Beyond surprisal: GPT-derived attention metrics offer additional explanatory power in predicting the N400 during naturalistic reading

Sven Terpstra, Willem Zuidema, Marianne de Heer Kloots
Amsterdam Brain & Cognition; Institute for Logic, Language, and Computation; University of Amsterdam
s.terpstra@outlook.com, W.H.Zuidema@uva.nl, m.l.s.deheerkloots@uva.nl

The N400 component of the EEG signal is a well-established neural correlate of real-time language comprehension, sensitive to a range of lexical and contextual variables. While earlier studies have linked measures of word predictability (such as language model surprisal) to N400 amplitude, the N400 is known to reflect a broader array of cognitive processes than lexical expectation alone. For reading time prediction, prior work has found significant increases in language model psychometric predictive power based on metrics that go beyond surprisal, by capturing incremental changes in model attention patterns across timesteps (Oh & Schuler, 2022; Ryu & Lewis, 2025). In this study, we examine whether additional metrics derived from GPT language models, particularly attention-based measures, can provide complementary predictive value for N400 amplitude.

We evaluate language model N400 predictivity on EEG data from the RaCCooNs dataset of Dutch naturalistic reading (Frank & Aumeistere, 2024), which contains data from 37 participants reading 200 Dutch sentences each. We fed the same sentences to four GPT-based language models (including monolingual and multilingual variants), and extracted surprisal, lexical entropy, and attention-derived metrics (the three metrics with highest explanatory value in a reading times study by Oh and Schuler, 2022). To compare the predictive power of these different metrics, we fit mixed-effect regression models with language model-based metrics as fixed-effect predictors and word- and participant-level random effects, including word frequency, word position and word length as covariates of no interest. Next we examined the quality of these model fits across metrics.

We show that GPT surprisal robustly predicts N400 amplitude during naturalistic reading of Dutch (Figure 1). Crucially, we demonstrate for the first time that attention-based metrics, namely Normalized Attention Entropy (NAE), ΔNAE, and Manhattan Distance (MD), offer significant additional explanatory power beyond surprisal and lexical entropy in predicting the N400 (Figure 2), based on analyses of two representative GPT models. These findings demonstrate that attention metrics from Transformer models can potentially serve as powerful and cognitively informative predictors of neural language processing in a naturalistic context.

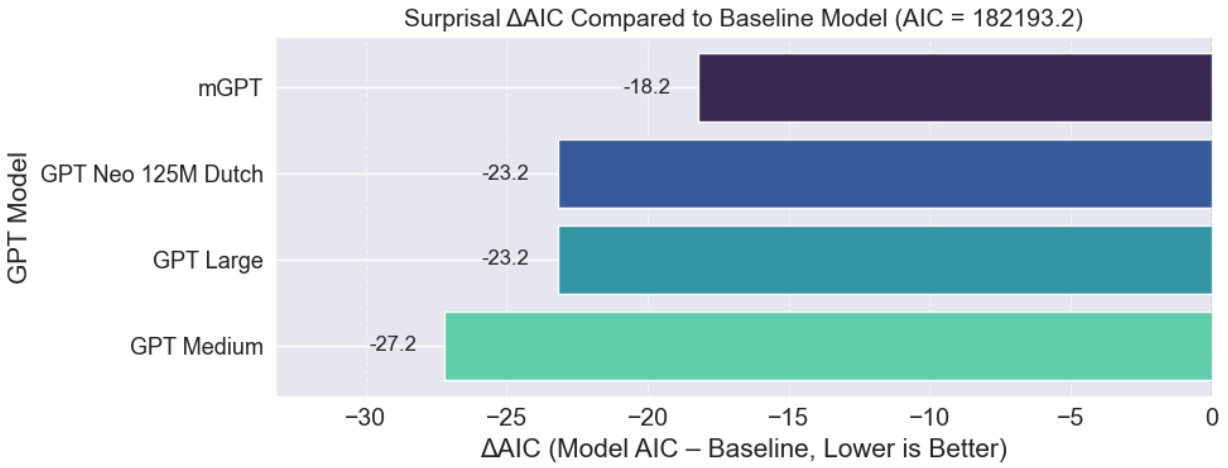

Figure 1. Among the surprisal predictors tested, GPT Medium surprisal yields the greatest improvement in AIC score. The bars represent the decrease in AIC score for a mixed effects model when adding surprisal as a predictor compared to a model that only contains control variables; A larger decrease in AIC indicates improved model fit and suggests better generalizability to similar unseen data.

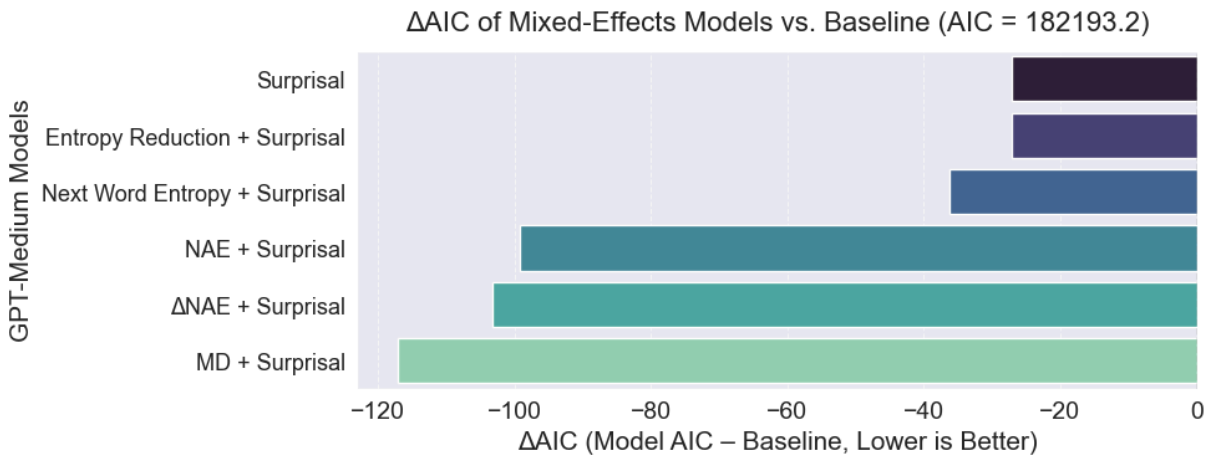

Figure 2. Including attention-based metrics extracted from the final layer of GPT Medium as predictors yields larger AIC improvements for mixed effects models compared to lexical entropy measures. Bars represent the change in Akaike Information Criterion (ΔAIC) relative to a baseline model with only control variables; more negative values indicate better model fit.

References

Frank, S. L., & Aumeistere, A. (2024). An eye-tracking-with-EEG coregistration corpus of narrative sentences. *Language Resources and Evaluation*, *58*(2), 641-657.

Oh, B. D., & Schuler, W. (2022). Entropy-and Distance-Based Predictors From GPT-2 Attention Patterns Predict Reading Times Over and Above GPT-2 Surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 9324-9334).

Ryu, H.R. & Lewis, R.L. (2025), Memory for prediction: A Transformer-based theory of sentence processing. *Journal of Memory and Language,* Volume 145.