LLMs Electrified: Early and deep layers differentially correlate with the N400 and P600 in language comprehension

Benedict Krieger (bkrieger@lst.uni-saarland.de)[1], **Matthew W. Crocker**[1] **& Harm Brouwer**[2]
[1]Department of Language Science and Technology, Saarland University
[2]Department of Cognitive Science and Artificial Intelligence, Tilburg University

Abstract

Recent research has examined the extent to which large language models (LLMs) can model the temporal dynamics of online language comprehension, as indexed by the N400 and P600 components of the event-related potential (ERP) signal. To better understand whether the internal representations of LLMs are consistent with distinct stages of comprehension, we employ representational similarity analysis (RSA) on a German ERP study, that found the N400 to be sensitive to association and expectancy, and the P600 to be sensitive to expectancy alone. We find that earlier layers show a stronger correlation with association, whereas deeper layers are more strongly correlated to expectancy. Similarly, correlations to the N400 are stronger at earlier and intermediate layers, consistent with its sensitivity to association, while correlations with both the N400 and P600 continue to increase in deeper layers, reflecting the influence of expectancy on both components. These results are consistent with independent stages of processing proposed by neurocognitive theories, such as Retrieval-Integration theory, suggesting that LLMs may contribute to our understanding of the temporal dynamics of language comprehension - as indexed by ERPs - at a more mechanistic level.

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

Introduction

Large language models (LLMs) are not explicitly designed to be models of human comprehension. Yet, their unprecedented ability to understand and produce natural language has led to a strong interest among researchers in assessing the extent to which LLMs can be used to model various neural and behavioral correlates of human comprehension (Goldstein et al., 2025; Hu et al., 2025; Kuribayashi et al., 2025), such as the N400 and the P600, the two most prominent language-related components of the event-related potential (ERP) signal.

These components are characterized by different functional sensitivities to the linguistic properties of the context. While both the N400 and the P600 are modulated by expectancy, only the N400 is sensitive to association (Aurnhammer et al., 2021, 2023; Delogu et al., 2019; Thornhill & Van Petten, 2012).[1] In particular, fully crossing association and expectancy in a 2x2 factorial design, Aurnhammer et al. (2021) observed additive effects of association and expectancy in the N400, but only a main effect of expectancy in the P600 (Fig. 1, left).

These different sensitivities have motivated neurocognitive theories that link the components to distinct stages of processing. Retrieval-Integration theory for example, proposes that the N400 indexes the retrieval of the meaning of the incoming word from long term memory, and the P600 indexes the compositional integration of this meaning into a mental representation of what is being communicated (Brouwer et al., 2012, 2017). As a consequence, the N400 is predicted to be sensitive to both the association and the expectancy of a word in context, whereas the P600 is predicted to be sensitive to expectancy only. Critically, retrieval and integration continuously proceed in cycles, updating the mental utterance representation with each new incoming word, and forming expectations about the upcoming words.

The notion that expectations play an essential role in human comprehension is also widely supported by instantiations of surprisal theory, which posits that the overall processing effort of a word is proportional to its contextual expectancy (Hale, 2001; Levy, 2008). Crucially, a substantial part of this evidence stems from operationalizations of surprisal through modern LLMs (see Wilcox et al., 2023). Moreover, LLMs have been shown to yield internal representations that correlate well with brain activity during language comprehension, and that have been used successfully in mapping models which relate these representations to brain activity (see Brouwer, 2026), even though LLMs fundamentally differ in certain aspects from the human language processor, such as requiring large amounts of training data and processing the preceding context non-sequentially.

LLMs are deep neural networks trained on the self-supervised task of next word prediction. Futrell and Mahowald (2025) argue that it is the difficulty of the prediction task that may lead to a partial convergence of the human brain and LLMs, as an instance of the so-called contravariance principle (Cao & Yamins, 2024). Since meaning retrieval and integration are central stages in the formation of next-word expectations during comprehension, and modern LLMs excel at next-word prediction, the question arises, if LLMs may also instantiate mechanisms that resemble those of meaning retrieval and integration. To the extent that LLMs and the brain indeed converge on a similar set of core strategies, we should expect to observe evidence for similar stages of comprehension in LLM architectures, as those indexed by neurobehavioral correlates of processing in the brain.

[1]Although association and expectancy are often confounded in naturalistic data, they are distinct, such that a strongly associated continuation may be unexpected (see Krieger et al., 2025 for a discussion).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

While there are a number of techniques for the mechanistic interpretation of LLMs, a direct mechanistic investigation of their internal representations remains challenging, as these representations are highly distributed in nature, may fulfill different roles at different depths of layers, and may form complex circuits to perform specific sub-tasks (Wang et al., 2022). By using correlational analyses, we aim to identify the extent to which representations at different layers are sensitive to established behavioral predictors like association and expectancy that are known to differentially modulate the N400 and P600. To the extent that these sensitivities are observed, we also consider to what degree they reflect the temporal organization posited by neurocognitive models such as RI-theory.

Indeed, assessing LLM representations with respect to how sensitive they are to contextual manipulations of association and expectancy will allow us to identify candidate representations for retrieval and integration, which are suitable for a deeper mechanistic investigation. Krieger et al. (2025) have found LLM surprisal, which is collected from the output layer, to exhibit sensitivity to both association and expectancy (see Fig. 1, right). The question thus arises whether, and if so where in the model, individual sensitivities to association and expectancy emerge, and to which extent they can model the N400 and P600, and hence be informative about the temporal dynamics of human comprehension.

Motivated by the correspondence between the layer time of LLMs and the time course of human online comprehension found by previous research (Goldstein et al., 2025; Kuribayashi et al., 2025), we hypothesize that: (1) earlier layers are more sensitive to contextual association than later layers, are correlated with the N400 response, and potentially reflect mechanistic processes of meaning retrieval, and (2) deeper layers are more sensitive to expectancy than earlier layers, are correlated with the P600 response, and potentially reflect mechanistic processes of meaning integration. On the dataset of Aurnhammer et al. (2021) – an ERP study that contributed to disentangling the differential influences of association and expectancy on the N400 and the P600 – we use representational similarity analysis (RSA; Kriegeskorte et al., 2008) to assess how the representational spaces of ERP responses, behavioral measures of association and expectancy, and `GerPT-2 large` representations at different depths of layers relate to each other.

Method

The German study by Aurnhammer et al. (2021) features 120 items in four conditions, fully crossing contextual association and expectancy in a context manipulation design (Fig. 2), yielding 480 unique stimuli in total. Expectancy was manipulated through the selectional restrictions of the main verb: "sharpened/ate ... the axe", and was operationalized via cloze probabilities in a separate norming study. Association was manipulated through the lexical content of an intervening adverbial clause: "... before he the [wood stacked]/[movie watched], ... the axe", and was operationalized in a norming

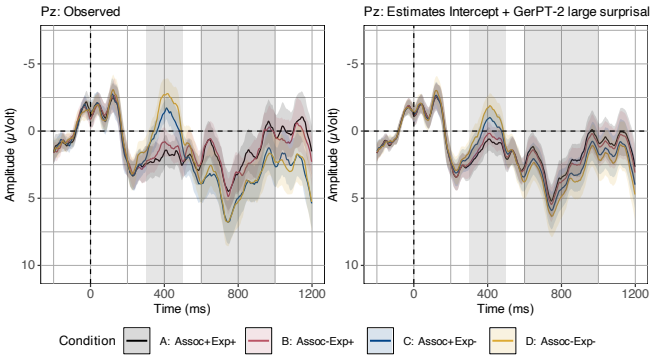

Figure 1: ERP study by Aurnhammer et al. (2021). Left: observed ERPs. Additive effects of association and expectancy in the N400, effect of expectancy in the P600. Right: ERP re-estimation in Krieger et al. (2025), `GerPT-2 large` surprisal showing sensitivity to both association and expectancy.

study, where participants rated the strength of association between the target word and the content words of the adverbial clause on a Likert-scale. The authors observed additive effects of association and expectancy on the N400, but crucially, only a main effect of expectancy on the P600 (Fig. 1, left). We use the publicly available EEG and behavioral data in our analysis. Furthermore, we collect the internal representations of the target words of all stimuli for all 36 layers of `GerPT-2 large` (774 million parameters; Minixhofer, 2020).

Representational similarity analysis

We use RSA to investigate the relationship between the three distinct data types: ERPs (N400 and P600), behavioral measures (association and expectancy), and LLM representations (in all 36 layers). A graphical overview of our approach is shown in Fig. 2. Given a fixed set of stimuli, each data type has its own representational space in which the 480 stimuli are embedded (Fig. 2-1). RSA abstracts away from the individual spaces to a common, comparable format: a representational similarity matrix (RSM). Within each datatype, we construct RSMs by calculating the similarity between all pairs of stimulus representations, and tabulating the results into a square matrix (Fig. 2-2). Given our total of 480 stimuli, the RSMs obtain a dimensionality of 480 x 480. As the RSMs are symmetric, we compare the different spaces by calculating Spearman rank correlations between the upper triangles for all pairs of RSMs (Fig. 2-3). In order to assess which correlations are significant, we conduct a stimulus-label randomization test (Kriegeskorte et al., 2008; Nili et al., 2014). Below, we describe our procedure in more detail for each data type, including relevant preprocessing steps.

ERP data In our analysis, we focus on the ERP-response at the 15 centro-posterior electrodes at which effects were consistently observed in both time windows (C3, Cz, C4, CP5, CP1, CP2, CP6, P7, P3, Pz, P4, P8, O1, Oz, O2). We define the N400 to range from 300-500 ms, and the P600 to range

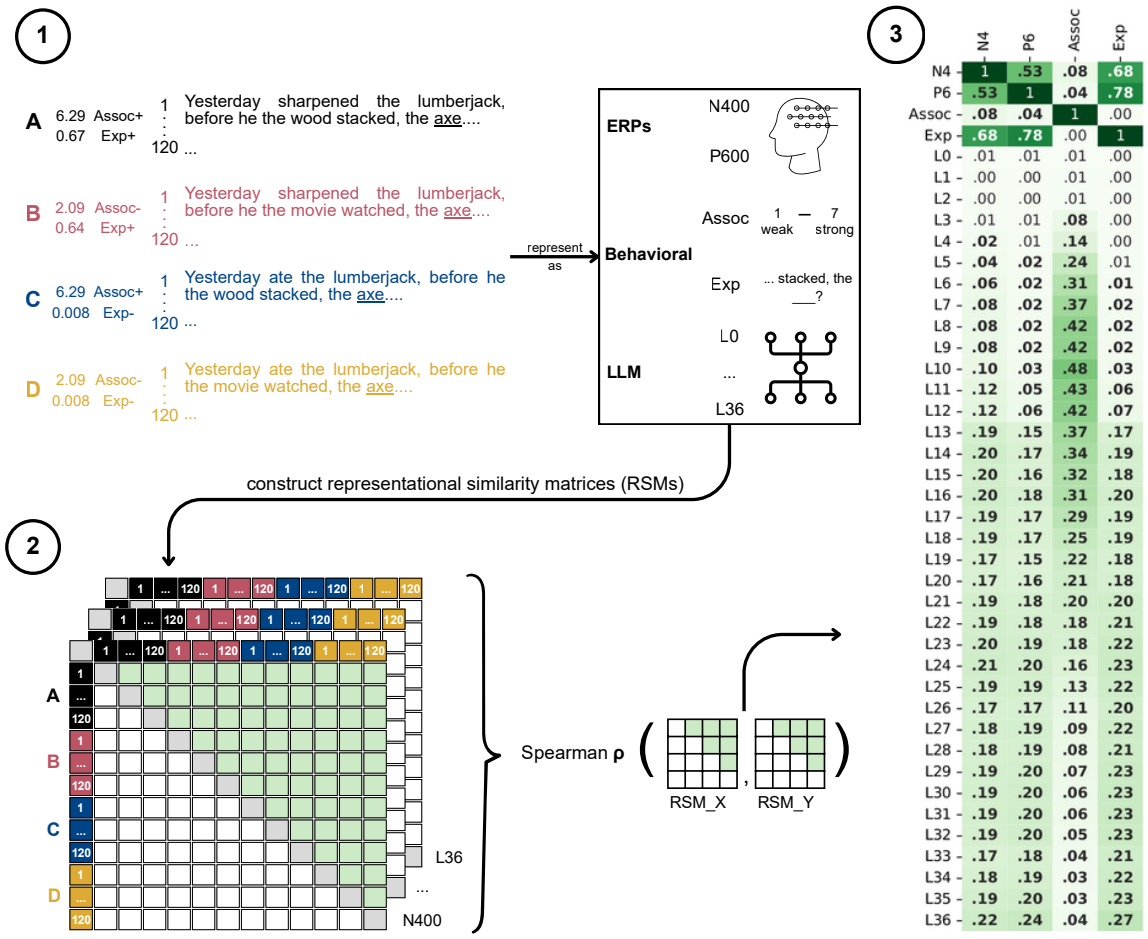

Figure 2: **(1)** The four conditions of Aurnhammer et al. (2021) cross contextual association and expectancy (example item transliterated from German, mean association ratings and cloze probabilities across all items). All stimuli are represented in terms of their target word's average neural response (N400/P600), behavioral measures (association and expectancy), and layer representations in `GerPT-2 large`. **(2)** Representational similarity matrices (RSMs) are constructed for all representation types by comparing all stimulus responses to each other. **(3)** Spearman rank correlations between the upper triangles of all pairs of (symmetric) RSMs reflect the similarities between the different representational spaces (bold-face indicates significance).

from 600-1000 ms post target word onset. For each time window, we construct an RSM from the average ERP response in that time window across subjects on a by-stimulus basis. Importantly, we expect the RSMs to capture the effect structure observed by Aurnhammer et al. (2021), in terms of how distinguishable the representations of the different conditions are to each other. In the N400 RSM, we expect the representations to differ for all inter-condition comparisons. In contrast, in the P600 RSM, we expect only the unexpected conditions (Assoc+Exp− and Assoc−Exp−) to substantially differ from the expected conditions (Assoc+Exp+ and Assoc−Exp+).

RSMs are often computed for each subject separately. However, this presupposes a study design in which all subjects have seen all stimuli. Given the factorial design of Aurnhammer et al. (2021), this is not the case: subjects were presented with only one condition per item to avoid repetition effects. Moreover, some trials had to be rejected due to artifacts. Therefore, the original design precludes us from computing per-subject RSMs. Instead, for each time window and electrode, we compute one average voltage per stimulus across all subjects. However, as each stimulus is only seen by a few subjects (m=8, sd=2), the resultant averages have a high variance, rendering it difficult to observe any systematic differences between conditions in the RSMs.

In order to account for this issue, we apply two preprocessing steps to the ERP data, *prior* to computing RSMs. First, we re-estimate all voltages using regression-based ERP estimation (rERPs; Brouwer, Delogu, and Crocker, 2021; Smith and Kutas, 2015). For each subject, electrode, and timestamp, we fit a simple linear regression model over all trials. As predictors, we use the by-stimulus association rating for the noun of the intervening adverbial clause, and the smoothed and log-transformed cloze probability of the target word:[2]

[2]The log-tranformation of Cloze probability has been shown to provide a closer fit to the ERP data in Aurnhammer et al. (2021). Cloze values are smoothed by adding 0.01.

$$Y = \beta_0 + \beta_1 * \text{association} + \beta_2 * log(\text{Cloze}) + \epsilon \quad (1)$$

This combination of predictors has also been used by Aurnhammer et al. (2021), and has been shown to approximate the entire ERP signal with only small residual error. After fitting the models, we obtain their forward estimates as a new dataset, maintaining the dimensionality of the original data. This effectively factors out the unexplained variance ϵ that remains from re-estimating the ERP signal with the predictors tied to the experimental manipulations, while accounting for by-subject variability. To further aid the interpretability of the RSA, we center the rERP estimates by subtracting the mean across trials within each subject. This has the effect that within each subject, a given trial is either more negative or positive than average within each of the time-windows.

On the re-estimated and centered data, we then compute the by-stimulus voltage averages for each time window and electrode. Since a stimulus is defined by the combination of an item and a condition, we group the data points by item and condition, and take their mean at each of the 15 electrodes. This results in a dataset of 480 15-dimensional stimulus representations. In order to more easily obtain a bounded similarity metric, we transform values to be bounded between zero and one by applying min-max scaling. As a result, the maximum possible Euclidean distance in this 15-dimensional representational space becomes $\sqrt{15}$.

Finally, we compute the Euclidean distance for all pairs of stimuli, and normalize the resulting unbounded distances by this maximum possible distance, in order to obtain a distance which is bounded between zero and one. We then convert these distances to similarities by subtracting them from one, and enter them into a symmetric square RSM.

Behavioral measures The predictors central to the experimental manipulations in Aurnhammer et al. (2021) are the by-stimulus association ratings of the target word, and the smoothed log-transformed Cloze probability of the target word. Per stimulus, there is a single mean association rating, computed from the individual ratings of the participants of the separate norming study conducted by Aurnhammer et al. (2021). Note that the association rating for each stimulus is identical in the two associated conditions (high), as well as the two unassociated conditions (low). Similarily, there is a single Cloze probability per stimulus, computed as the proportion of participants who filled in the target word in the Cloze task (cf. Taylor, 1953). We note, that the one-dimensionality of the behavioral measures necessitates a univariate approach in computing RSMs (see Bruera et al., 2023 for another univariate application of RSA in the context of concreteness ratings).

We compute an RSM for each association and expectancy. For a bounded similarity metric, we apply min-max scaling to each of the predictors across the 480 stimuli. We then take the absolute difference for all pairs of stimuli – which now ranges between zero and one – and convert the distances to similarities by subtracting from one.

LLM representations We present `GerPT-2 large` with each stimulus up until and including the target word. For each target, we extract the internal representations from the residual stream at the output of the input embedding layer (L0) and each of the 36 subsequent transformer blocks (L1-L36). These 1280-dimensional vectors represent the state of the model after the attention and feed-forward updates of each block have been integrated. In cases where the target word is split into multiple subwords by the tokenizer, we compute an average representation across the subwords. Importantly, the study by Aurnhammer et al. (2021) applies a context manipulation design. Therefore, within a single item, the same target word appears in all four conditions. This implies that the target word representation at the input layer starts out almost identical in all four conditions of an item. Throughout the layers, representations become increasingly contextualized. In autoregressive LLMs such as `GerPT-2 large`, this means that information from previous tokens is integrated into later tokens. At the final layer, the representation of the last token of the sequence, i.e., in our case the target word of the stimulus, then encodes contextual information of the entire sequence (Cassani et al., 2023; Ethayarajh, 2019).

The distributed meaning representations are highly sensitive to the lexical content of the context. Due to how the contexts of the items are construed, the overall lexical material is very similar within items. Using multidimensional scaling (MDS), it thus becomes apparent that representations within items form clusters (Fig. 3, left). The representations within the clusters may still be distinguishable, however, observing representational differences between conditions becomes difficult. To account for this observation, we center the representations on a by-item basis. That is, we compute the average representation per item, which we then subtract from each target word representation within the item. In this way, we abstract away from the specific lexical material of the items to a representation which emphasizes the representational differences resulting from the experimental manipulations of the context, namely association and expectancy.[3] On the centered representations, we then use cosine similarity to compute a separate RSM for each LLM layer.

Rank correlations After constructing RSMs for ERPs, behavioral measures, and LLM representations, we compare the representational spaces to each other by computing Spearman rank correlations between the upper triangles for all pairwise combinations of RSMs. In order to assess which of the obtained rank correlations are statistically significant, we conduct a stimulus-label randomization test (Kriegeskorte et al., 2008; Nili et al., 2014). For each pair of RSMs, we first record the observed correlation coefficient. Then, we keep one RSM fixed, and permute the stimulus labels of the other RSM. We record the correlation coefficient between the fixed and the permuted RSM, and repeat this process (n = 10,000) to generate a null distribution of permuted correlation

[3] Note that the syntactic structure is constant across items.

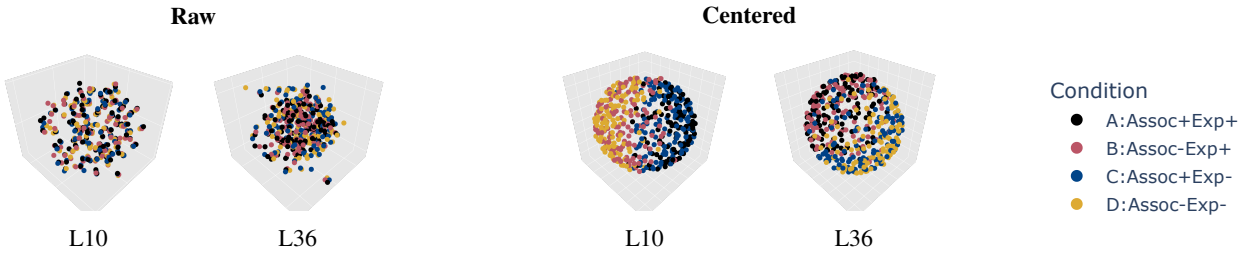

Figure 3: Visualization of `GerPT-2 large` representations through multidimensional scaling at example layers 10 and 36. Left: raw representations form item-specific clusters, rendering the observation of condition-related differences difficult. Right: per item, subtracting the mean item representation from all stimulus representations emphasizes condition-related differences.

coefficients. We compute a one-tailed p-value as the proportion of the permuted correlation coefficients that are greater than or equal to the observed correlation coefficient, and treat all correlation coefficients with Bonferroni-corrected (within representational space) p-values < 0.05 as significant.

Results

Panel 3 in Fig. 2 shows the rank correlations which we obtain from our analysis.[4] Significant correlations, as determined by the stimulus-label randomization test, are printed in bold. Below, we first describe the correlations between the behavioral and the ERP RSMs. We then assess the correlation trajectories from earlier to deeper layer LLM RSMs on the one hand, and both behavioral and ERP RSMs on the other.

Behavioral and ERP RSMs The association and the expectancy RSMs are unrelated, which is expected given the orthogonality of these predictors in the factorial 2x2 design of Aurnhammer et al. (2021). The overall stronger correlation of expectancy compared to association with both ERP components reflects the main effect of expectancy in the original data. The expectancy RSM shows a stronger correlation with the P600 RSM (ρ = .78) relative to the N400 RSM (ρ = .68). Conversely, the association RSM shows a stronger correlation with the N400 RSM (ρ = .08) relative to the P600 RSM (ρ = .04). This pattern is consistent with the differential sensitivities of the ERP components to association and expectancy. There is a moderate correlation between the N400 and P600 RSMs (ρ = .53). All observed correlations are significant.

LLM RSMs The correlation between the association RSM and the LLM RSMs starts to become significant at layer 3. The correlation continues to increase until its maximum at layer 10 (ρ = .48), and then decreases until the final layer (ρ = .04). The correlation between the expectancy RSM and the LLM RSMs starts to become significant at layer 6. The correlation continues to increase until its maximum at layer 36 (ρ = .27), with a noticeable rise at layer 13 (ρ = .17).

The correlations between the ERP RSMs and the LLM RSMs overall increase until the final 36th layer. The correlation between the N400 RSM and the LLM RSMs already starts to become significant at layer 4, while the correlation between the P600 RSM and the LLM RSMs becomes significant at layer 5. In earlier layers (3-13), the magnitude of the N400 correlations is notably stronger than the P600 correlations (e.g., cf. layer 10: N400 ρ = .10, P600 ρ = .03). While the correlation between the N400 RSM and the LLM RSMs reaches its maximum at the final 36th layer (ρ = .22), correlations that are close to this maximum can already be observed at some of the earlier layers: layers 14 - 16 and 23 (ρ = .20), and layer 24 (ρ = .21). In sum, until layer 24, we observe that the N400 correlation is numerically greater than the P600 correlation. At layers 25 and 26 the correlation coefficients are equal. Then, from layer 27 until the final layer, the P600 correlation remains greater than the N400 correlation.

In a follow-up analysis using `Leo-LM` (Plüster, 2023) – a Llama-2 based LLM with 40 hidden layers, 13B parameters, and a different training dataset – we observe similar results: an early peak and subsequent descent of association, a later increase of expectancy towards the deepest layers, a stronger correlation of earlier layers with the N400, and a stronger correlation of deeper layers with the P600.

Discussion

We have used RSA to investigate if and to what extent layers at different depths in `GerPT-2 large` are differentially sensitive to contextual association and expectancy. Our aim was to identify representations that have the potential to index distinct stages of meaning retrieval and integration, which RI-theory has linked to the N400 and P600 components of the event-related potential: while the retrieval stage is facilitated by both association and expectancy and modulates the N400, the integration stage is facilitated by expectancy alone and modulates the P600 (Brouwer et al., 2017).

For each item in the dataset of Aurnhammer et al. (2021), which fully crossed association and expectancy in a 2x2 factorial design, we have computed RSMs from average N400 and P600 responses, behavioral measures of association and expectancy, and LLM representations for the target words from all (36) layers of `GerPT-2 large`. By computing rank correlations on the upper triangles for all pairs of these RSMs, we have then compared the different representational spaces

[4] We omit presenting the correlations between layer RSMs, since these are not of interest for our research question.

to each other.

We observe that the correlation between the association and the LLM RSMs rises and peaks early, and then descends until the final layer. In contrast, the correlation between the expectancy and the LLM RSMs starts to increase at later layers, and continues to rise until the final layer. Importantly, this finding supports our hypothesis that earlier layers are more sensitive to association, while deeper layers are more sensitive to expectancy. This observation is also reflected in the three-dimensional MDS-visualization of the internal representations of `GerPT-2 large` in Fig. 3 (right): after applying the by-item centering to abstract away from the specific lexical material of the items, representations at layer 10 reveal the clustering of the associated (right) versus unassociated (left) conditions, while representations at layer 36 form two major groups that reflect the expected (upper left) versus unexpected (lower right) conditions.

As discussed in the introduction, the Aurnhammer study demonstrated that while the N400 is modulated by both association and expectancy, the P600 is modulated by expectancy alone, and that these components have been argued to index lexical retrieval and semantic integration processes, respectively (Brouwer et al., 2012). We therefore hypothesized that earlier layers in the LLM — to the extent that they are consistent with retrieval — should exhibit a mixed sensitivity to association and expectancy, and may thus be better able to capture the N400. By contrast, deeper layers – which should increasingly reflect expectancy alone – should better index integration, and thus the P600.

Indeed, assessing the correlation trajectories between the ERP RSMs and the LLM RSMs, we find that the N400 correlations are stronger than the P600 correlations in early and middle layers, while the P600 correlations are stronger than the N400 correlations in the deeper layers. Critically, the correlations to the P600 continuously increase until the final layer, reflecting both the decreasing correlation with association and the increasing correlation with expectancy. Despite the decreasing correlation with association, the correlations of the N400 with the LLM RSMs continue to be strong in deeper layers. This suggests that while the correlations between the N400 and LLM RSMs in earlier layers may be predominantly driven by association, the correlations in deeper layers are driven by expectancy.

Importantly, Brouwer, Delogu, Venhuizen, and Crocker (2021) distinguish between two different types of expectancy: while retrieval – as indexed by the N400 – is modulated by the expectancy of *word meaning*, integration – as indexed by the P600 – is modulated by the expectancy of *utterance meaning*. Our finding that middle LLM layers correlate well with the N400 is thus consistent with the notion of retrieving word meaning from semantic long-term memory (Kutas & Federmeier, 2000). Critically, this finding also aligns with the observation that word representations from middle LLM layers best capture explicit semantic associations, and are generally best suited to model lexical semantics (Cassani et al., 2023, p.28). By contrast, representations in deeper layers have been shown to increasingly encode contextual information necessary to capture the meaning of the entire sequence. Since we observe that the deepest LLM layers are correlated with the P600 – to a stronger degree than with the N400 – it thus appears plausible, that the LLM-internal processes which lead to those representations may resemble integration processes during language comprehension.

In sum, the differential sensitivity to association and expectancy that we observe across the layers of the LLM, is consistent with the view that LLMs may converge on a solution similar to that proposed by neurocognitive accounts of the N400 and P600, such as RI-theory, which not only predicts this pattern, but does so in a single-stream architecture similar to that of LLMs, rather than appealing to multi-stream mechanisms (see Brouwer et al., 2012 for a review). Early and middle layers are consistent with retrieval (N400), while the deepest layers are consistent with integration (P600). Our results motivate a deeper, more mechanistic investigation of each of these layer groups, which RSA has enabled us to identify. While we here focus on evaluating the internal representations of the model, such a deeper investigation may focus on the different roles that have been proposed for distinct architectural components, such as multilayer perceptrons serving as a storage for factual knowledge, and attention heads moving relevant information between tokens (Elhage et al., 2021; Geva et al., 2023; Meng et al., 2022).

Conclusion

We have examined how differentially sensitive the internal representations at different layers in `GerPT-2 large` are to manipulations of contextual association and expectancy, and how this sensitivity may influence their ability to model the N400 and P600 components of the event-related potential. Based on a study that disentangled the influences of association and expectancy on these ERP components, we have used RSA to compare representational spaces derived from behavioral measures, average ERP responses, and LLM representations. We have found early LLM layers to strongly reflect association with the context, and deep layers to capture expectancy. While the sensitivity to expectancy continuously increases until the final layer, the sensitivity to association decreases after an early peak. Our results suggest that early and middle layers may have the potential to better approximate meaning retrieval as indexed the N400, and that deeper layers may better approximate meaning integration as indexed by the P600. Our work constitutes an important next step in identifying more mechanistic LLM-derived linking hypotheses to the ERP components, and thus answering the question to which extent LLMs can contribute to our understanding of the temporal dynamics of online language comprehension.

Acknowledgments

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

232722074—SFB 1102. The first author would also like to thank Stan Bergey for informative and valuable discussions on the topic of representational similarity analysis.

References

Aurnhammer, C., Delogu, F., Brouwer, H., & Crocker, M. (2023). The P600 as a continuous index of integration effort. *Psychophysiology*. https://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. https://doi.org/10.1371/journal.pone.0257430

Brouwer, H. (2026). Mapping meaning in the brain's language. *Cortex*, *194*, 12–21. https://doi.org/10.1016/j.cortex.2025.10.012

Brouwer, H., Crocker, M., Venhuizen, N., & Hoeks, J. (2017). A Neurocomputational Model of the N400 and the P600 in Language Processing. *Cognitive Science*, *41*(S6), 1318–1352. https://doi.org/10.1111/cogs.12461

Brouwer, H., Delogu, F., & Crocker, M. W. (2021). Splitting event-related potentials: Modeling latent components using regression-based waveform estimation. *European Journal of Neuroscience*, *53*(4), 974–995. https://doi.org/10.1111/ejn.14961

Brouwer, H., Delogu, F., Venhuizen, N., & Crocker, M. (2021). Neurobehavioral Correlates of Surprisal in Language Comprehension: A Neurocomputational Model. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.615538

Brouwer, H., Fitz, H., & Hoeks, J. (2012). Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research*, *1446*, 127–143. https://doi.org/10.1016/j.brainres.2012.01.055

Bruera, A., Tao, Y., Anderson, A., Çokal, D., Haber, J., & Poesio, M. (2023). Modeling Brain Representations of Words' Concreteness in Context Using GPT-2 and Human Ratings. *Cognitive Science*, *47*(12), e13388. https://doi.org/10.1111/cogs.13388

Cao, R., & Yamins, D. (2024). Explanatory models in neuroscience, Part 2: Functional intelligibility and the contravariance principle. *Cognitive Systems Research*, *85*, 101200. https://doi.org/10.1016/j.cogsys.2023.101200

Cassani, G., Bianchi, F., Attanasio, G., Marelli, M., & Guenther, F. (2023). Meaning Modulations and Stability in Large Language Models: An Analysis of BERT Embeddings for Psycholinguistic Research. https://doi.org/10.31234/osf.io/b45ys

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

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, *1*(1), 12.

Ethayarajh, K. (2019). How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55–65. https://doi.org/10.18653/v1/D19-1006

Futrell, R., & Mahowald, K. (2025). How Linguistics Learned to Stop Worrying and Love the Language Models. *Behavioral and Brain Sciences*, 1–98. https://doi.org/10.1017/S0140525X2510112X

Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). Dissecting Recall of Factual Associations in Auto-Regressive Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12216–12235. https://doi.org/10.18653/v1/2023.emnlp-main.751

Goldstein, A., Ham, E., Schain, M., Nastase, S. A., Aubrey, B., Zada, Z., Grinstein-Dabush, A., Gazula, H., Feder, A., Doyle, W., Devore, S., Dugan, P., Friedman, D., Brenner, M., Hassidim, A., Matias, Y., Devinsky, O., Siegelman, N., Flinker, A., ... Hasson, U. (2025). Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. *Nature Communications*, *16*(1), 10529. https://doi.org/10.1038/s41467-025-65518-0

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. *Proceedings of NAACL 2001*, *2*. https://doi.org/10.3115/1073336.1073357

Hu, J., Lepori, M. A., & Franke, M. (2025). Signatures of human-like processing in Transformer forward passes. https://doi.org/10.48550/ARXIV.2504.14107

Krieger, B., Brouwer, H., Aurnhammer, C., & Crocker, M. W. (2025). On the limits of LLM surprisal as a functional explanation of the N400 and P600. *Brain Research*, *1865*, 149841. https://doi.org/10.1016/j.brainres.2025.149841

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249. https://doi.org/https://doi.org/10.3389/neuro.06.004.2008

Kuribayashi, T., Oseki, Y., Taieb, S. B., Inui, K., & Baldwin, T. (2025). Large Language Models Are Human-Like Internally. *Transactions of the Association for Computational Linguistics*, *13*, 1743–1766. https://doi.org/10.1162/TACL.a.58

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*(12), 463–470. https://doi.org/10.1016/S1364-6613(00)01560-6

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

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT [arXiv:2202.05262]. *Advances in Neural Information Processing Systems*, *36*.

Minixhofer, B. (2020). GerPT2: German large and small versions of GPT2. https://doi.org/10.5281/ZENODO.5509984

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis (A. Prlic, Ed.). *PLoS Computational Biology*, *10*(4), e1003553. https://doi.org/10.1371/journal.pcbi.1003553

Plüster, B. (2023). LeoLM: Igniting German-Language LLM Research. https://laion.ai/blog/leo-lm/

Smith, N., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, *52*. https://doi.org/10.1111/psyp.12317

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism & Mass Communication Quarterly*, *30*, 415–433.

Thornhill, D. E., & Van Petten, C. (2012). Lexical versus conceptual anticipation during sentence processing: Frontal positivity and N400 ERP components. *International Journal of Psychophysiology*, *83*(3), 382–392. https://doi.org/10.1016/j.ijpsycho.2011.12.007

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. https://doi.org/10.48550/ARXIV.2211.00593

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612