumberjack, before he the wood stacked, the axe...
Assoc−Exp+	2.09	-	0.64	Yesterday sharpened the lumberjack, before he the movie watched, the axe...
Assoc+Exp−	6.29	-	0.008	Yesterday ate the lumberjack, before he the wood stacked, the axe...
Assoc−Exp−	2.09	-	0.008	Yesterday ate the lumberjack, before he the movie watched, the axe...

Mean association ratings between adverbial clause noun and target, cloze probabilities for target. English translation preserving German word order.

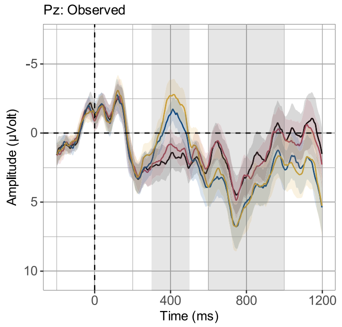


Fig. 3. **Study 1** (40 participants), experimental conditions, mean human ratings across items, example item and observed ERPs.

In the N400 window, the stimuli elicited additive modulations from both expectancy and association. Keeping one of the two properties constant, a decrease in the other led to an increased negativity. Crucially, in the P600 window a main effect of expectancy was observed, but the manipulation of association did not elicit a difference. Hence, in order to model the N400 response in the rERPs, LLM surprisal values need to reflect sensitivity to both association and expectancy, but – in contrast – they are required to be insensitive to association to capture the P600. Fig. 4 shows surprisal densities in the first, the rERP forward estimates in the second, residuals in the third, and t- and p-values in the last row.

N400. Inspecting the densities in the top row of Fig. 4, we can observe that surprisal values by all three LLMs appear to reflect sensitivity to both association and expectancy. For LeoLM and GerPT-2 large the contrast of unexpected versus expected (Assoc+Exp– & Assoc–Exp– vs. Assoc+Exp+ & Assoc–Exp+) is more pronounced than the contrast of un-associated versus associated (Assoc–Exp– vs. Assoc+Exp–, Assoc–Exp+ vs. Assoc+Exp+). This is not the case for GerPT-2 surprisal values. Consequently, we observe that the rERP forward estimates (Fig. 4, middle row) match the observed ERPs (Fig. 3) qualitatively well when entering LeoLM or GerPT-2 large surprisal into the regression models.

However, the surprisal values from LeoLM show only a small difference between unassociated versus associated in the expected conditions (Assoc–Exp+ vs. Assoc+Exp+). Therefore, unlike in the observed pattern, hardly any N400 difference is predicted between these conditions in the rERPs. GerPT-2 surprisal values appear to reflect the contrast of association well, but the contrast of expectancy however only to a smaller extent when compared to the other LLMs. In the rERPs, surprisal values from this LLM only capture the overall ordering of differences, and provide the worst fit to the human data, which is also reflected in the largest residuals for GerPT-2 in both time windows.

P600. The contrast of the higher mean surprisal in the unexpected conditions relative to the lower mean surprisal in the expected conditions – observable in the densities of LeoLM and GerPT-2 large – leads to rERP forward estimates that predict a difference of expectancy in this time window. This prediction matches the observed ERPs (Fig. 3). However, the additional sensitivity to association, which is reflected in the surprisal values, leads to the prediction of a small P600 difference of association in the unexpected conditions (Assoc–Exp– vs. Assoc+Exp–). Entering surprisal values from GerPT-2 leads to rERPs that predict only minimal P600 differences between all conditions.

Summary. Surprisal by LeoLM and GerPT-2 large was a significant predictor throughout most time samples in both time windows, while surprisal by GerPT-2 was only significant in the N400 and initial time samples in the P600. In the N400 window, however, **Study 1** revealed additive influences of association and expectancy, i.e., decreasing either association or expectancy led to a stronger negativity when keeping the other property constant. This means modeling this response requires LLM surprisal values to reflect sensitivity to not only expectancy but association as well – which is what we observe in the densities for all three LLMs. While resulting in rERP forward estimates that overall match the observed voltages qualitatively (LeoLM and GerPT-2 large), this raises the question of how well LLMs estimate true surprisal, as true surprisal is insensitive to association. By contrast, P600 amplitude was shown here to be sensitive to expectancy but insensitive to association, as only an effect between the expected and unexpected conditions was observed. While in the rERPs this difference of expectancy is captured best with surprisal from the larger LLMs (LeoLM and GerPT-2 large), their additional sensitivity to association leads to the prediction of a small difference of association in the unexpected conditions.

3.3. Association overrides expectancy in the N400 but not P600

Study 2 showed, that in the N400 time window, association and expectancy may not necessarily lead to additive effects. An example item of this study is presented in Fig. 5, alongside mean human judgments of association, plausibility and cloze across items. Central to the design is a manipulation of plausibility that is determined by world event knowledge: while entering a restaurant and then opening a menu is plausible (Assoc+Exp+), leaving a restaurant and then opening a menu is implausible and less expected (Assoc+Exp–), which also holds for entering an apartment and opening a menu (Assoc–Exp–). However, the conditions Assoc+Exp+ and Assoc+Exp– share the same strong contextual association between the prime noun *restaurant* and the target *menu*, while *apartment* and *menu* in condition Assoc–Exp– are only weakly associated. Crucially, no N400 effect was observed between Assoc+Exp+ and Assoc+Exp–, despite the decreased plausibility and cloze probability of *menu* in Assoc+Exp–. Only Assoc–Exp–, being both implausible and weakly associated, elicited a stronger negativity relative to the other conditions. While the implausible condition Assoc+Exp– did not elicit an N400 effect relative to Assoc+Exp+, due to the strong contextual association in both conditions, a P600 effect was observed instead. Although Delogu et al. (2019) predicted a centroparietal P600 effect in the other implausible condition Assoc–Exp– relative to Assoc+Exp+ as well, such an effect was only observed at occipital electrodes, due to the sustained negativity which extended into the P600 window. When correcting for spatio-temporal overlap, the predicted P600 effect was observed (Brouwer et al., 2021a; Delogu et al., 2021).

If mean LLM surprisal differs sufficiently in conditions Assoc+Exp+ and Assoc+Exp–, it will predict a difference between these conditions in the rERPs, which does not match the observed data. If mean LLM surprisal does not differ between these conditions, no difference will be predicted in the rERPs. This would match the observed data but also show that LLM surprisal is sensitive to association and not purely estimating surprisal. In order to assess LLM surprisal in the P600 window, a correction of component overlap is required first, showing a positivity in both conditions Assoc+Exp– and Assoc–Exp– relative to Assoc+Exp+. Predicting a positivity in Assoc+Exp– and Assoc–Exp– relative to Assoc+Exp+ consequently requires mean surprisal to be higher in these conditions.

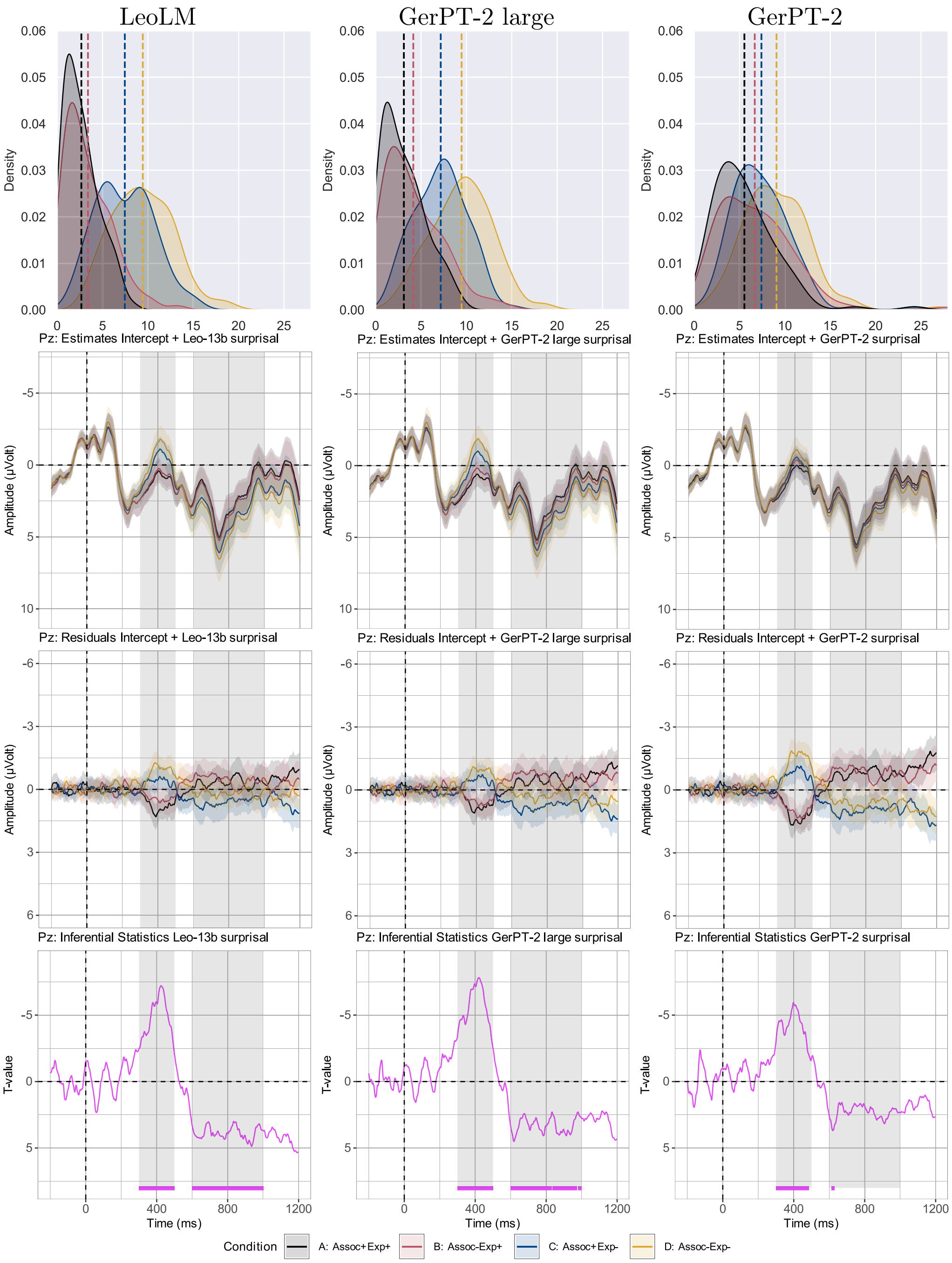


Fig. 4. **Study 1** surprisal densities (row 1), rERP forward estimates (row 2), rERP residuals (row 3), t-values and significant corrected p-values (row 4).

Condition	Assoc.	Plaus.	Cloze	Example Item
Assoc+Exp+	6.32	6.28	0.38	John entered the restaurant. Before long, he opened the menu...
Assoc+Exp−	6.32	2.42	0.13	John left the restaurant. Before long, he opened the menu...
Assoc−Exp−	1.56	1.93	0.008	John entered the apartment. Before long, he opened the menu...

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

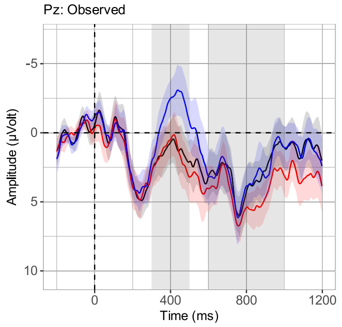


Fig. 5. **Study 2** (26 Participants), experimental conditions, mean human ratings across items, example item and observed ERPs.

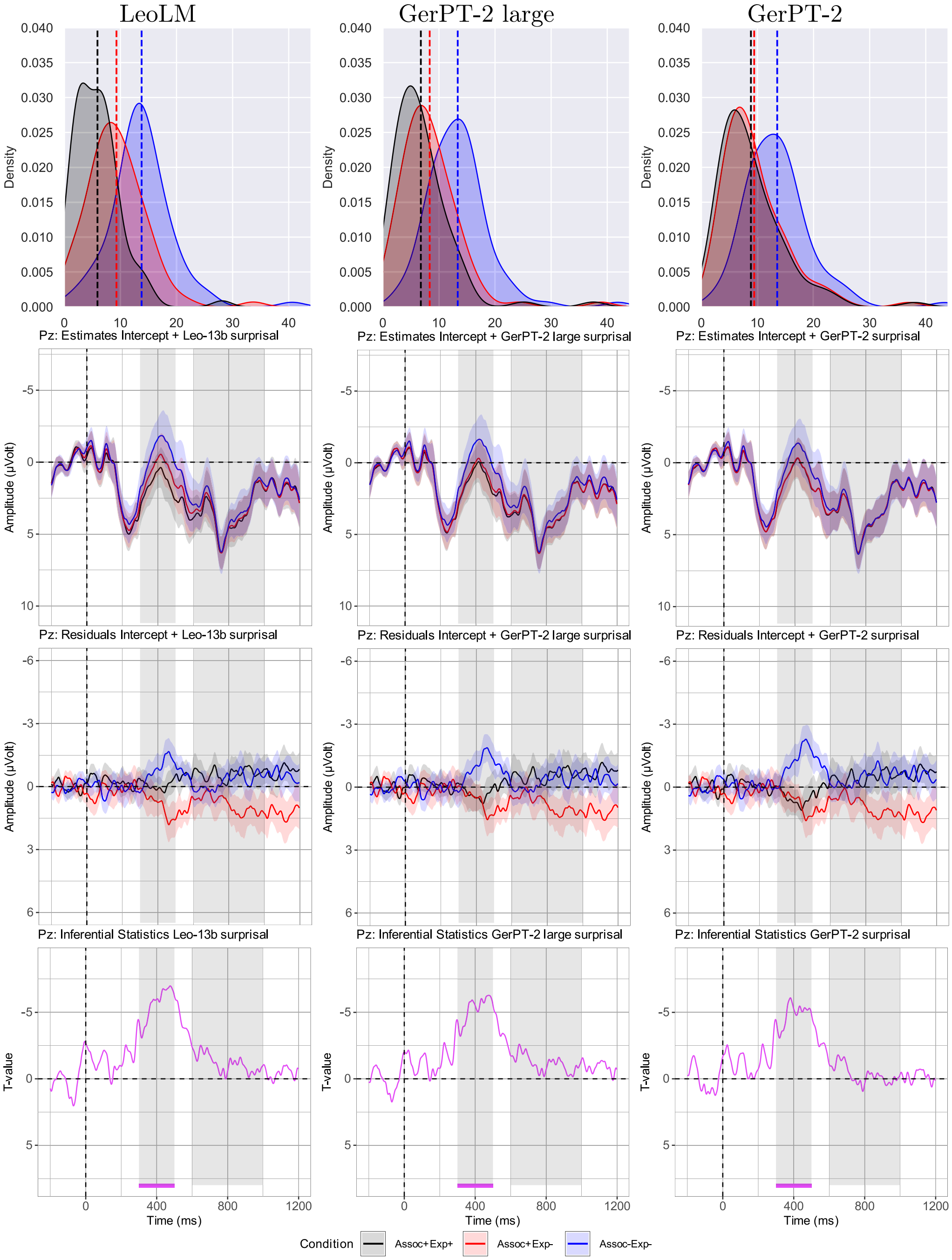


Fig. 6. **Study 2** surprisal densities (row 1), rERP forward estimates (row 2), rERP residuals (row 3), t-values and significant corrected p-values (row 4).

N400. As can be observed in the top row of Fig. 6, the LLMs diverge in how strongly their mean surprisal differs between conditions Assoc+Exp+ and Assoc+Exp−. There is a noticeable difference for LeoLM, which is less pronounced for GerPT-2 large and almost unnoticeable for GerPT-2. Consequently, computing rERP forward estimates with LeoLM surprisal values leads to the prediction of a stronger negativity in the N400 in condition Assoc+Exp− relative to Assoc+Exp+, aligning with judgments of plausibility and expectancy, even though no N400 but rather a P600 effect was observed between these conditions (see Fig. 5). By contrast, entering GerPT-2 surprisal into the rERP analysis does not predict any difference between the conditions, which is in line with association and the observed ERPs. GerPT-2 large appears to fall in between the other LLMs, that is, its surprisal values lead to only a minimally stronger negativity. In sum, the largest LLM (LeoLM) produces surprisal values patterning with plausibility and expectancy – predicting a difference in Assoc+Exp− relative to Assoc+Exp+ in the rERPs which was not observed in the ERP data – while using the smaller GPT-2 models (GerPT-2 large and GerPT-2) leads to surprisal values patterning with association, predicting no difference between Assoc+Exp+ and Assoc+Exp− in the rERPs – which is the pattern that was observed.

P600. A positivity in Assoc+Exp−, but not Assoc−Exp−, relative to Assoc+Exp+ was observed in the original data, even though both conditions were implausible. Thus, modeling the observed data would require LLM surprisal to be high in the implausible condition Assoc+Exp−, and lower in the other implausible condition Assoc−Exp− as well as the plausible baseline Assoc+Exp+. As confirmed by Brouwer et al. (2021a) and Delogu et al. (2021, 2025), the absence of the positivity in Assoc−Exp− was due to component overlap: the preceding

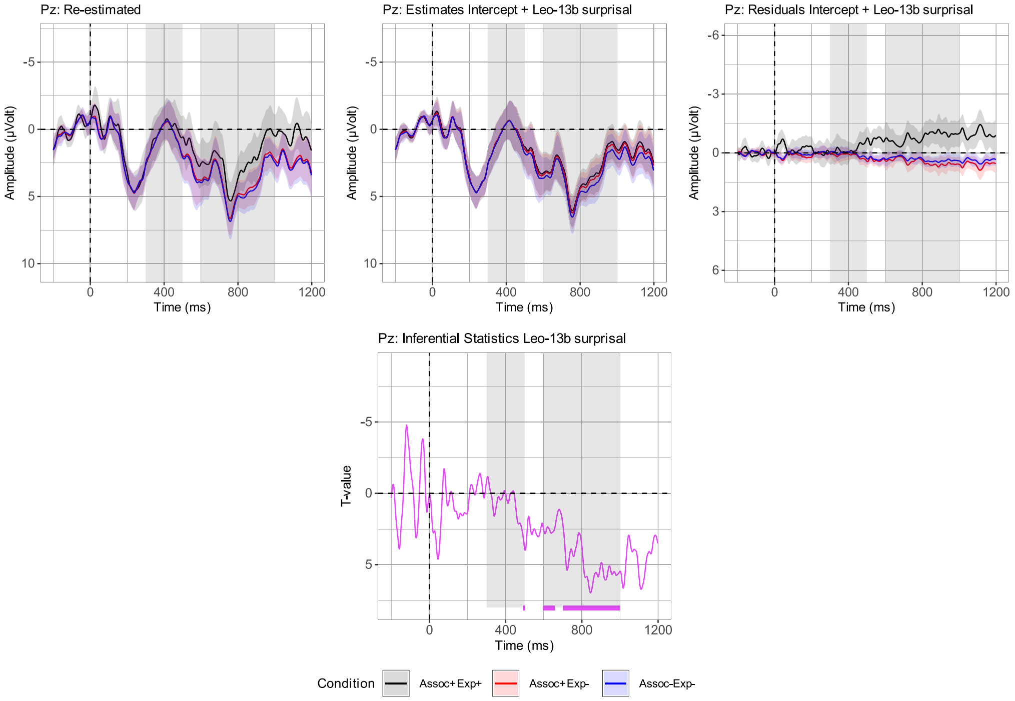


Fig. 7. **Study 2** (26 participants), rERPs re-estimating observed data with the model specified in Eq. (3) with association set to its mean (left), rERPs re-estimating this data with LeoLM surprisal (middle) and corresponding residuals (right).

strong negativity in this condition sustained throughout the ERP epoch and concealed the subsequent positivity. To account for this phenomenon, we first isolate the influence of plausibility and re-estimate the data, as described in the Method section. This reveals a positivity in both implausible conditions Assoc+Exp− and Assoc−Exp− relative to Assoc+Exp+, as can be seen in the left panel of Fig. 7. The rERP analysis is then conducted on the re-estimated data, potentially allowing LLM surprisal to predict the increased positivity in the implausible conditions in this time window.

We only present the forward estimates and residuals for LeoLM in Fig. 7, where close inspection reveals a minimally stronger positivity in condition Assoc−Exp−. This difference is hardly noticeable for the forward estimates based on the other LLM's surprisal values. Indeed, the residuals clearly show that the plausibility effect (Assoc+Exp− & Assoc−Exp− relative to Assoc+Exp+) is not adequately modeled in the corrected ERPs.

Summary. Surprisal by all three LLMs was a significant predictor throughout the N400 (see Fig. 6), and partially in the P600 time-window (see Fig. 7, where we focus our analysis on LeoLM). The N400 effect pattern observed in this study, however, poses a challenge for LLM surprisal: the less expected condition Assoc+Exp− did not elicit an increased negativity, due to the strong semantic association between the target and the context. Surprisal values obtained with LeoLM, our largest LLM in terms of model complexity and training data size, pattern with human judgments of plausibility and expectancy, thus predicting an increased negativity in the rERPs that was not observed in the N400, but rather in the P600. In contrast, surprisal values obtained with the smallest LLM (GerPT-2) pattern with association rather than expectancy, predicting the (observed) absence of this difference in the N400.

The implausible and associated condition Assoc+Exp− elicited a P600 relative to the plausible and associated baseline. The implausible and un-associated condition Assoc−Exp− did not lead to an observable P600 difference due to component overlap. Evaluating LLM surprisal on re-estimated ERPs that account for component overlap and show a positivity for both implausible conditions relative to the baseline, we find that none of the LLMs yields surprisal values that allow to capture this difference in the rERPs, although surprisal values from the largest LLM appear to predict a slightly increased positivity for the implausible and unassociated condition.

3.4. *The P600 as continuous index of plausibility-driven expectancy*

Study 3 found that repeated priming of the target word can lead to the absence of any N400 effects between gradually less plausible conditions. Note, that the pre-N400 negativity which can be observed in condition Exp−, can be explained by the high cloze probability of the distractor, and is hence akin to a mismatch negativity (see Aurnhammer et al., 2023 for a discussion). Instead, gradually decreased plausibility elicited an increasingly positive P600 amplitude. In this study, a context paragraph repeats the target word multiple times to maximally prime its meaning.[7] The subsequent target sentence then offers a continuation in which the target word is either plausible (Exp+), less plausible (Exp−) or implausible (Exp−−). See Fig. 8 for mean plausibility ratings, cloze probability and an example item. The graded plausibility manipulation is achieved by continuously decreasing the plausibility of the target word being the object of the preceding verb: "dismissed ... the *tourist*" Exp+ vs. "weighed ... the *tourist*" Exp− vs. "signed ... the *tourist*" Exp−−. Due to the repetition priming noted above, the association between the target and the context is equally strong in all three conditions and no N400 effects were elicited. Instead, the gradual decrease in plausibility led to an increasingly positive amplitude in the P600 window, with the strongest positivity for condition Exp−− > Exp− > Exp+. In order to model the N400 window in the rERPs, mean LLM surprisal would need to be equal in all three conditions. In contrast, to model the P600 window, mean surprisal should be highest in Exp−− and gradually lower in Exp− and Exp+.

N400. Inspecting the surprisal densities in the top row of Fig. 9, we observe that for all three LLMs, mean surprisal is higher with an increased spread in the less plausible and implausible conditions (Exp−

[7] The stimulus materials were based on materials by Nieuwland and Van Berkum (2005).

[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.*

Condition	Assoc.	Plaus.	Cloze	Example Item
Exp+	-	5.84	0.8	Then dismissed the lady the tourist...
Exp−	-	3.69	0.09	Then weighed the lady the tourist...
Exp−−	-	1.42	0.02	Then signed the lady the tourist...

Mean plausibility ratings for stimuli up until including the target, cloze probabilities for target.

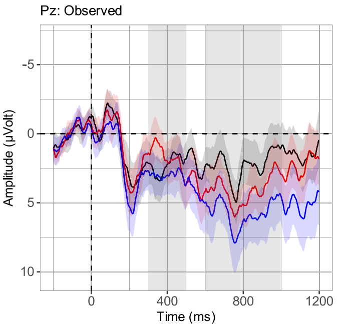
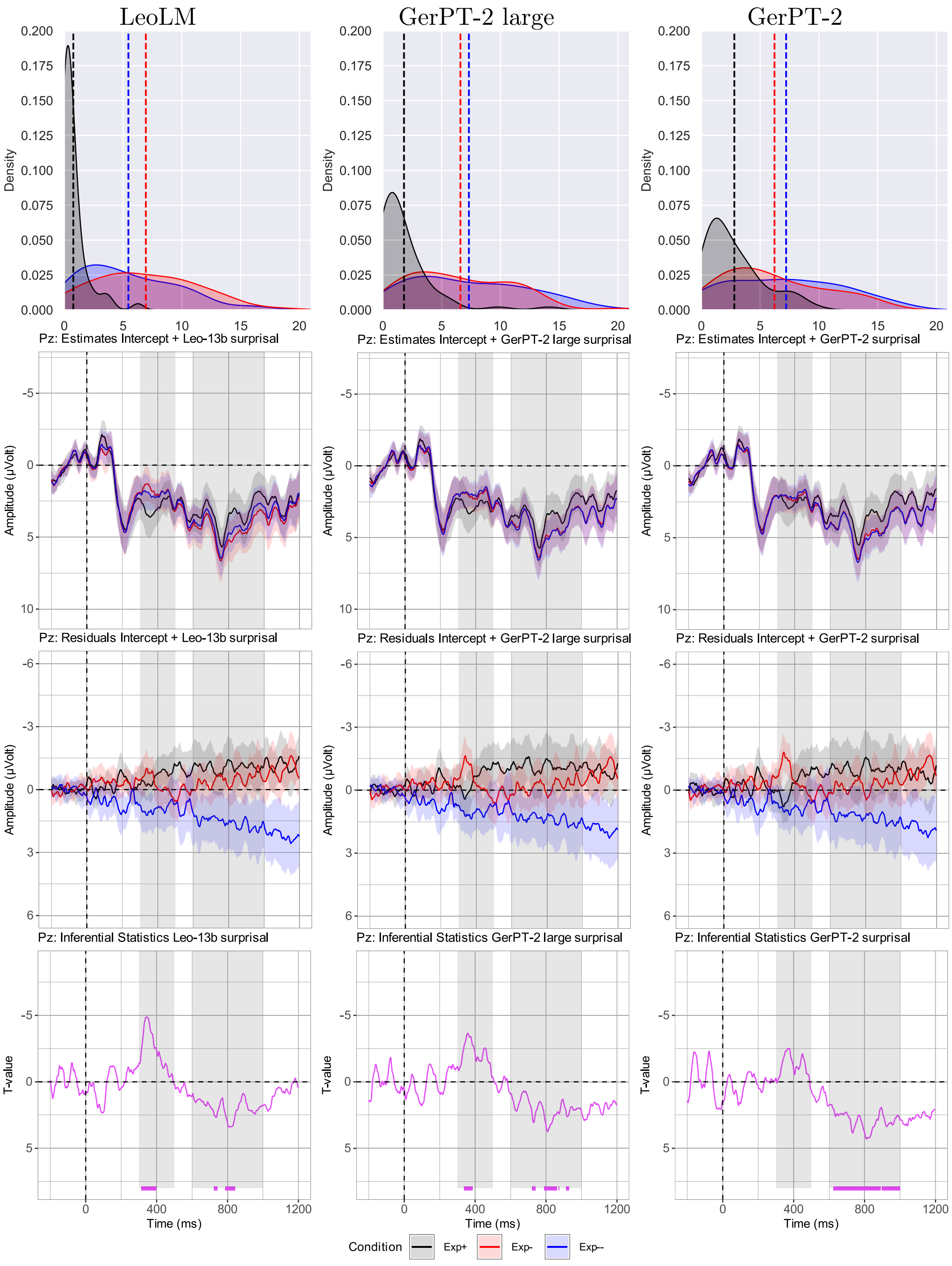

Pz: Observed

Fig. 8. **Study 3** (30 participants), experimental conditions, mean human ratings across items, example item and observed ERPs.


LeoLM

GerPT-2 large

GerPT-2

Pz: Estimates Intercept + Leo-13b surprisal

Pz: Estimates Intercept + GerPT-2 large surprisal

Pz: Estimates Intercept + GerPT-2 surprisal

Pz: Residuals Intercept + Leo-13b surprisal

Pz: Residuals Intercept + GerPT-2 large surprisal

Pz: Residuals Intercept + GerPT-2 surprisal

Pz: Inferential Statistics Leo-13b surprisal

Pz: Inferential Statistics GerPT-2 large surprisal

Pz: Inferential Statistics GerPT-2 surprisal

Fig. 9. **Study 3** surprisal densities (row 1), rERP forward estimates (row 2), rERP residuals (row 3), t-values and significant corrected p-values (row 4).

and Exp−−) relative to the plausible baseline Exp+.[8] In the rERPs, this leads to the prediction of an unobserved increased negativity for conditions Exp− and Exp−− when using surprisal computed with LeoLM and GerPT-2 large. This negativity is less noticeable for the forward estimates computed with GerPT-2 surprisal.

P600. As the densities in the top row of Fig. 9 show, the difference in mean LLM surprisal is small between conditions Exp− and Exp−−, although higher in both conditions relative to Exp+. Consequently, in the rERPs, LLM surprisal predicts an increased positivity for the conditions with decreased plausibility. However, the gradedness of this effect is not captured. There is no prediction of an increased positivity in condition Exp−− relative to Exp−.

Summary. Surprisal by LeoLM and GerPT-2 large was a significant predictor in some time samples in the N400. In the P600, surprisal by GerPT-2 was a significant predictor throughout most time samples, while surprisal by LeoLM and GerPT-2 large was a significant predictor in some time samples. However, no N400 effects were observed in the increasingly less plausible conditions relative to the plausible baseline, due to the repeated priming of the target word which leads to a strong association with the context. Modeling this absence of differences is again challenging for LLM surprisal, as it would require the LLM to assign a similar probability to the target word in all conditions, but a good language model should have learned to assign increasingly lower probability to increasingly unexpected continuations. LLM surprisal is higher in the conditions Exp− and Exp−− relative to Exp+. Consequently, the higher LLM surprisal for conditions Exp− and Exp−− still predicts a small negativity for the less plausible and implausible conditions, which was not observed. In the P600 window, the gradually decreased plausibility elicited a continuously increased amplitude (Exp−− > Exp− > Exp+). However, mean LLM surprisal only differs slightly between the less plausible and implausible conditions. Thus, in the rERPs it predicts a difference of plausibility but not the gradedness of this difference.

4. Discussion

There is broad empirical support for expectation-based accounts of language comprehension as formalized by surprisal in Eq. (1). Given the remarkable performance of recent large language models which directly operationalize surprisal, we sought to examine the relationship between LLM surprisal and two of the most salient ERP indices of language comprehension: the N400 and P600. Importantly, any empirical evaluation of the relationship defined in Eq. (1) faces a many-to-many mapping problem: there are multiple measures of cognitive effort, and many possible operationalizations of surprisal.

With regard to the left-hand side of Eq. (1), reading times, eye movements, or ERP components offer different measures of processing effort that may more or less reliably reflect the full extent of the difficulty which comprehenders experience when processing a word. The N400 and the P600, for example, both index processing in a manner which is differentially sensitive to word expectancy. Beyond contextual expectancy, the N400 is also sensitive to simple contextual association, to the extent that association can attenuate (Aurnhammer et al., 2021) and even override (Aurnhammer et al., 2023) any expectancy effects in this time window. This is consistent with Delogu et al. (2019), as well as ample evidence from reversal anomalies demonstrating the absence of an N400 for words that are unexpected based on all linguistic and world knowledge (see Kuperberg et al., 2007; Bornkessel-Schlesewsky and Schlesewsky, 2008; Brouwer et al., 2012 for reviews), and the broader insensitivity of the N400 to syntactically unexpected words, which are rather indexed by the P600, as discussed below (see Gouvea et al., 2010 for a review). Taken together, these findings present a serious challenge to the claim that the N400 indexes true surprisal. By contrast, the P600 is insensitive to association, and known to robustly respond to words that are unexpected based on morpho-syntactic, semantic, and pragmatic constraints (see Gouvea et al., 2010; Brouwer et al., 2012; Hoeks and Brouwer, 2014 for reviews). Furthermore, Aurnhammer et al. (2023) demonstrate that the P600 is continuously sensitive to plausibility grounded in world-knowledge. Therefore, we consider whether the P600 may provide a more reliable, yet underexplored, index of true surprisal compared to the N400.

Turning to the right-hand side of the equation sign in Eq. (1), LLMs based on transformer architectures (Vaswani et al., 2017) have become a focus of recent research into expectation-based processing. In addition to being trained on the vast amounts of linguistic data necessary to accurately reflect the distributional properties of language, their performance also suggests they capture plausibility – including aspects of world and event knowledge – suggesting they are better at approximating true surprisal than previous language model architectures. While surprisal estimates from LLMs have indeed been found to provide a close fit to a range of neurobehavioral processing indices (e.g., see de Varda et al., 2023), much of the supportive evidence is based on correlating LLM surprisal with language from naturalistic corpora. While such correlations underline the robustness of surprisal theory, our primary focus is on which ERP component provides the best index of true surprisal, as well as how well this is operationalized by LLMs. We therefore motivated the evaluation of three controlled datasets which are particularly revealing about the differential response of the two components to manipulations of expectancy and association, as these are not easily dissociable in naturalistic data. We analyzed the surprisal estimates from three current German LLMs on these datasets in order to gain insight into the generality of our findings.

4.1. LLM surprisal as predictor of the N400 or the P600

Since its discovery by Kutas and Hillyard (1980), the N400 has repeatedly been found to be sensitive to expectancy manipulations: unexpected words generally elicit an increased negativity compared to expected words. Consequently, previous research has tested whether N400 amplitude can be predicted by LLM surprisal on both naturalistic (e.g., Merkx and Frank, 2021) and experimental data (e.g., Michaelov et al., 2024). Overall, a significant negative correlation has been observed. However, among numerous linguistic and non-linguistic features, the N400 is also sensitive to contextual semantic association (Kutas and Federmeier, 2011). Association is not concerned with grammaticality or plausibility, i.e., surprising continuations that are ungrammatical and/or implausible may be highly associated with the context and hence attenuate processing effort reflected in the N400. Therefore, we argue that these findings challenge the role of the N400 as a reliable index of surprisal, as surprisal by definition is unaffected by simple association (Levy, 2008).

In the ERP studies which we evaluated, association either attenuated expectancy effects in the N400 (**Study 1**), or eliminated them completely (**Study 2 and 3**). LLM surprisal was able to capture the additive N400 effects of association and expectancy observed in **Study 1**. This finding implies that not only the N400 but also LLMs are to some extent sensitive to association. This result is also in line with previous results reported by Michaelov and Bergen (2022) about the stimuli of Metusalem et al. (2012): when two continuations were matched for a cloze probability of zero, the less associated continuation led to higher LLM surprisal.

Cases in which association eliminates expectancy effects remain challenging for LLMs, however. In **Study 2**, the condition Assoc+Exp− is less expected than Assoc+Exp+. Yet, no N400 effect of expectancy was observed between these conditions, due to the target word being highly associated with the context in both conditions. Only the condition that is both unassociated and unexpected (Assoc−Exp−) elicited an increased negativity relative to the other conditions. The same logic

[8] We note that LeoLM shows the highest mean surprisal for Exp− instead of Exp−−.

applies to the N400 data of **Study 3**, where the conditions are either expected, less expected, or unexpected, but the strong association of the target to the context across conditions eliminated any expectancy effects. Thus, to correctly model N400 amplitude, LLMs would need to assign equal likelihood to target words in each condition despite their varying expectancy. The extent to which LLM surprisal does this, however, would suggest a divergence from true surprisal.

Such a divergence could reflect the extent to which LLMs can learn the relevant world/event knowledge that determines the plausibility – and thus expectancy – differences of the conditions in **Study 2 and 3**, which would be necessary for them to assign lower probabilities to less plausible continuations. Kauf et al. (2023) investigated this question, and found that LLMs are consistently sensitive to sentences describing impossible states-of-affairs, but not necessarily to those describing possible but unlikely ones. While this suggests mixed sensitivity of current LLMs to plausibility based on world knowledge, one would expect better language models to capture such plausibility-driven expectations and be less affected by association. As LLMs are trained to minimize perplexity on naturalistic data, smaller LLMs tend to have a higher perplexity. As a consequence, such models may be poorer at distinguishing plausible and implausible words that have not been adequately observed during training, and we speculate that they may rather rely on simple association. Indeed, our results are in line with this, as in our analysis of **Study 2** we find that surprisal by the small and medium-sized GPT-2 models captures the observed absence of a difference best, while surprisal by the largest LLM predicts an unobserved negativity for Assoc+Exp− relative to Assoc+Exp+. Analogously, the smallest LLM predicts the least amount of difference between all conditions in the data of **Study 3**. The sensitivity of LLM surprisal to not only expectancy but also association found for controlled experimental data in the present study may thus explain the previously observed correlation between the N400 and LLM surprisal in naturalistic data.

We also assessed the ability of LLM surprisal to account for differences between conditions in the P600. The P600 has been found to be sensitive to syntactic, semantic and pragmatic expectancy (see Gouvea et al., 2010; Brouwer et al., 2012; Hoeks and Brouwer, 2014 for reviews). Crucially, as the P600 is insensitive to association, we hypothesize that this component offers a more direct index of true surprisal. In the ERP studies we evaluated, P600 effects were elicited by manipulating semantic plausibility, as determined by script-knowledge (**Study 2**), or by selectional restrictions of verbs (**Study 1 and 3**). LLM surprisal was able to account for the increased positivity elicited by the more salient expectancy manipulations in **Study 1** and **Study 3**. However, it could neither capture the graded plausibility effects in **Study 3**, nor the script-knowledge violations in **Study 2**. As discussed above, the inability of LLM surprisal to completely capture the P600 in these two controlled studies may be due to (a) the influence of association on LLM surprisal, and/or (b) variability in the ability of LLMs to learn relevant world/event-knowledge plausibility constraints on expectancy. This is consistent with our observation that the surprisal from the largest LLM – which is less sensitive to association and more likely to approximate true surprisal – performed best in capturing the P600.

While the aim of the present study was not to assess the specific parameters of the models considered, the differences in their performance naturally lead to the question of which LLM features contribute to their *psychometric quality*, that is, how well their surprisal estimates match neurobehavioral processing data (de Varda and Marelli, 2023; Wilcox et al., 2023a). Factors that have been shown to influence model behavior include the amount of training data the LLM has been exposed to, the number of its trainable parameters, hidden layers, attention heads and also the amount of previous context which it considers during prediction. While it was initially assumed that psychometric quality linearly increases with decreasing perplexity (Goodkind and Bicknell, 2018), further research found conflicting results. LLMs which are smaller with regard to their architecture, training data size or training duration, i.e., LLMs with higher perplexity, were found to provide a closer fit to human reading times than larger ones (Oh and Schuler, 2023b; Shain et al., 2024). Oh et al. (2024) argued that this effect may arise due to larger LLMs becoming overly accurate in predicting the probability of rare open-class words, thus underestimating reading times for these words. Moreover, Kuribayashi et al. (2022) found that LLMs which were more constrained in their context size, also led to a closer reading time fit. Furthermore, de Varda and Marelli (2023) observed that surprisal from LLMs of different sizes predict early versus late eye-tracking measures differentially well, such that earlier measures were better predicted by smaller, and later measures better predicted by larger LLMs. In sum, the fit of LLM surprisal to various neurobehavioral processing indices is complex, with numerous factors contributing to LLM performance, and any conclusions about the effect of LLM parameters on psychometric fit with ERPs remain to be explored in future experiments.

4.2. Reconciling LLM behavior with the functional interpretation of ERPs

The focus of the present investigation has been to assess the degree to which surprisal, as estimated at the output layer of an LLM, can explain observed N400 and P600 effects in language comprehension. Due to the partial orthogonality of N400 and P600 responses to various aspects of the next word, it is not possible for LLM surprisal to fully explain both of these components. Indeed, the differential sensitivity of the N400 and P600 – as exemplified by the studies evaluated here – have underpinned the functional interpretation of the components with regard to which mechanistic processes they index. Retrieval-Integration (RI) theory, for instance, posits that 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., 2012; Brouwer and Hoeks, 2013; Brouwer et al., 2017; Venhuizen and Brouwer, 2025). RI theory predicts facilitated retrieval of the meaning of the next word if it is contextually expected by, or associated with, the prior context, resulting in an attenuated N400 amplitude. The cost of integrating this retrieved word meaning into an unfolding utterance representation, by contrast, is predominantly determined by linguistic and plausibility constraints, such that unexpected words entail greater effort, resulting in larger P600 amplitude.

One approach to reconciling such a mechanistic account of the cognitive processes underlying the N400 and the P600, is to look for correlates of these processes and their assumed representation in different internal layers of an LLM. For instance, as earlier layers of an LLM are closer to the input word embeddings, the computational and representational dynamics at these layers may be closer to those underlying the N400. On the other hand, computations and representations closer to the final layers may be more reflective of utterance-level integrative processes, and thus better capture the computations and representations underlying P600. Indeed, this perspective is consistent with the hypothesis that more shallow LLM representations align better with earlier, and deeper representations better with late processing indices (Kuribayashi et al., 2025). This hypothesis could, for instance, be tested by applying the *tuned lens method* (Belrose et al., 2023), which reveals model predictions about upcoming input at different LLM layers, other than just the output layer.

Alternatively, methods from the field of mechanistic interpretability may be leveraged (see Rai et al., 2024 for a review). Research in this area has begun to shed light on the mechanistic roles of different LLM representations, such as the multilayer-perceptron layers and attention heads, and how they combine to form circuits which are specialized in fulfilling sub-goals during next-word prediction (Geva et al., 2023; Wang et al., 2022). On a final note, building on work by Kauf et al. (2023), probing methods may be harnessed to investigate the degree to which different LLMs layers are sensitive to, for instance, association, plausibility and world/event knowledge.

5. Conclusion

Expectations regarding the next word play a central role in language comprehension, such that listeners process expected words with greater ease than less expected ones. Surprisal theory formalizes this relationship by positing that the cognitive effort required to process a word is proportional to its negative log-probability in context. Critically, the true expectancy of a word should in principle reflect all relevant determinants of what words are likely to appear next — including distributional, linguistic, and world knowledge-based plausibility constraints. LLMs trained on next word prediction directly compute richly contextualized surprisal estimates which, when combined with their deceptively human-like language capabilities, has motivated their consideration as plausible models of human comprehension at both the computational and algorithmic level. We here evaluated the degree to which LLM surprisal aligns with the two most salient neural correlates of comprehension – the N400 and the P600 – which are differentially sensitive to the semantic association and contextual expectancy of a word.

Critically, while previous studies have established a link between the N400 and LLM surprisal, these results are predominantly based on evaluations of naturalistic data, in which association and expectancy are confounded. By focusing the present evaluation on factorial designs that tease apart association and expectancy, we challenge the validity of the N400 as an index of true surprisal. While we find that LLM surprisal does indeed not fully align with the N400, we do observe that it is sensitive to association, especially in smaller models. Hence, to the extent that LLM surprisal properly models the N400, it is in fact a poor model of true surprisal. Due to this sensitivity to association and/or the inability to learn world/event knowledge constraints – at least for the LLMs considered here – we find that LLM surprisal also does not fully align with the P600. Importantly, as the P600 is insensitive to association, patterns with expectancy in a continuous manner, and is more broadly sensitive to syntactic, semantic and pragmatic expectancy, we argue that the P600 is a better index of true surprisal. We therefore posit that – when the full complex of N400 and P600 responses as revealed by controlled manipulations is taken into account – we should regard the P600, rather than the N400, as the neural correlate of true surprisal. Indeed, we advocate more generally for the importance of evaluating LLMs against controlled experimental data that more fully reveal the sensitivity of relevant ERP components to a range of expectancy manipulations, thus complementing more naturalistic data. Finally, given our observations that the surprisal from the largest LLM performed best in capturing the P600, we hypothesize that better performing LLMs will be both less sensitive to association and better embody the full-range of linguistic and world knowledge constraints that determine true surprisal.

CRediT authorship contribution statement

Benedict Krieger: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Harm Brouwer:** Writing – review & editing, Visualization, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Christoph Aurnhammer:** Writing – review & editing, Visualization, Supervision, Software, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Matthew W. Crocker:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Acknowledgments

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

Data availability

Code and data required to reproduce the analyses are publicly available at https://github.com/benedict-krieger/llm-surprisal-rerps.

References

Akaike, H., 1998. Information theory and an extension of the maximum likelihood principle. In: Parzen, E., Tanabe, K., Kitagawa, G. (Eds.), Selected Papers of Hirotugu Akaike. Springer New York, New York, NY, pp. 199–213. http://dx.doi.org/10.1007/978-1-4612-1694-0_15.

Aurnhammer, C., Delogu, F., Brouwer, H., Crocker, M.W., 2023. The P600 as a continuous index of integration effort. Psychophysiology 60 (9), e14302. http://dx.doi.org/10.1111/psyp.14302.

Aurnhammer, C., Delogu, F., Schulz, M., Brouwer, H., Crocker, M., 2021. Retrieval (N400) and integration (P600) in expectation-based comprehension. PLoS One 16 (9), 1–31. http://dx.doi.org/10.1371/journal.pone.0257430.

Aurnhammer, C., Frank, S., 2019. Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. Neuropsychologia 134, 107198. http://dx.doi.org/10.1016/j.neuropsychologia.2019.107198.

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., Steinhardt, J., 2023. Eliciting latent predictions from transformers with the tuned lens. URL https://arxiv.org/abs/2303.08112, arXiv:2303.08112.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 289–300.

Bornkessel-Schlesewsky, I., Schlesewsky, M., 2008. An alternative perspective on "semantic P600" effects in language comprehension. Brain Res. Rev. 59 (1), 55–73. http://dx.doi.org/10.1016/j.brainresrev.2008.05.003.

Brouwer, H., Crocker, M.W., 2017. On the proper treatment of the N400 and P600 in language comprehension. Front. Psychol. 8, http://dx.doi.org/10.3389/fpsyg.2017.01327.

Brouwer, H., Crocker, M., Venhuizen, N.J., Hoeks, J., 2017. A neurocomputational model of the N400 and the P600 in language processing. Cogn. Sci. 41 (S6), 1318–1352. http://dx.doi.org/10.1111/cogs.12461.

Brouwer, H., Delogu, F., Crocker, M.W., 2021a. Splitting event-related potentials: Modeling latent components using regression-based waveform estimation. Eur. J. Neurosci. 53, 974–995. http://dx.doi.org/10.1111/ejn.14961.

Brouwer, H., Delogu, F., Venhuizen, N.J., Crocker, M.W., 2021b. Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. Front. Psychol. 12, 615538. http://dx.doi.org/10.3389/fpsyg.2021.615538.

Brouwer, H., Fitz, H., Hoeks, J., 2010. Modeling the noun phrase versus sentence coordination ambiguity in Dutch: Evidence from surprisal theory. In: Hale, J.T. (Ed.), Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics, Uppsala, Sweden, pp. 72–80, URL https://aclanthology.org/W10-2009.

Brouwer, H., Fitz, H., Hoeks, J., 2012. Getting real about semantic illusions: Rethinking the functional role of the P600 in language comprehension. Brain Res. 1446, 127–143. http://dx.doi.org/10.1016/j.brainres.2012.01.055.

Brouwer, H., Hoeks, J.C., 2013. A time and place for language comprehension: mapping the N400 and the P600 to a minimal cortical network. Front. Hum. Neurosci. 7, http://dx.doi.org/10.3389/fnhum.2013.00758, URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2013.00758.

Caucheteux, C., Gramfort, A., King, J.R., 2023. Evidence of a predictive coding hierarchy in the human brain listening to speech. Nat. Hum. Behav. 7 (3), 430–441. http://dx.doi.org/10.1038/s41562-022-01516-2, URL https://www.nature.com/articles/s41562-022-01516-2.

Caucheteux, C., King, J.R., 2022. Brains and algorithms partially converge in natural language processing. Commun. Biol. 5 (1), 134. http://dx.doi.org/10.1038/s42003-022-03036-1, URL https://www.nature.com/articles/s42003-022-03036-1.

Chomsky, N., 1965. Aspects of the Theory of Syntax, 50th ed. The MIT Press, URL http://www.jstor.org/stable/j.ctt17kk81z.

Cong, Y., Chersoni, E., Hsu, Y.Y., Lenci, A., et al., 2023. Are language models sensitive to semantic attraction? A study on surprisal. Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.starsem-1.13, URL https://aclanthology.org/2023.starsem-1.13/.

Contreras Kallens, P., Kristensen-McLachlan, R.D., Christiansen, M.H., 2023. Large language models demonstrate the potential of statistical learning in language. Cogn. Sci. 47 (3), e13256. http://dx.doi.org/10.1111/cogs.13256, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.13256.

de Varda, A.G., Marelli, M., 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Toronto, Canada, pp. 139–149. http://dx.doi.org/10.18653/v1/2023.acl-short.14, URL https://aclanthology.org/2023.acl-short.14/.

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. Behav. Res. Methods http://dx.doi.org/10.3758/s13428-023-02261-8.

Delogu, F., Aurnhammer, C., Brouwer, H., Crocker, M.W., 2025. On the biphasic nature of the N400-P600 complex underlying language comprehension. Brain Cogn. 186, 106293. http://dx.doi.org/10.1016/j.bandc.2025.106293.

Delogu, F., Brouwer, H., Crocker, M.W., 2019. Event-related potentials index lexical retrieval (N400) and integration (P600) during language comprehension. Brain Cogn. 135, 103569. http://dx.doi.org/10.1016/j.bandc.2019.05.007.

Delogu, F., Brouwer, H., Crocker, M.W., 2021. When components collide: Spatiotemporal overlap of the N400 and P600 in language comprehension. Brain Res. 1766, 147514. http://dx.doi.org/10.1016/j.brainres.2021.147514.

Demberg, V., Keller, F., 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition 109 (2), 193–210. http://dx.doi.org/10.1016/j.cognition.2008.07.008, URL https://www.sciencedirect.com/science/article/pii/S0010027708001741.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.

Earley, J., 1970. An efficient context-free parsing algorithm. Commun. ACM 13 (2), 94–102.

Ehrlich, S.F., Rayner, K., 1981. Contextual effects on word perception and eye movements during reading. J. Verb. Learn. Verb. Beh. 20 (6), 641–655. http://dx.doi.org/10.1016/S0022-5371(81)90220-6.

Ettinger, A., Linzen, T., Marantz, A., 2014. The role of morphology in phoneme prediction: Evidence from MEG. Brain Lang. 129, 14–23. http://dx.doi.org/10.1016/j.bandl.2013.11.004, URL https://linkinghub.elsevier.com/retrieve/pii/S0093934X13002216.

Fernandez Monsalve, I., Frank, S.L., Vigliocco, G., 2012. Lexical surprisal as a general predictor of reading time. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 398–408, URL https://aclanthology.org/E12-1041/.

Frank, S.L., Fernandez Monsalve, I., Thompson, R.L., Vigliocco, G., 2013. Reading time data for evaluating broad-coverage models of English sentence processing. Behav. Res. Methods 45 (4), 1182–1190. http://dx.doi.org/10.3758/s13428-012-0313-y, URL http://link.springer.com/10.3758/s13428-012-0313-y.

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 Lang. 140, 1–11. http://dx.doi.org/10.1016/j.bandl.2014.10.006.

Frank, S.L., Willems, R.M., 2017. Word predictability and semantic similarity show distinct patterns of brain activity during language comprehension. Lang. Cogn. Neurosci. 32 (9), 1192–1203. http://dx.doi.org/10.1080/23273798.2017.1323109.

Geva, M., Bastings, J., Filippova, K., Globerson, A., 2023. Dissecting recall of factual associations in auto-regressive language models. In: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp. 12216–12235. http://dx.doi.org/10.18653/v1/2023.emnlp-main.751, URL https://aclanthology.org/2023.emnlp-main.751/.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S.A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., Dugan, P., Melloni, L., Reichart, R., Devore, S., Flinker, A., Hasenfratz, L., Levy, O., Hassidim, A., Brenner, M., Matias, Y., Norman, K.A., Devinsky, O., Hasson, U., 2022. Shared computational principles for language processing in humans and deep language models. Nature Neurosci. 25 (3), 369–380. http://dx.doi.org/10.1038/s41593-022-01026-4.

Goodkind, A., Bicknell, K., 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In: Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics. CMCL 2018, Association for Computational Linguistics, Salt Lake City, Utah, pp. 10–18. http://dx.doi.org/10.18653/v1/W18-0102, URL https://aclanthology.org/W18-0102.

Gouvea, A.C., Phillips, C., Kazanina, N., Poeppel, D., 2010. The linguistic processes underlying the P600. Lang. Cogn. Process. 25 (2), 149–188. http://dx.doi.org/10.1080/01690960902965951.

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, pp. 1–8, URL https://aclanthology.org/N01-1021/.

Hoeks, J.C., Brouwer, H., 2014. Electrophysiological research on conversation and discourse processing. In: Holtgraves, T.M. (Ed.), The Oxford Handbook of Language and Social Psychology. Oxford University Press, http://dx.doi.org/10.1093/oxfordhb/9780199838639.013.024.

Hoeks, J.C., Stowe, L.A., Doedens, G., 2004. Seeing words in context: the interaction of lexical and sentence level information during reading. Cogn. Brain Res. 19 (1), 59–73. http://dx.doi.org/10.1016/j.cogbrainres.2003.10.022.

Hu, J., Levy, R., Degen, J., Schuster, S., 2023. Expectations over unspoken alternatives predict pragmatic inferences. Trans. Assoc. Comput. Linguist. 11, 885–901. http://dx.doi.org/10.1162/tacl_a_00579, URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00579/116994/Expectations-over-Unspoken-Alternatives-Predict.

Katzir, R., 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. Biolinguistics 17, e13153. http://dx.doi.org/10.5964/bioling.13153, URL https://bioling.psychopen.eu/index.php/bioling/article/view/13153.

Kauf, C., Ivanova, A.A., Rambelli, G., Chersoni, E., She, J.S., Chowdhury, Z., Fedorenko, E., Lenci, A., 2023. Event knowledge in large language models: The gap between the impossible and the unlikely. Cogn. Sci. 47 (11), e13386. http://dx.doi.org/10.1111/cogs.13386.

Kim, A., Osterhout, L., 2005. The independence of combinatory semantic processing: Evidence from event-related potentials. J. Mem. Lang. 52 (2), 205–225. http://dx.doi.org/10.1016/j.jml.2004.10.002.

Kuperberg, G.R., Kreher, D.A., Sitnikova, T., Caplan, D.N., Holcomb, P.J., 2007. The role of animacy and thematic relationships in processing active English sentences: Evidence from event-related potentials. Brain Lang. 100 (3), 223–237. http://dx.doi.org/10.1016/j.bandl.2005.12.006.

Kuribayashi, T., Oseki, Y., Brassard, A., Inui, K., 2022. Context limitations make neural language models more human-like. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, pp. 10421–10436. http://dx.doi.org/10.18653/v1/2022.emnlp-main.712, URL https://aclanthology.org/2022.emnlp-main.712/.

Kuribayashi, T., Oseki, Y., Taieb, S.B., Inui, K., Baldwin, T., 2025. Large language models are human-like internally. arXiv preprint arXiv:2502.01615.

Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62, 621–647. http://dx.doi.org/10.1146/annurev.psych.093008.131123.

Kutas, M., Hillyard, S.A., 1980. Reading senseless sentences: Brain potentials reflect semantic incongruity. Science 207 (4427), 203–205. http://dx.doi.org/10.1126/science.7350657, URL https://www.science.org/doi/abs/10.1126/science.7350657, arXiv:https://www.science.org/doi/pdf/10.1126/science.7350657.

Kutas, M., Hillyard, S.A., 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 161–163. http://dx.doi.org/10.1038/307161a0.

Levy, R., 2008. Expectation-based syntactic comprehension. Cognition 106 (3), 1126–1177. http://dx.doi.org/10.1016/j.cognition.2007.05.006.

Luck, S.J., 2005. An Introduction to the Event-Related Potential Technique. MIT Press.

Malisz, Z., Brandt, E., Möbius, B., Oh, Y.M., Andreeva, B., 2018. Dimensions of segmental variability: Interaction of prosody and surprisal in six languages. Front. Commun. 3, 25. http://dx.doi.org/10.3389/fcomm.2018.00025, URL https://www.frontiersin.org/article/10.3389/fcomm.2018.00025/full.

Manning, C.D., Clark, K., Hewitt, J., Khandelwal, U., Levy, O., 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. Proc. Natl. Acad. Sci. 117 (48), 30046–30054. http://dx.doi.org/10.1073/pnas.1907367117, URL https://pnas.org/doi/full/10.1073/pnas.1907367117.

Marr, D., 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. MIT Press.

Meister, C., Cotterell, R., 2021. Language model evaluation beyond perplexity. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, pp. 5328–5339. http://dx.doi.org/10.18653/v1/2021.acl-long.414.

Merkx, D., Frank, S., 2021. Human sentence processing: Recurrence or attention? In: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics. Association for Computational Linguistics, pp. 12–22. http://dx.doi.org/10.18653/v1/2021.cmcl-1.2, Online.

Metusalem, R., Kutas, M., Urbach, T.P., Hare, M., McRae, K., Elman, J.L., 2012. Generalized event knowledge activation during online sentence comprehension. J. Mem. Lang. 66 (4), 545–567. http://dx.doi.org/10.1016/j.jml.2012.01.001.

Michaelov, J.A., Bardolph, M., Coulson, S., Bergen, B., 2021. Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude? In: Proceedings of the 43th Annual Meeting of the Cognitive Science Society. pp. 300–306. http://dx.doi.org/10.48448/tj50-1h53.

Michaelov, J.A., Bardolph, M.D., Van Petten, C.K., Bergen, B.K., Coulson, S., 2024. Strong prediction: Language model surprisal explains multiple N400 effects. Neurobiol. Lang. 5 (1), 107–135.

Michaelov, J.A., 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. http://dx.doi.org/10.18653/v1/2020.conll-1.53.

Michaelov, J.A., 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. http://dx.doi.org/10.18653/v1/2022.conll-1.2.

Michaelov, J.A., Coulson, S., Bergen, B.K., 2023. So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. IEEE Transactions on Cognitive and Developmental Systems 15 (3), 1033–1042. http://dx.doi.org/10.1109/TCDS.2022.3176783.

Minixhofer, B., 2020. GerPT2: German large and small versions of GPT2. http://dx.doi.org/10.5281/zenodo.5509984, URL https://github.com/bminixhofer/gerpt2.

Minixhofer, B., Paischer, F., Rekabsaz, N., 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In: Carpuat, M., de Marneffe, M.-C., Meza Ruiz, I.V. (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Seattle, United States, pp. 3992–4006. http://dx.doi.org/10.18653/v1/2022.naacl-main.293, URL https://aclanthology.org/2022.naacl-main.293/.

Mitchell, J., Lapata, M., Demberg, V., Keller, F., 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 196–206, URL https://aclanthology.org/P10-1021/.

Nair, S., Resnik, P., 2023. Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship?. In: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp. 11251–11260. http://dx.doi.org/10.18653/v1/2023.findings-emnlp.752, URL https://aclanthology.org/2023.findings-emnlp.752/.

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. Cogn. Brain Res. 24 (3), 691–701. http://dx.doi.org/10.1016/j.cogbrainres.2005.04.003.

Oh, B.D., Schuler, W., 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In: Bouamor, H., Pino, J., Bali, K. (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, Singapore, pp. 1915–1921. http://dx.doi.org/10.18653/v1/2023.findings-emnlp.128, URL https://aclanthology.org/2023.findings-emnlp.128/.

Oh, B.D., Schuler, W., 2023b. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? Trans. Assoc. Comput. Linguist. 11, 336–350. http://dx.doi.org/10.1162/tacl_a_00548, URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00548/115371/Why-Does-Surprisal-From-Larger-Transformer-Based.

Oh, B.D., Yue, S., Schuler, W., 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In: Graham, Y., Purver, M. (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, St. Julian's, Malta, pp. 2644–2663, URL https://aclanthology.org/2024.eacl-long.162/.

Piantadosi, S.T., 2023. Modern language models refute Chomskys approach to language. In: Gibson, E., Poliak, M. (Eds.), From fieldwork to linguistic theory: A tribute to Dan Everett. Language Science Press Berlin, Germany, pp. 353–414. http://dx.doi.org/10.5281/zenodo.12665933.

Pimentel, T., Meister, C., 2024. How to compute the probability of a word. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp. 18358–18375. http://dx.doi.org/10.18653/v1/2024.emnlp-main.1020, URL https://aclanthology.org/2024.emnlp-main.1020/.

Plüster, B., 2023. LeoLM: Igniting german-language LLM research. URL https://laion.ai/blog/leo-lm/.

Rai, D., Zhou, Y., Feng, S., Saparov, A., Yao, Z., 2024. A practical review of mechanistic interpretability for transformer-based language models. arXiv preprint arXiv:2407.02646.

Schrimpf, M., Blank, I.A., Tuckute, G., Kauf, C., Hosseini, E.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E., 2021. The neural architecture of language: Integrative modeling converges on predictive processing. Proc. Natl. Acad. Sci. 118 (45), e2105646118. http://dx.doi.org/10.1073/pnas.2105646118.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., Levy, R., 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. Proc. Natl. Acad. Sci. 121 (10), e2307876121. http://dx.doi.org/10.1073/pnas.2307876121, URL https://pnas.org/doi/10.1073/pnas.2307876121.

Smith, N.J., Kutas, M., 2015. Regression-based estimation of ERP waveforms: I. the rERP framework. Psychophysiology 52 (2), 157–168. http://dx.doi.org/10.1111/psyp.12317.

Smith, N.J., Levy, R., 2008. Optimal processing times in reading: A formal model and empirical investigation. In: Love, B.C., McRae, K., Sloutsky, V.M. (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society. Cognitive Science Society, URL https://escholarship.org/uc/item/3mr8m3rf.

Speer, R., 2022. rspeer/wordfreq: v3.0. Zenodo, http://dx.doi.org/10.5281/zenodo.7199437.

Stolcke, A., 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. In: Hirschberg, J. (Ed.), Comput. Linguist. 21 (2), 165–201, URL https://aclanthology.org/J95-2002/.

Taylor, W.L., 1953. "Cloze procedure": A new tool for measuring readability. Journal. Mass Commun. Q. 30 (4), 415–433.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS '17, Curran Associates Inc., Red Hook, NY, USA, pp. 6000–6010.

Venhuizen, N.J., Brouwer, H., 2025. Referential retrieval and integration in language comprehension: An electrophysiological perspective. Psychol Rev http://dx.doi.org/10.1037/rev0000530.

Venhuizen, N.J., Crocker, M.W., Brouwer, H., 2019. Expectation-based comprehension: Modeling the interaction of world knowledge and linguistic experience. Discourse Process. 56 (3), 229–255. http://dx.doi.org/10.1080/0163853X.2018.1448677.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., Steinhardt, J., 2022. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. arXiv preprint arXiv:2211.00593.

Wilcox, E.G., Gauthier, J., Hu, J., Qian, P., Levy, R., 2020. On the predictive power of neural language models for human real-time comprehension behavior. http://dx.doi.org/10.48550/ARXIV.2006.01912, URL https://arxiv.org/abs/2006.01912. Version Number: 1.

Wilcox, E.G., Meister, C., Cotterell, R., Pimentel, T., 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Singapore, pp. 7503–7511. http://dx.doi.org/10.18653/v1/2023.emnlp-main.466, URL https://aclanthology.org/2023.emnlp-main.466.

Wilcox, E.G., Pimentel, T., Meister, C., Cotterell, R., Levy, R.P., 2023b. Testing the predictions of surprisal theory in 11 languages. Trans. Assoc. Comput. Linguist. 11, 1451–1470. http://dx.doi.org/10.1162/tacl_a_00612, URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00612/118718/Testing-the-Predictions-of-Surprisal-Theory-in-11.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A., 2020. Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Association for Computational Linguistics, pp. 38–45. http://dx.doi.org/10.18653/v1/2020.emnlp-demos.6, Online, URL https://aclanthology.org/2020.emnlp-demos.6/.

Xu, H., Nakanishi, M., Coulson, S., 2024. Revisiting joke comprehension with surprisal and contextual similarity: Implication from N400 and P600 components. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 46, URL https://escholarship.org/uc/item/01n9j76q.