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

Benedict Krieger (bkrieger@lst.uni-saarland.de), Matthew W. Crocker, Harm Brouwer

Recent research has posed the question to which extent large language models (LLMs) can model the temporal dynamics of online language comprehension [1-3], as indexed by the components of the event-related potential (ERP) signal. Critically, while the N400 and the P600 components are both sensitive to expectancy and hence can to some degree be estimated using LLM surprisal [2,4]. The N400 component, however, is also sensitive to association, while the P600 is not [5-6]. This differential sensitivity has motivated neurocognitive theories that link these components to functionally distinct subprocesses of comprehension [7]. To better understand whether LLMs similarly reflect these subprocesses, we exploit representational similarity analysis (RSA; [8]) to investigate if and to what degree layers at different depths are (differentially) sensitive to association and expectancy, with the aim of identifying more mechanistically motivated linking hypotheses with the N400 and P600 components [2,7].

For each item in the German dataset of Aurnhammer et al. [5], which fully crossed association and expectancy in a 2x2 factorial design, we compute representational similarity matrices (RSMs) for neural responses (an average ERP response for each component), behavioral measures of association and expectancy, and LLM representations for the target words from all (36) layers of GerPT-2 large (see Fig. 1-1 and 1-2). RSMs compare all stimulus responses to each other, thus reflecting the internal representational geometry of each space (ERPs, behavioral measures, LLM). To quantify the representational similarity between spaces, we then compute rank correlations between the upper triangles of all RSMs (Fig. 1-3).

Crucially, we find that the correlation between the association RSM and the RSMs from the LLM increases during early layers – peaking at layer 10 – and then decreases until the final 36th layer. In contrast, the correlation between the expectancy RSM and the LLM RSMs starts to notably increase at layer 13, and continues to increase until the final layer. Indeed, this indicates that representations in earlier LLM layers are more sensitive to simple association, while representations in deeper layers are more sensitive to true expectancy. Similarly, expectancy is most strongly correlated with the P600 compared to the N400, while association is better correlated with the N400 than the P600. The overall greater magnitude of the expectancy correlations highlights its importance for both components in this particular study, and is consistent with the overall correlation pattern of the components to the LLM layers; that is, while for both components the correlations increase from earlier to deeper layers, the N400 shows higher correlations with earlier layers, consistent with the additive effects of expectancy and association on this component observed by Aurnhammer et al. [5].

Our results thus reveal the ability of LLM representations at different layers to better model the differential sensitivity of the N400 and P600 to properties such as association and expectancy, in a manner that LLM surprisal alone cannot [2]. The enhanced influence of association we observe at earlier layers is also consistent with subprocesses motivated independently by neurocognitive models [9], and suggests that LLMs may contribute to our understanding of the temporal dynamics of language comprehension – as indexed by ERPs – at a more mechanistic level [7].

Figures

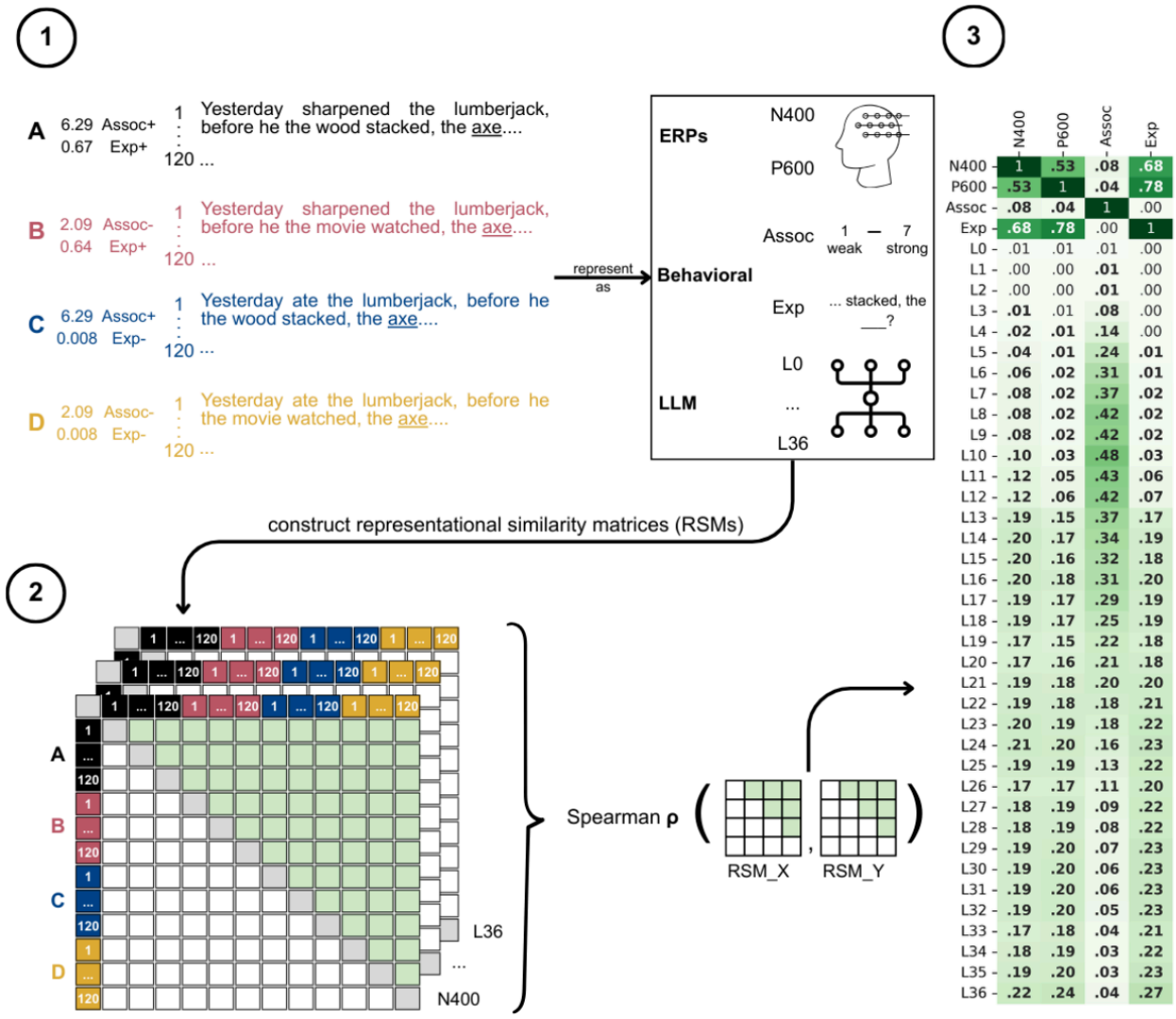

Fig. 1. (1) The four conditions of [5] cross contextual association and expectancy (example item transliterated from German, mean association ratings and cloze probabilities across all items). All stimuli are represented in terms of their target word's average neural response (N400/P600), behavioral measures (association and expectancy), and layer activations 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. Significant correlations ($p < 0.05$), as determined by a stimulus-label randomization test, are printed in bold.

References

[1] Hu *et al.*, (2025), *arXiv*; **[2]** Krieger *et al.*, (2025), *Brain Res.*; **[3]** Kuribayashi *et al.*, (2025), *arXiv*; **[4]** Michaelov *et al.*, (2024), *Neurobiol. Lang.*; **[5]** Aurnhammer *et al.*, (2021), *PLOS ONE*; **[6]** Delogu *et al.*, (2019), *Brain Cogn.*; **[7]** Brouwer, (2026), *Cortex*; **[8]** Kriegeskorte *et al.*, (2008), *Front. Sys. Neurosci.*; **[9]** Brouwer *et al.*, (2017), *Cognit. Sci.*